DOCUMENT RESUME

ED 329 100 FL 018 408

AUTHOR Stansfield, Charles W.; Kenyon, Dorry Mann
TITLE Development of Semi-Direct Tests of Oral Proficiency in Hausa, Hebrew, Indonesian and Portuguese.
INSTITUTION Center for Applied Linguistics, Arlington, Va.
SPONS AGENCY Office of International Education (ED), Washington, DC.
PUB DATE 90
CONTRACT G008740397
NOTE 109p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Hausa; *Hebrew; *Indonesian; Language Proficiency; *Language Tests; *Oral Language; *Test Construction; Test Format; Test Manuals; Test Validity; Uncommonly Taught Languages

ABSTRACT

This project extended the application of a model for development of semi-direct tests of oral proficiency, originally developed for Chinese, to a diverse set of less commonly taught languages spanning various language families and representing diverse cultural backgrounds. This second, final report covers development of tests for Hebrew, Hausa, and Indonesian, each described separately. An introductory section gives an overview of the project and describes the prototypical form of the semi-direct test. For each language, the test development process is outlined in terms of major project activities, test form trials, validation studies, subject response to the test on questionnaires, and test operationalization, a process involving printing, reproduction of masters, and development of test manuals and examinee handbooks. A financial status report for the project is also included. The substantial appendixes include tests, therefore they are not appended. (MSE)

Development of Semi-Direct Tests of Oral Proficiency in Hausa, Hebrew, Indonesian and Portuguese

Final Project Report for

Grant No. G008740397

Office of International Research and Studies

U.S. Department of Education

Charles W. Stansfield

Project Director

Dorry Mann Kenyon

Test Development Coordinator

Center for Applied Linguistics
1118 22nd Street, NW
Washington, DC 20037

Table of Contents

1. Introduction 1

   1.1 Overview 1
   1.2 Proto-typical Format of the Semi-direct Tests 2

2. Hebrew Speaking Test 4

   2.1 Major Project Activities 4
   2.2 Trialing of the Test Forms 5
   2.3 Validation Study 6
   2.4 USA Study 9
   2.5 Israeli Study 19
   2.6 Subject Response to the Test 27
   2.7 Operationalization of the Test 38

3. Indonesian Speaking Test 40

   3.1 Major Project Activities 40
   3.2 Trialing of the Test Forms 40
   3.3 Validation Study 42
   3.4 Subject Response to the Test 53
   3.5 Operationalization of the Test 64

4. Hausa Speaking Test 65

   4.1 Major Project Activities 65
   4.2 Trialing of the Test Forms 65
   4.3 Validation Study 67
   4.4 Subject Response to the Test 82
   4.5 Operationalization of the Test 87

References 88

Financial Status Report 89

APPENDICES

A-1: Hebrew Speaking Test Official Test Manual
A-2: Hebrew Speaking Test Examinee Handbook
A-3: Data Collections Forms Used in the Trialing of the HeST
A-4: Questionnaire for Participants in the Hebrew Speaking Test Validation Study
A-5: Comments from the Participants in the Hebrew Speaking Test Validation Study

B-1: Indonesian Speaking Test Official Test Manual
B-2: Indonesian Speaking Test Examinee Handbook
B-3: Data Collections Forms Used in the Trialing of the IST

B-4: Questionnaire for Participants in the Indonesian Speaking Test Validation Study
B-5: Comments from the Participants in the Indonesian Speaking Test Validation Study

C-1: Hausa Speaking Test Official Test Manual
C-2: Hausa Speaking Test Examinee Handbook
C-3: Data Collections Forms Used in the Trialing of the HaST
C-4: Questionnaire for Participants in the Hausa Speaking Test Validation Study
C-5: Comments from the Participants in the Hausa Speaking Test Validation Study
C-6: Data Collection Form for the HaST Validation Study

1. INTRODUCTION

This is the final report for the project entitled "Development of Semi-Direct Tests of Oral Proficiency in Hausa, Hebrew, Indonesian and Portuguese." The goal of the project was to extend the application of a model for the development of semi-direct tests of oral proficiency, originally used by the Center for Applied Linguistics (CAL) in the development of the proto-typical Chinese Speaking Test (CST) (Clark, 1986; Clark and Li, 1986), to a diverse set of less commonly taught languages spanning various language families and representing diverse cultural backgrounds. The year one project report covers the development of the Portuguese Speaking Test (PST). For further information on that project, readers are referred to Stansfield and Kenyon (1988). This report covers the development of the semi-direct tests in Hebrew, Indonesian and Hausa. Each will be treated in a separate section.

1.1 OVERVIEW

The past decade has witnessed a major theoretical and practical development in the field of foreign language assessment. This development is the application of a "proficiency" orientation in the testing of foreign language competence. At the forefront in foreign language proficiency measures is the oral proficiency interview, a direct face-to-face evaluation of the foreign language learner's competence conducted by trained interviewers and raters. In the government setting, the testing committee of the Interagency Language Roundtable (ILR) has been spearheading the movement. In academia, the American Council on the Teaching of Foreign Languages (ACTFL) has been overseeing its extension into American college and university programs.

For the less commonly taught languages (LCTLs), however, the practical problems of organization and economics often impede having adequate numbers of individuals available to test competency via the live interview. Thus, semi-direct testing (using recorded and printed stimuli and recorded responses) is an efficient and feasible approach to proficiency measurement in the LCTLs. This

approach eliminates the need to try to sustain a costly and labor intensive face-to-face (direct) oral proficiency interview program for low-volume languages whose enrollment figures may be unstable from year to year. At the same time, it ensures the benefits derived from a continual assessment program as the impetus for competency-based learning for students of the LCTLs.

1.2 PROTO-TYPICAL FORMAT OF THE SEMI-DIRECT TESTS

Each of the three separate test development projects described in this report began with the same format that had been used successfully in the development of the CST and PST. However, as the development for each project continued, the format in each case was modified to reflect concerns specific to the testing of that target language. Because the tests share a common format, that format is presented here. Modifications will be outlined in the section on the development of the individual test.

There are three components in each test: the Master Test Tape, the Test Booklet, and the Examinee Response Tape. (The last is a blank cassette on which the examinee's responses are recorded.) The Master Test Tape begins with the reading of the general test directions, which the examinee can follow on the cover of the Test Booklet. The test then continues with the following types of questions:

1. Personal Conversation.

   This section corresponds to the "warm-up" section of the direct interview. In this section, the examinee listens to conversational questions about his/her family, education, hobbies, etc. in the test language and responds to each question as it is asked. There are 10 to 13 such questions on each form. This is the only section in which the test language is used on the tape.

   For each of the following question types, the examinee is given between 15 and 60 seconds to prepare an answer before being required to speak. Time for giving an answer ranges from 45 seconds to 2 minutes.

2. Giving Directions.

The examinee is shown a pictorial map in the test booklet and is instructed to give directions between two points on the map in a realistic, contextualized situation.

3. Detailed Description.

The examinee is shown a drawing in the test booklet and is instructed to describe the picture in as much detail as possible. Each picture contains not only a variety of objects but also of actions. This question is also contextualized so that the examinee knows the specific audience being addressed and the purpose of the description.

4. Picture Sequences.

The examinee is instructed to speak in a narrative fashion about a sequence of four or five pictures shown in the test booklet. There are three questions of this type; in general, one each for past, present and future time narration. Again all questions are contextualized so that the examinee is given a specific audience and a specific reason for the narration.

5. Topical Discourse.

The examinee is instructed to talk about selected topics involving different discourse strategies. These strategies include explaining a process, supporting an opinion and talking about a hypothetical situation. There are five or six such topics, each printed in the test booklet.

6. Situations.

The examinee reads a printed description of a real-life situation in which a specified audience and communicative task are identified. The examinee is then instructed to carry out the specified task. There are five such situation questions on each form, with tasks ranging from making simple requests to giving an informal toast.

2. HEBREW SPEAKING TEST

2.1 MAJOR PROJECT ACTIVITIES

The development of the Hebrew Speaking Test (HeST) was carried out under the direction of Charles W. Stansfield, who served as Project Director, with assistance from Dorry Kenyon, who served as CAL's Test Development Coordinator for this project. The day-to-day activities were carried out at Tel Aviv University, Tel Aviv, Israel, under the supervision of Elana Shohamy, Hebrew Testing Specialist, assisted by Claire Gordon, Test Development Specialist. Dr. Shohamy remained in constant communication with CAL via electronic mail in every step of the project. In addition to Shohamy and Gordon, a local test development committee was formed in Tel Aviv which included one experienced teacher of Hebrew as a Foreign Language at the university level, Ms. Shoshana Brosh (Tel Aviv University) and Dr. Iris Geva (The Technion, Haifa), an experienced Hebrew language teaching materials developer. Ms. Laura Greenberg completed the local team as the artist for the test.

In addition to CAL staff and the local test development committee members, three leading professors of Hebrew involved with the proficiency testing movement served as external reviewers during the development of the HeST: Ruth Gollan (Brandeis University), Shmuel Bolozky (University of Massachusetts, Amherst) and Adina Ofek (Jewish Theological Seminary). These individuals provided feedback on the draft versions of the test forms.

The local test development committee met on a regular basis from November, 1988, to January, 1989, to develop the specific items for the test, based on the question types used in the semi-direct test of Chinese and Portuguese. It was decided to develop two versions of the test, each in two forms. One version (known as the USA Version) was to be intended to be used with examinees in North America, who may not have the type of cultural background knowledge an extended stay in Israel would bring, and the other version (known as the Israeli Version) for English speaking learners of Hebrew resident in Israel. The difference, then, between the two is only in the amount of background

information that is assumed of the examinee.

Of the three tests developed in year two, the HeST deviated the least from the proto-typical semi-direct format used in these tests. One reason for this is that Israeli culture is closer to the Western culture of Brazil and Portugal (as opposed to Indonesian or Hausa culture), thus necessitating fewer changes. In the personal conversation, however, there was no way to avoid the fact that in Hebrew the term "you" carries gender markings. Thus, each of the four Master Test Tapes exists in a Male and a Female version (i.e., different versions intended for a male or female examinee). The HeST retained all of the five picture items and all of the topic and situation items. Sample items may be found in the HEST Examinee Handbook, located in Appendix A-2.

2.2 TRIALING OF THE TEST FORMS

Each of the two versions of the HeST were trialed separately on examinees from the respective intended populations. The two forms of the USA Version were piloted in the Washington, DC, area and the two forms of the Israeli Version at Tel Aviv University, Israel. The subjects involved in the trialing in each case represented three different levels of proficiency: Intermediate, Advanced and Superior.

The purpose of the trialing was to ensure that the questions were clear, understandable and working as intended as well as to check the appropriateness of the pause times allotted on the tape for examinee responses. The trialing in Washington, DC, using 13 volunteer examinees from the Hebrew program of George Washington University and Hebrew language students from the Jewish Community Center in Rockville, Maryland, was conducted by Elana Shohamy with the assistance of Dorry Kenyon and Charles W. Stansfield. The subjects took the test on an individual basis using two tape-recorders. The trialing in Israel was conducted in a language lab at Tel Aviv University by Elana Shohamy and Claire Gordon with the assistance of Shoshana Brosh. Participants in the trialing in Israel were predominantly American exchange students at Tel Aviv

University.

At each location, upon completion of the test, examinees responded to a detailed questionnaire about it (Appendix A-3). When possible, they were also questioned about the test in person. In most cases the students were observed while taking the test by a Hebrew speaking member of the test development committee who took notes on students' performance and also responded to a second questionnaire about the test (Appendix A-3).

Feedback on the USA versions of the test was also provided by members of the External Review Board who listened to some of the examinee performances on the trial version.

All feedback from the trialing was summarized and the test development committee met to discuss revisions. Some modifications were necessary in the questions; in most cases this involved clarification of ambiguous items in the tasks and minor revisions in the pictures. The original pauses were adjusted--they were lengthened or shortened in various items. One important result of the trialing was the decision to prepare two versions of the final Hebrew warm-up conversation, one for female test takers and one for male test-takers, as discussed above. This decision did not alter the wording of the scripts significantly.

2.3 VALIDATION STUDY

Similar to the studies conducted for the CST and PST, a research study was designed and carried out to validate each of the four forms (in two versions) of the test. The study sought to answer the following questions:

1. Can this test, which involves spoken responses in Hebrew, be scored reliably by different raters?

2. For each version of the HeST, are the two separate forms of the test interchangeable, i.e., do they produce similar examinee results independently of the particular form administered?

3. Do the recorded responses produce the same score as a regular live interview for any given examinee?

To answer these questions, a research design was prepared involving 40 subjects. Two parallel validation studies were conducted: the two forms of the USA Version were validated in the USA on 20 university students learning Hebrew, while the validation of the two forms of the Israeli Version was conducted in Israel with 20 English speaking students studying Hebrew. Each subject was administered the two versions of the appropriate test and the Oral Proficiency Interview (OPI). The design controlled for order of administration, with half of the subjects of each form receiving form A first and form B second, and the other half in reverse order. In all cases the OPI was administered before the administration of the HeST. The design also attempted to control for proficiency level; students from three different class levels were selected for participation.

10 examinees from Brandeis University and 10 from the University of Massachusetts, Amherst, participated in the USA study. The OPI was given to subjects from the University of Massachusetts by Shmuel Bolozky, while Ruth Gollan administered the OPI to subjects from Brandeis. At the time of the study, both Bolozky and Gollan had received training from ACTFL in the oral proficiency interviewing technique, and Gollan had been certified by ACTFL in English-as-a-Second-Language. In each case the OPI was recorded to be scored at a later date. The HeST was administered locally, either individually or in a language lab.

The Israeli study involved 20 American undergraduate students who had completed a year or more of Hebrew study at their respective home universities before coming to study at Tel Aviv University. The OPI in Israel was administered by Elana Shohamy, who has had experience in administering Oral Interviews in a variety of settings. Again, the interviews were tape recorded, to be rated at a later date by the two Israeli raters. Some of the

subjects took the HeST in the language lab and others on an individual basis.

Four raters, two in the USA and two in Israel, were used in the design. Ruth Gollan and Shmuel Bolozky were the raters for the USA study and Miriam Shachar and Ziona Snir were the raters in Israel. Both Ms. Shachar and Ms. Snir are experienced Hebrew teachers and received intensive training in using the ACTFL guidelines prior to the rating of the speech samples of the validation study. In both studies, the ratings of all the tapes were done independently and in random order, and subjects were rated anonymously; however, raters scored all the oral interviews before proceeding to rate the HeST tapes. After all the ratings were completed, subjects were sent their test results in the mail: the scores of the two raters on the live interview and on each of the HeST versions.

To proceed with the empirical analysis of the ratings, scores on both the live interview and the tape-based semi-direct tests converted to a scale combining both ACTFL and ILR rating scales with weights assigned as follows:

ACTFL/ILR Level	Coded as:
Novice-Low	0.2
Novice-Mid	0.5
Novice-High	0.8
Intermediate-Low	1.0
Intermediate-Mid	1.5
Intermediate-High	1.8
Advanced	2.0
Advanced-Plus	2.8
Superior/Level 3	3.0
"High Superior"/Level 3+-5	3.8

This system of score coding is based on the ILR 0 to 5 rating scale and is intended to assign an appropriate numerical value to the proficiency level descriptions. For example, proficiency at an Advanced-Plus level is characterized by many of the same features as at the Superior/3 level, though the examinee cannot sustain the performance. Thus, the numerical interpretation falls closer to 3.0 than mid-way between the two, as may be expected.

The several tables below provide descriptive statistics, interrater reliabilities and parallel-form reliability data obtained in the two studies. The USA study results will be presented first, followed by the results of the Israeli study. Note: UV refers to the USA Version and IV refers to Israeli Version, A refers to Form A and B refers to Form B; Rater 1 (USA) is Bolozky, Rater 2 (USA) is Gollan, Rater 1 (Israel) is Shachar, and Rater 2 (Israel) is Snir.

2.4 USA Study

Table 2.1 shows the mean score, standard deviation and other basic statistics for the ratings assigned by each of the two raters to subject performances on each of the semi-direct test forms and on the live interview.

Table 2.1
Descriptive Statistics for Scoring Levels Assigned
Tape and Live Tests (USA Study)

Test Form	Minimum Score	Maximum Score	Mean	Standard Deviation
UVA (n=20)				
Rater 1	1.0	3.0	2.01	0.74
Rater 2	0.8	3.0	1.90	0.65
UVB (n=20)				
Rater 1	0.8	3.0	2.00	0.74
Rater 2	0.8	3.0	1.84*	0.62
USA-Interview (n=20)				
Rater 1	1.0	3.0	2.00	0.69
Rater 2	1.0	3.0	2.03	0.64

* The difference in these paired means was significant at the $p<.05$ level.

Table 2.2 shows the frequency of ratings given by the two raters on the live interviews (n=20) and on both forms (n=40).

Table 2.2
Frequency Distributions

USA Live Interview Ratings (Rater 1)

Rating	Frequency	Percent
1	3	15.0
1.5	4	20.0
1.8	2	10.0
2	5	25.0
2.8	3	15.0
3	3	15.0

USA Live Interview Ratings (Rater 2)

Rating	Frequency	Percent
1	1	5.0
1.5	7	35.0
1.8	2	10.0
2	4	20.0
2.8	3	15.0
3	3	15.0

USA-Rating of HeST Test Forms (Rater 1)

Rating	Frequency	Percent
0.8	1	2.5
1	5	12.5
1.5	10	25.0
1.8	7	17.5
2	3	7.5
2.8	7	17.5
3	7	17.5

USA-Rating of HeST Test Forms (Rater 2)

Rating	Frequency	Percent
0.8	2	5.0
1	3	7.5
1.5	12	30.0
1.8	9	22.5
2	5	12.5
2.8	6	15.0
3	3	7.5

These statistics indicate that each form of the test was taken by a group of examinees that varied widely in oral language proficiency. Similarly, the ratings assigned by both raters reflected this range. The mean scores for each rater were very similar, indicating that the raters were almost equal in their degree of severity. However, Rater 2 was slightly more severe than Rater 1 in rating Form UVB of the HeST, although this difference was minuscule, i.e., only .16 of a level on the ILR scale. Using a t-test for the difference between paired means, this difference was statistically significant ($p<.05$).

The degree of agreement between the absolute ratings may be seen from the following three cross-tab diagrams. First, Table 2.3 presents the ratings of Rater 1 (down) against the ratings of Rater 2 (across) for the live interview.

Table 2.3

USA - Crosstabulations of Live Interview Ratings (n=20)

Rater 1 (down)/ / Rater 2 (across) Frequency	1	1.5	1.8	2	2.8	3	Total
1	1	2	0	0	0	0	3
1.5	0	4	0	0	0	0	4
1.8	0	1	1	0	0	0	2
2	0	0	1	4	0	0	5
2.8	0	0	0	0	3	0	3
3	0	0	0	0	0	3	3
Total	1	7	2	4	3	3	20

For the live interview, there was total agreement in 80% of the ratings. Of the four cases of disagreement, Rater 1 was more generous in 2 cases and Rater 2 was more generous in the other 2. There were no cases in which disagreement in the rating was more than one step away on the rating scale, and none of them crossed ACTFL level boundaries.

Table 2.4 presents the ratings of Rater 1 (down) against Rater

2 (across) for HeST Form UVA.

Table 2.4

Crosstabulations of HeST Form UVA Ratings (20 ratings)

Rater 1 (down) / / Rater 2 (across) Frequency	0.8	1	1.5	1.8	2	2.8	3	Total
0.8	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	3
1.5	0	0	5	0	0	0	0	5
1.8	0	0	1	3	0	0	0	4
2	0	0	0	1	0	0	0	1
2.8	0	0	0	0	2	1	0	3
3	0	0	0	0	0	2	2	4
Total	1	1	7	4	2	3	2	20

From Table 2.4 we see that the agreement of the absolute ratings was again quite high. There was total agreement in 60% of the 20 HeST Form UVA ratings. In six of the eight cases of disagreement, Rater 1 was more generous while Rater 2 was more generous in only two cases. For none of the ratings was the disagreement more than one step away on the rating scale. Only three of the disagreements crossed ACTFL level boundaries.

Table 2.5 presents the ratings of Rater 1 (down) against Rater 2 (across) for HeST Form UVB.

Table 2.5

Crosstabulations of HeST Form UVB Ratings (20 ratings)

Rater 1 (down) / Rater 2 (across)

Frequency	0.8	1	1.5	1.8	2	2.8	3	Total
0.8	1	0	0	0	0	0	0	1
1	0	2	0	0	0	0	0	2
1.5	0	0	4	1	0	0	0	5
1.8	0	0	1	2	0	0	0	3
2	0	0	0	2	0	0	0	2
2.8	0	0	0	0	3	1	0	4
3	0	0	0	0	0	2	1	3
Total	1	2	5	5	3	3	1	20

From Table 2.5 we see that the agreement of the absolute ratings was again relatively high. There was total agreement in 55% of the 20 HeST Form UVB ratings. Rater 2 was more generous in eight of the nine cases of disagreement; Rater 1 was more generous in only one case. Again, for none of the ratings was the disagreement more than one step away on the rating scale. There were only four cases of disagreements crossing ACTFL level boundary lines.

The tables above show that Rater 1 tended to be slightly more generous than Rater 2, especially on the taped forms in the difference between Advanced (2.0) and Advanced-Plus (2.8), where there were 5 instances on the 40 tapes where Rater 1 awarded an examinee a 2.8 while Rater 2 gave the examinee a 2.0.

Interrater reliabilities (Pearson product-moment correlations) between the ratings assigned by Rater 1 and those assigned by Rater 2 for the two semi-direct test forms and for the live interview are shown in Table 2.6 below.

Table 2.6
Interrater Reliabilities

Test Form	Correlation
UVA (n=20)	.92
UVB (n=20)	.92
Interview (n=20)	.97

These interrater reliabilities are all uniformly high across the two test forms and the live interview. Interrater reliability was not adversely affected by the semi-direct test format. This suggests that the HeST elicits a sample of speech as ratable as the live interview.

On performance-based tests such as the HeST, there is an increased concern for test-retest reliability. This form of reliability measures the degree of inconsistency in examinee performance on two separate administrations of the same test. The amount of inconsistency reflects the degree to which the test score may be confounded by such inconsistency. Therefore, it is important to examine this factor. However, on a test with a limited number of questions such as the HeST, it is not wise to administer the same test twice, since the first sitting will serve to instruct the examinee in the task at hand. (For a thorough discussion of this "reactivity effect," see Stansfield and Ross, 1988, p. 174.) Under such circumstances, it is preferable to administer different forms of the test while still using the same rater to score the performance. This type of reliability is known as parallel-form reliability, which is the degree of correlation between scores on two forms of the test.

Parallel-form reliabilities for the same subject taking two different test forms, with the same rater scoring both forms, are shown in Table 2.7.

Table 2.7
Parallel-Form Reliabilities (Same Rater)

	Rater 1	Rater 2
Forms UVA and UVB (n=20)	.99	.93

The statistics indicate that the parallel form reliability of the HeST is very high. With the first rater, the parallel-form reliability was nearly perfect (.99). With a different rater, Rater 2, the parallel-form reliability was also very high (.94). Such favorable statistics provide strong support for the proposition that each form of the HeST elicits a sample of speech that is uniformly challenging to the examinee. The fact that the parallel-form reliability was high for two different raters supports the claim that the sample of speech elicited by different forms is equally ratable.

In summary, the evidence from Table 2.7 warrants the conclusion that natural variations in examinee oral language performance are adequately controlled for by the HeST format.

Table 2.8 shows parallel-form reliabilities for subjects taking two different test forms, with each form scored by a different rater.

Table 2.8
Parallel Form Reliabilities (Different Forms and Raters)

Rater/Form Combination	Correlation
Rater 1/Form UVA - Rater 2/Form UVB (n=20)	.92
Rater 1/Form UVB - Rater 2/Form UVA (n=20)	.92

This type of parallel-form reliability involves error that can be attributed to natural variation in examinee speech, error that can be attributed to differences in test form, and error that can be attributed to differences in raters. Thus, it may be viewed as a lower-bound estimate of the reliability of an HeST score. Again

the reliabilities here are high, even under these severe conditions (different forms and different raters).

Correlations of semi-direct test scores with the live face-to-face interview are given in Table 2.9 below. These correlations are evidence of the validity of the HeST as a surrogate live interview.

Table 2.9
Correlations with Live Interview

Rater/Form	Rater 1/Interview	Rater 2/Interview
Rater 1/Form UVA (n=20)	.96	.94
Rater 1/Form UVB (n=20)	.96	.94
Rater 2/Form UVA (n=20)	.93	.94
Rater 2/Form UVB (n=20)	.92	.90
All Matched Interviews/Forms (80 pairs)	.93	

Again, the correlations are all high. The average correlation based on 80 pairs of ratings (20 subjects x 2 HeST forms x 2 ratings, correlated with the score assigned for the live interview) was .93. Such results support the claim that the HeST is a valid measure of oral language proficiency that can be substituted for a live interview.

The degree of agreement in absolute ratings given on the live interview with ratings given on the same examinee's HeST may be seen from the following cross-tab diagram. In Table 2.10 all 80 pairs of interview ratings (down) with HeST ratings (across) are presented.

Table 2.10

Crosstabulations of Interview ratings by HeST ratings

Interview (down) / / HeST (across) Frequency	0.8	1	1.5	1.8	2	2.8	3	Total
0.8	0	0	0	0	0	0	0	0
1	3	5	0	0	0	0	0	8
1.5	0	3	18	1	0	0	0	22
1.8	0	0	3	5	0	0	0	8
2	0	0	1	10	5	2	0	18
2.8	0	0	0	0	3	8	1	12
3	0	0	0	0	0	3	9	12
Total	3	8	22	16	8	13	10	80

From the table we see that in 62.5% of the cases there was an absolute agreement between the two ratings. In only one case, in which an examinee received a 1.5 on the HeST and a 2.0 on the interview, was the disagreement in the rating more than one step away on the rating scale. For all of the remaining ratings, the disagreement was only one away step on the scale. In 4 cases (13% of the disagreements), the score on the HeST was above that awarded on the interview. In 26 cases (87% of the disagreements), the score on the interview was the higher of the two. Thus, besides the high correlations documented above, the absolute values given to examinees on both the live interview and the HeST were extremely close. In only 18 cases (22.5%), did the disagreement cross an ACTFL level boundary, with 10 of these cases involving the awarding of a 1.8 on the HeST and a 2.0 on the interview.

As a general summary of the statistical information above, it may be stated that both forms of the semi-direct test reveal high interrater reliabilities, with Pearson product-moment correlations at .92. Parallel form reliabilities are also very high, even under the most "severe" conditions (i.e., different raters rating two different forms), where correlations are at .92; with the same rater, correlations range from .93 to .99. The correlations with

the live interview are also very high; with the same rater they range from .90 to .96 and with different raters from .92 to .94.

2.5 ISRAELI STUDY

Table 2.11 shows the mean score, standard deviation and other basic statistics for the ratings assigned by each of the two raters to subject performances on each of the semi-direct test forms and on the live interview.

Table 2.11
Descriptive Statistics for Scoring Levels Assigned Tape and Live Tests

Test Form	Minimum Score	Maximum Score	Mean	Standard Deviation
IVA (n=20)				
Rater 1	0.8	3.8	2.26	0.86
Rater 2	0.8	3.8	2.36	1.00
IVB (n=20)				
Rater 1	0.8	3.8	2.28	0.85
Rater 2	0.8	3.8	2.35	0.87
Israel-Interview (n=20)				
Rater 1	0.5	3.8	2.19	0.90
Rater 2	0.5	3.8	2.18	0.88

None of the differences in these paired means was significant at the $p<.05$ level.

Table 2.12 shows the frequency of ratings given by the two raters on the live interviews (n=20) and on both HeST forms (n=40).

Table 2.12
Frequency Distributions

Israeli Live Interview Ratings (Rater 1)

Rating	Frequency	Percent
0.5	1	5.0
0.8	1	5.0
1	1	5.0
1.5	2	10.0
1.8	5	25.0
2	1	5.0
2.8	2	10.0
3	6	30.0
3.8	1	5.0

Israeli Live Interview Ratings (Rater 2)

Rating	Frequency	Percent
0.5	1	5.0
1	2	10.0
1.5	3	15.0
1.8	3	15.0
2	2	10.0
2.8	3	15.0
3	5	25.0
3.8	1	5.0

Israel-Rating of HeST Test Forms (Rater 1)

Rating	Frequency	Percent
0.8	2	5.0
1	4	10.0
1.5	4	10.0
1.8	6	15.0
2	4	10.0
2.8	6	15.0
3	12	30.0
3.8	2	5.0

Israel-Rating of HeST Test Forms (Rater 2)

Rating	Frequency	Percent
0.8	2	5.0
1	3	7.5
1.5	5	12.5
1.8	5	12.5
2	6	15.0
2.8	4	10.0
3	9	22.5
3.8	6	15.0

As in the USA study, these statistics indicate that each form of the test was taken by a group of examinees that varied widely in oral language proficiency. As could be expected, average performance of this sample of native English speaking learners of Hebrew studying in the target language country was higher than that of the subjects sampled in the USA. In fact, it was necessary to use the rating of 3.8 ("High-Superior") in the Israeli study to distinguish those examinees who were clearly above an ILR level 3 (ACTFL Superior) from those who were at that level. The mean scores for each rater were very similar, indicating that the raters were almost equal in their degree of severity.

The degree of agreement between the absolute ratings may be seen from the following three cross-tab diagrams. First, Table 2.13 presents the ratings of Rater 1 (down) against the ratings of Rater 2 (across) for the live interview.

Table 2.13
Israel-Crosstabulations of Live Interview Ratings (n=20)

Rater 1 (down)// / Rater 2 (across)

Frequency	0.5	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.5	1	0	0	0	0	0	0	0	0	1
0.8	0	0	1	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	0	1
1.5	0	0	0	2	0	0	0	0	0	2
1.8	0	0	0	1	3	1	0	0	0	5
2	0	0	0	0	0	1	0	0	0	1
2.8	0	0	0	0	0	0	2	0	0	2
3	0	0	0	0	0	0	1	5	0	6
3.8	0	0	0	0	0	0	0	0	1	1
Total	1	0	2	3	3	2	3	5	1	20

For the live interview, there was total agreement in 85% of the ratings. Of the three cases of disagreement, Rater 1 was more generous than Rater 2; Rater 2 was more generous in one case. There were no cases in which disagreement in the rating was more

than one step away on the rating scale. Two of the three disagreements crossed ACTFL level boundaries.

Table 2.14 presents the ratings of Rater 1 (down) against Rater 2 (across) for HeST Form IVA.

Table 2.14
Israel-Crosstabulations of HeST Form IVA Ratings (20 ratings)

Rater 1 (down) / / Rater 2 (across)

Frequency	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.8	1	0	0	0	0	0	0	0	1
1	0	2	0	0	0	0	0	0	2
1.5	0	0	1	1	0	0	0	0	2
1.8	0	0	1	2	1	0	0	0	4
2	0	0	0	0	1	0	0	0	1
2.8	0	0	0	0	1	0	2	0	3
3	0	0	0	0	0	1	2	3	6
3.8	0	0	0	0	0	0	0	1	1
Total	1	2	2	3	3	1	4	4	20

From Table 2.14 we see that the agreement of the absolute ratings was again relatively high. There was total agreement in 50% of the 20 HeST Form IVA ratings. Of the 10 cases of disagreement, Rater 2 was more generous in 7; Rater 1 in 3. As in the USA study, in none of the ratings was the disagreement more than one step away on the rating scale. There was only one case in which the disagreement crossed an ACTFL level boundary.

Table 2.15 presents the ratings of Rater 1 (down) against Rater 2 (across) for HeST Form IVB.

Table 2.15
Israel - Crosstabulations of HeST Form IVB Ratings (20 ratings)

Rater 1 (down) / / Rater 2 (across)

Frequency	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.8	1	0	0	0	0	0	0	0	1
1	0	1	1	0	0	0	0	0	2
1.5	0	0	2	0	0	0	0	0	2
1.8	0	0	0	1	1	0	0	0	2
2	0	0	0	1	2	0	0	0	3
2.8	0	0	0	0	0	3	0	0	3
3	0	0	0	0	0	0	5	1	6
3.8	0	0	0	0	0	0	0	1	1
Total	1	1	3	2	3	3	5	2	20

From Table 2.15 we see that the agreement of the absolute ratings was again high. There was total agreement in 80% of the 20 HeST Form UVB ratings. In three of the four cases of disagreement, Rater 2 was the more generous with Rater 1 being more generous in only one case. Again, in none of the ratings was the disagreement more than one step away on the rating scale. Two of the four disagreements crossed ACTFL level boundaries.

The tables above show that Rater 2 tended to be more slightly more generous than Rater 1. This is especially apparent in three instances on Form IVA, where Rater 2 awarded three examinees a 3.8, while Rater 1 awarded them a 3.0.

Interrater reliabilities (Pearson product-moment correlations) between the ratings assigned by Rater 1 and those assigned by Rater 2 for the two semi-direct test forms and for the live interview are shown in Table 2.16 below.

Table 2.16
Interrater Reliabilities

Test Form	Correlation
IVA (n=20)	.93
IVB (n=20)	.97
Interview (n=20)	.99

Again, as in the USA study, the interrater reliabilities are all uniformly high across the test forms and the live interview. Interrater reliability was not adversely affected by the semi-direct test format. This again suggests that the HeST elicits a sample of speech as ratable as the live interview.

Parallel-form reliabilities for the same subject taking two different HeST forms, with the same Israeli rater scoring both forms, are shown in Table 2.17.

Table 2.17
Parallel-Form Reliabilities (Same Rater)

	Rater 1	Rater 2
Forms IVA and IVB (n=20)	.94	.94

The statistics indicate that the parallel form reliability of the HeST is very high, being .94 for each rater. Such a high correlation provides strong support for the proposition that each form of the HeST elicits a sample of speech that is uniformly challenging to the examinee. The fact that the parallel-form reliability was high for two different raters supports the claim that the sample of speech elicited by different forms is equally ratable.

In summary, the evidence from Table 2.17 warrants the conclusion that natural variations in examinee oral language performance are adequately controlled for by the HeST format.

Table 2.18 shows parallel-form reliabilities for subjects

taking two different test forms, with each form scored by a different rater.

Table 2.18
Parallel Form Reliabilities (Different Forms and Raters)

Rater/Form Combination	Correlation
Rater 1/Form IVA - Rater 2/Form IVB (n=20)	.91
Rater 1/Form IVB - Rater 2/Form IVA (n=20)	.94

This type of parallel-form reliability involves error that can be attributed to natural variation in examinee speech, error that can be attributed to differences in test form, and error that can be attributed to differences in raters. Thus, it may be viewed as a lower-bound estimate of the reliability of an HeST score. Again the reliabilities here are high, even under these severe conditions (different forms and different raters).

Correlations of semi-direct test scores with the live face-to-face interview are given in Table 2.19 below. These correlations are evidence of the validity of the HeST as a surrogate live interview.

Table 2.19
Correlations with Live Interview

Rater/Form	Rater 1/Interview	Rater 2/Interview
Rater 1/Form IVA (n=20)	.95	.95
Rater 1/Form IVB (n=20)	.91	.91
Rater 2/Form IVA (n=20)	.84	.84
Rater 2/Form IVB (n=20)	.87	.88
All Matched Interviews/Forms (80 pairs)	.89	

Again, the correlations are all relatively high. The average correlation based on 80 pairs of ratings (20 subjects x 2 HeST forms x 2 ratings, correlated with the score assigned for the live

interview) was .89. As in the USA study, these results support the claim that the HeST is a valid measure of oral language proficiency that can be substituted for a live interview.

The degree of agreement in absolute ratings given on the live interview with ratings given on the same examinee's HeST may be seen from the following cross-tab diagram. In Table 2.20 all 80 pairs of interview ratings (down) with HeST ratings (across) are presented.

Table 2.20
Israel-Crosstabulations of Interview ratings by HeST ratings

Interview (down) / / HeST (across) Frequency	0.5	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.5	0	4	0	0	0	0	0	0	0	4
0.8	0	0	2	0	0	0	0	0	0	2
1	0	0	5	1	0	0	0	0	0	6
1.5	0	0	0	5	4	1	0	0	0	10
1.8	0	0	0	3	7	6	0	0	0	16
2	0	0	0	0	0	2	0	3	1	6
2.8	0	0	0	0	0	0	3	5	2	10
3	0	0	0	0	0	1	7	11	3	22
3.8	0	0	0	0	0	0	0	2	2	4
Total	0	4	7	9	11	10	10	21	8	80

From the table we see that in 43.75% of the cases there was an absolute agreement between the ratings awarded by either rater on the live interviews and on the taped tests. In another 43.75% of the cases, the difference was only one away step on the rating scale. In 12.5% of the cases, there was a more serious disagreement between the ratings awarded an examinee for the live interview and that awarded the examinee's performance on one of the taped test forms. In one case, an examinee was awarded a 1.5 on the interview, but a 2.0 on the tape. In three cases, examinees received a 2 on the live interview but a 3 on the tapes; this was

reversed in one case. In two other cases, examinees on the tape were awarded a 3.8, whereas on the live interview they received a 2.8. Lastly, the worst case of disagreement was that in one case an individual was awarded a 2.0 on the live interview and a 3.8 on the taped test. An examination of this last case reveals that the raters were in agreement on the subject's live interview performance (2.0), Form IVB performance (3.0), and one awarded the subject a 3.0 on Form IVA while the other awarded the subject a 3.8. Unlike for the vast majority of the HeST examinees, it appears that this particular individual did indeed perform quite differently on the two different test formats.

Where there was a disagreement in ratings, in 32 cases (71% of the disagreements), a higher score was awarded on the HeST than on the interview; in 13 cases (29% of the disagreements), this order was reversed. This is in contrast to the USA study, in which the majority of cases received a higher score on the interview when there was a disagreement. In 19 cases (23.75% of the total) did the disagreements cross ACTFL level boundaries. This compares to 18 cases in the USA study.

As a general summary of the statistical information above, it may be stated that both forms of the semi-direct test reveal high interrater reliabilities, with Pearson product-moment correlations at .93 and .97. Parallel form reliabilities are also very high, even under the most "severe" conditions (i.e., different raters rating two different forms), where correlations were between .91 and .94; with the same rater, correlations were .94. The correlations with the live interview are also rather high; with either the same or different raters they range from .84 to .95.

2.6 SUBJECT RESPONSE TO THE TEST

In both the USA and Israeli studies, feedback information from the participants on various aspects of their experience with and opinions about both types of testing procedures were elicited by means of a short questionnaire (Appendix A-4). The questionnaire was given to the subjects directly after they completed the semi-

direct tests. In most cases they completed and returned it before leaving the testing room. In a few cases the questionnaires were returned at a later date. All subjects in each study completed the questionnaire for a 100% participation rate.

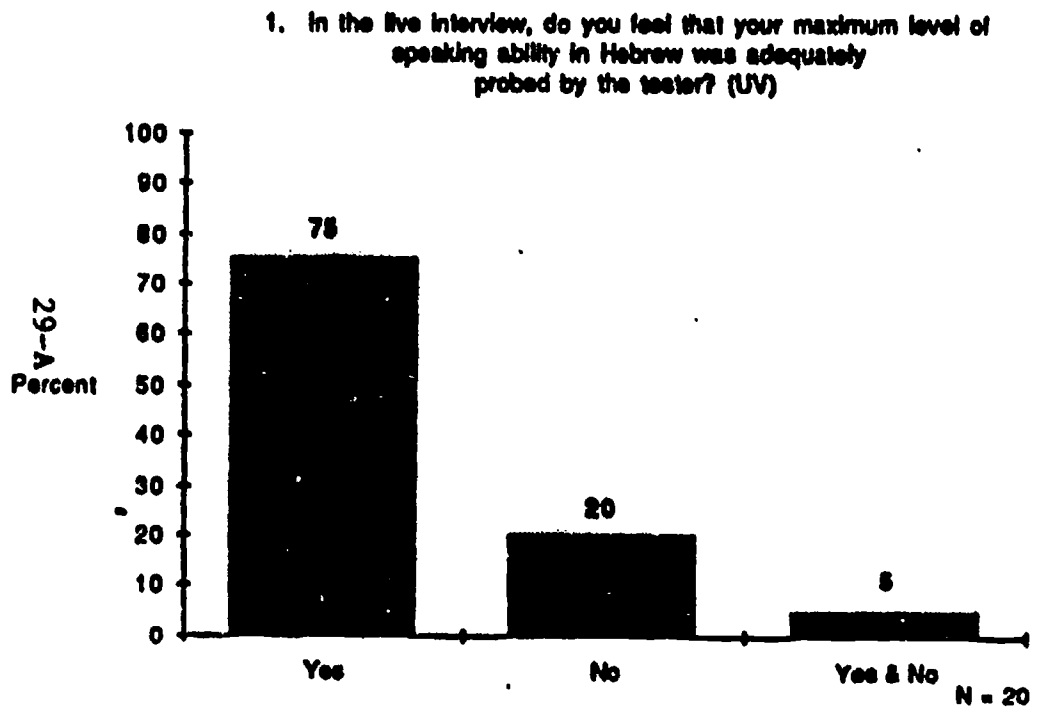
The answers to the examinee questionnaires are given in graphic summary form below. Written comments in response to the questionnaire are presented in Appendix A-5.

The first two questions sought to elicit from the subjects the extent to which they felt their Hebrew speaking ability had been probed by the two types of test: the live interview and the HeST.

(1) Over the course of the live interview, do you feel that your maximum level of speaking ability in Hebrew was adequately probed by the tester?

(2) Over the course of the taped test, do you feel that the descriptions, narratives, situations, and other types of questions in the test were adequate to probe your maximum level of speaking ability in Hebrew?

Figures 2.1 and 2.2 reveal that exactly the same high percentage of students responded positively to both questions in the Israeli study. In the USA study, the same high percentage responded positively to question 1 while for the taped test, a slightly lower percentage responded positively. This suggests that examinees in both countries held similar attitudes towards their testers and for the most part felt their speaking ability was being adequately tested by both test formats, i.e., there was little felt difference in the ability of the two test formats to test the depth and thoroughness of their present Hebrew speaking ability.


1. In the live interview, do you feel that your maximum level of speaking ability in Hebrew was adequately probed by the tester? (UV)
29-A
Percent
100
90
80
70
60
50
40
30
20
10
0
75
20
5
Yes
No
Yes & No
N = 20

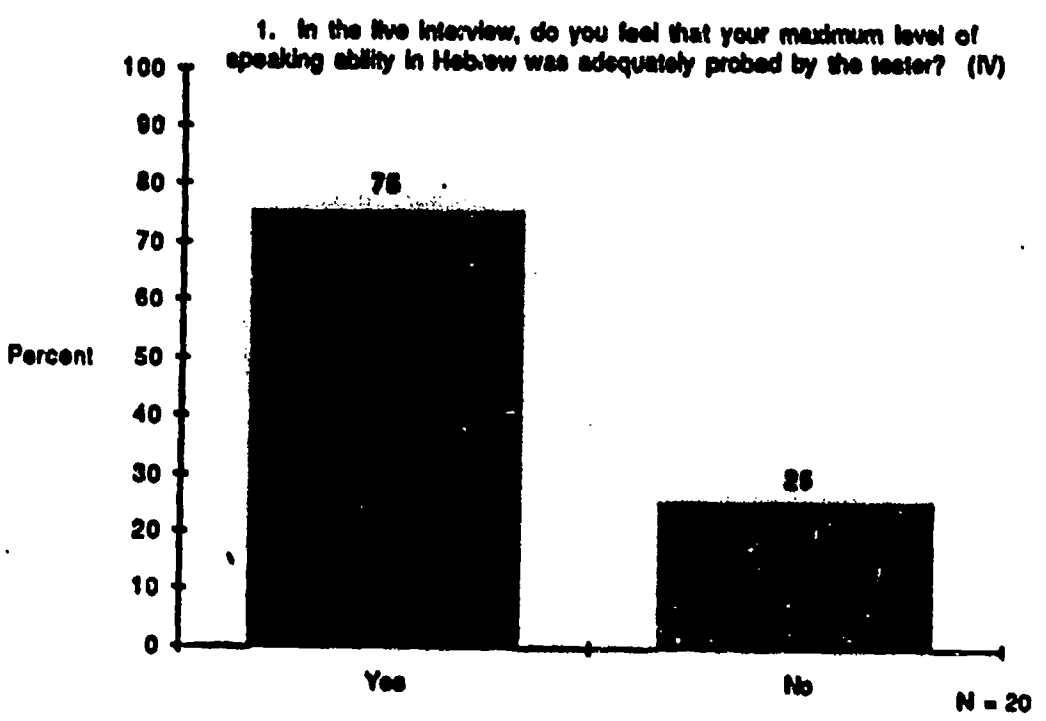
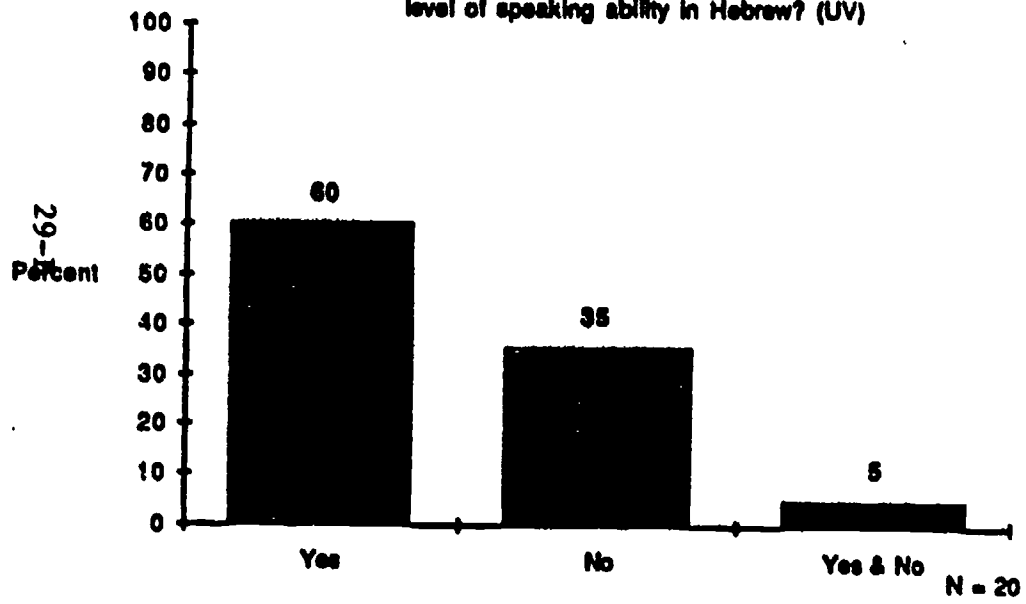

1. In the live interview, do you feel that your maximum level of speaking ability in Hebrew was adequately probed by the tester? (IV)
Percent
100
90
80
70
60
50
40
30
20
10
0
75
25
Yes
No
N = 20

Figure 2.1


2. During the taped test, did you feel that the descriptions, narratives, situations and other types of questions in the test were adequate to probe your maximum level of speaking ability in Hebrew? (UV)
Percent
100
90
80
70
60
50
40
30
20
10
0
60
35
5
Yes
No
Yes & No
N = 20

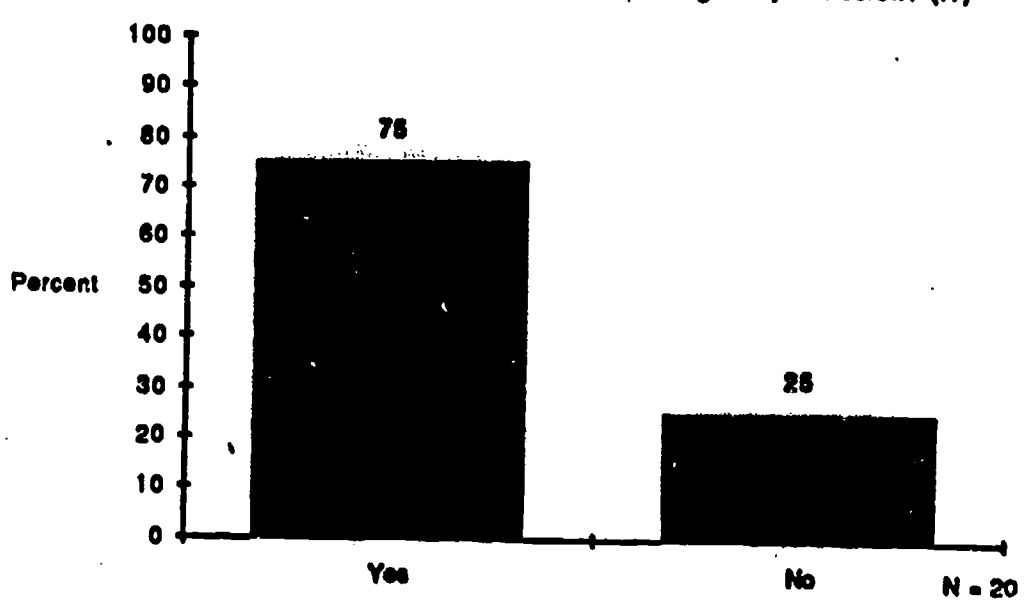

2. During the taped test, did you feel that the descriptions, narratives, situations and other types of questions in the test were adequate to probe your maximum level of speaking ability in Hebrew? (IV)
Percent
100
90
80
70
60
50
40
30
20
10
0
75
25
Yes
No
N = 20

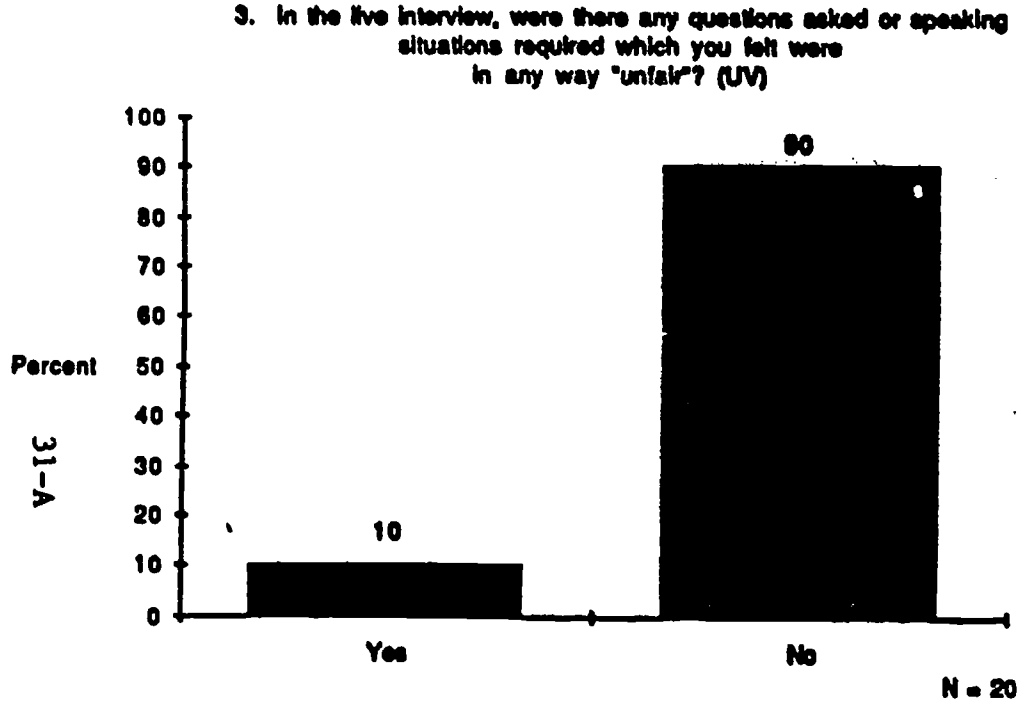
Figure 2.2

The next two questions focused on whether the subjects perceived any unfair questions on either test format.

(3) In the <u>live interview</u>, were there any questions asked or speaking situations required which you felt were in any way 'unfair'?

(4) In the <u>taped tests</u>, were there any picture/descriptions, narratives, situations, or other questions that you felt were in any way 'unfair'?

As shown in figure 2.3, in both studies a small minority (2 individuals in each study) felt there were unfair questions in the live interview, while a few more (3 in Israel and 4 in the USA) felt there were unfair questions on the HeST, as shown in Figure 2.4. This small number is impressive for a taped test which cannot adapt itself to the level or circumstances surrounding the testing of a particular examinees. In the taped test, the examinee is asked every question, whether it is too difficult or not. In any case, only a very low percentage of subjects felt there were 'unfair' questions on the taped test, and the differences between testing with the live interviewer or with the tape were minor.


3. In the live interview, were there any questions asked or speaking situations required which you felt were in any way "unfair"? (UV)
Percent
100
90
80
70
60
50
40
30
20
10
0
10
90
Yes
No
N = 20

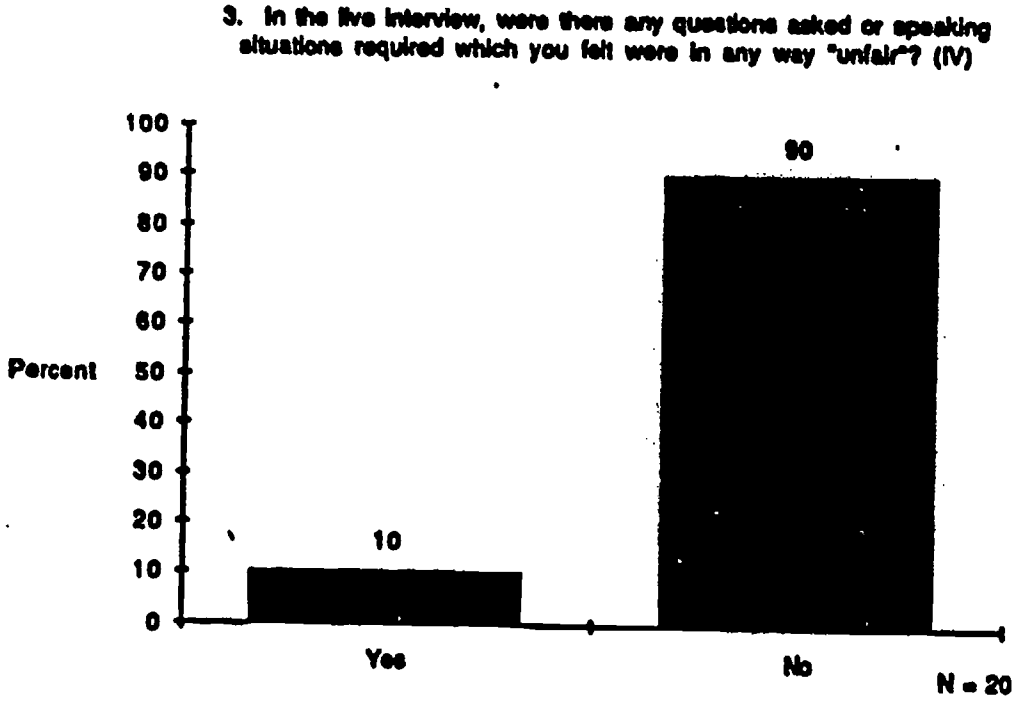
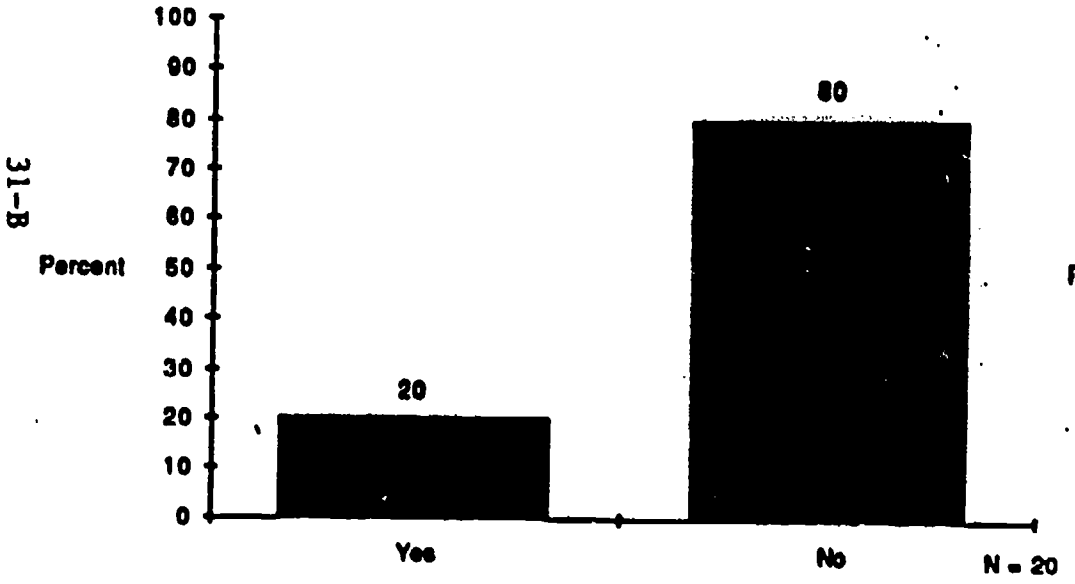

3. In the live interview, were there any questions asked or speaking situations required which you felt were in any way "unfair"? (IV)
Percent
100
90
80
70
60
50
40
30
20
10
0
10
90
Yes
No
N = 20

Figure 2.3

31-B



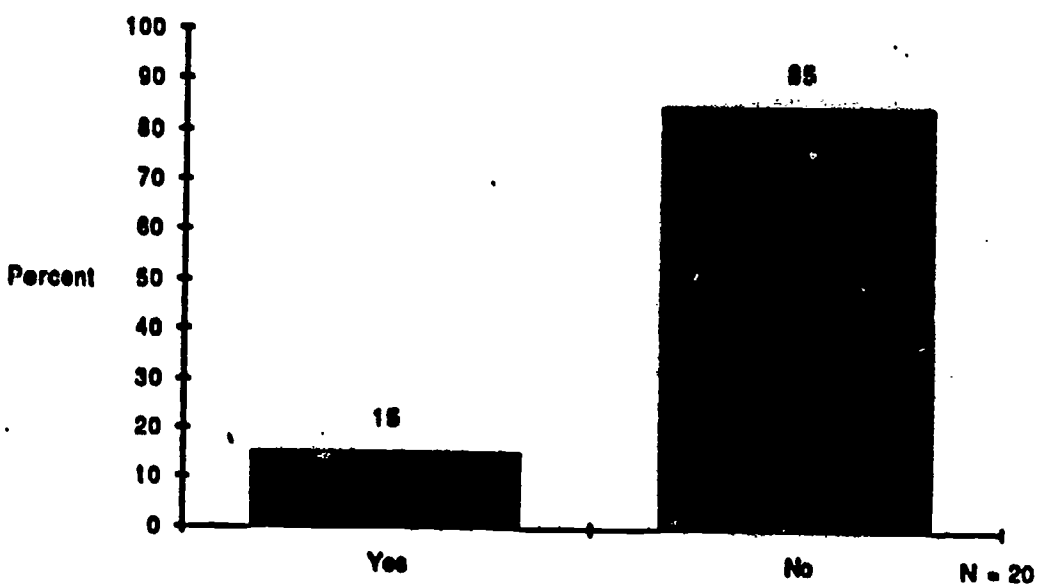


Figure 2.4

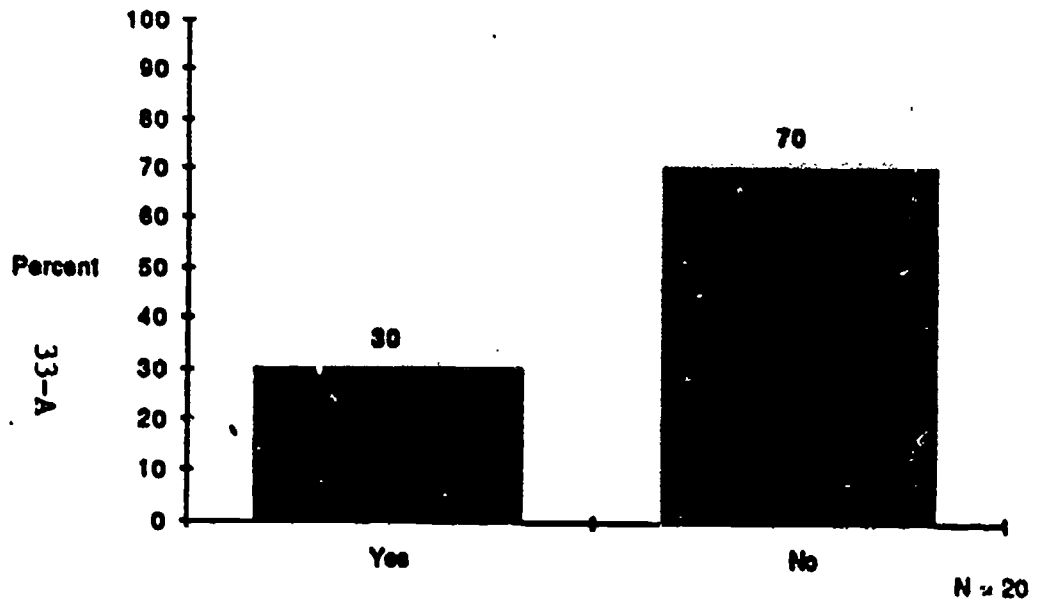
The next two questions focused on the subject's affective perceptions of the test.

(5) (A) Did you feel unduly nervous in the live interview?
(B) Did you feel unduly nervous in the taped tests?
(C) If you answered yes to both questions, in which of the two types of test did you feel more anxious or nervous?

(6) Which of the two types of tests (live interview or taped test) did you feel was more difficult?

Because the semi-direct mode of testing may be unfamiliar and perhaps 'unnatural' to students in general, it would not be unusual for a large percentage of the students in this study to feel more nervous in the taped test than in the live interview. In the USA study, 6 subjects answered they felt unduly nervous in the live interview, while 10 answered affirmatively for the taped test. However, of the five who answered yes to both questions (25% of the entire group), only one felt more nervous taking the taped test, while two felt more nervous taking the live interview (see Figures 2.5A, 2.5B and 2.5C). Perhaps because of their in country exposure to oral Hebrew, the subjects in the Israeli study were less nervous overall. Three reported nervousness in the live interview, while 5 reported feeling unduly nervous during the taped test. Interestingly, of the two who answered yes to both questions, neither was more anxious or nervous taking the taped test.

Question 6 focused on perceived difficulty. Despite the fact that subjects did approximately the same on both tests (see correlations above), a majority (70% in the USA study and 60% in the Israeli study, see Figure 2.6) of the subjects perceived the taped test as more difficult. Perhaps some of the individual comments are enlightening (see Appendix A-5); these seem to revolve around the timed pauses and discomfort in talking to a machine. It appears the 'unnatural' format contributed heavily to perceived difficulty.


5A. Did you feel unduly nervous in the live interview? (UV)
Percent
100
90
80
70
60
50
40
30
20
10
0
30
70
Yes
No
N = 20
33-A

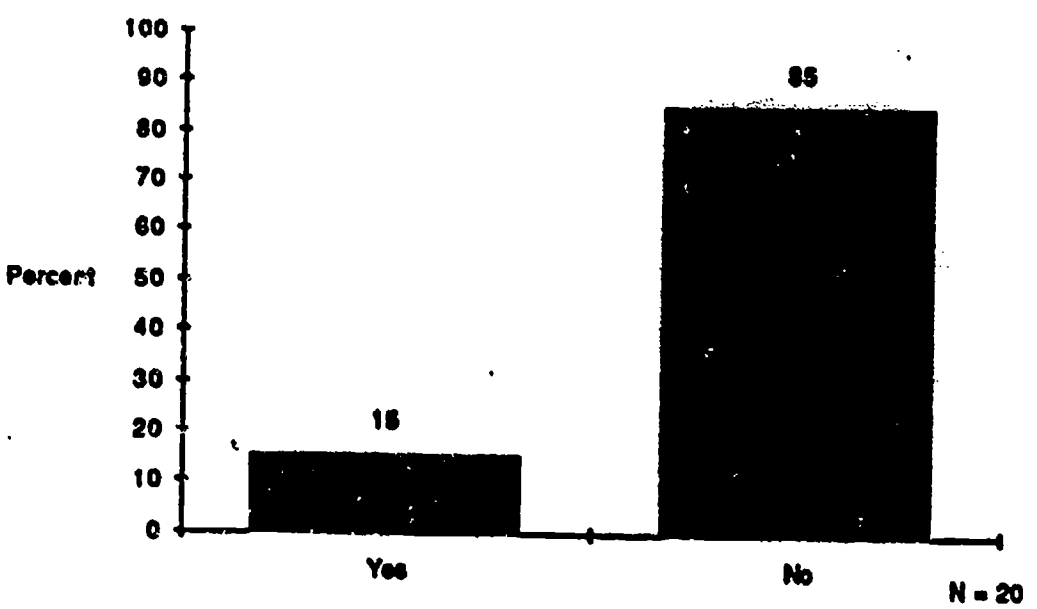
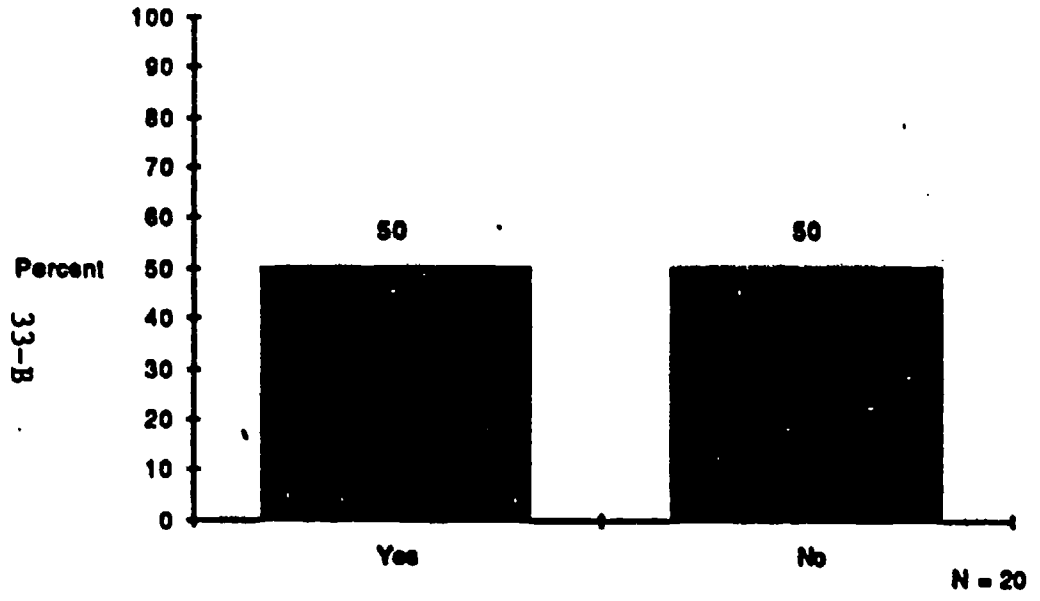

5A. Did you feel unduly nervous in the live interview? (IV)
Percent
100
90
80
70
60
50
40
30
20
10
0
15
85
Yes
No
N = 20

Figure 2.5A

5B. Did you feel unduly nervous in the taped tests? (UV)



33-B

5B. Did you feel unduly nervous in the taped tests? (IV)

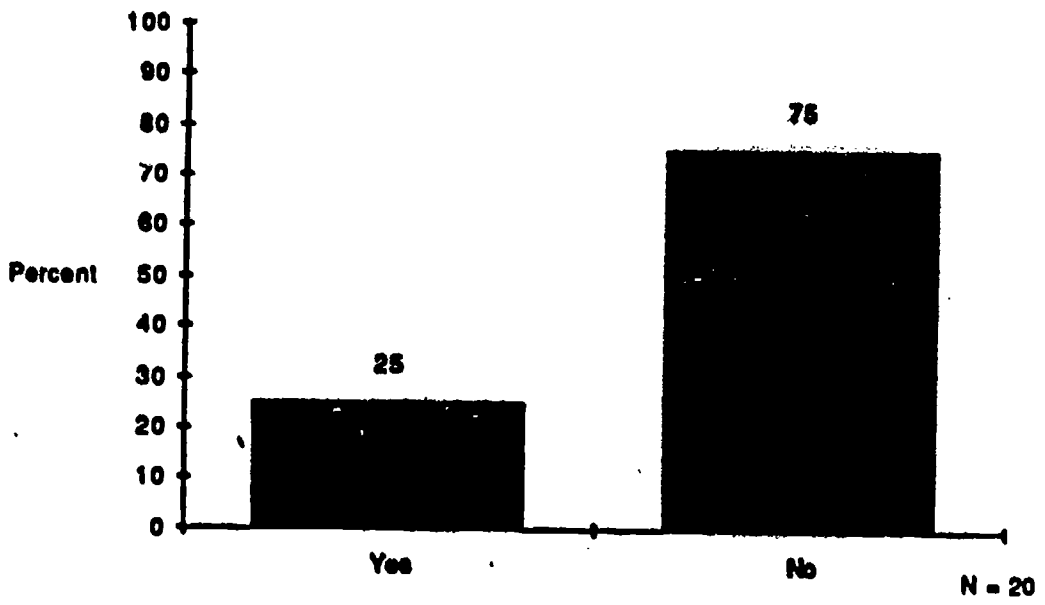
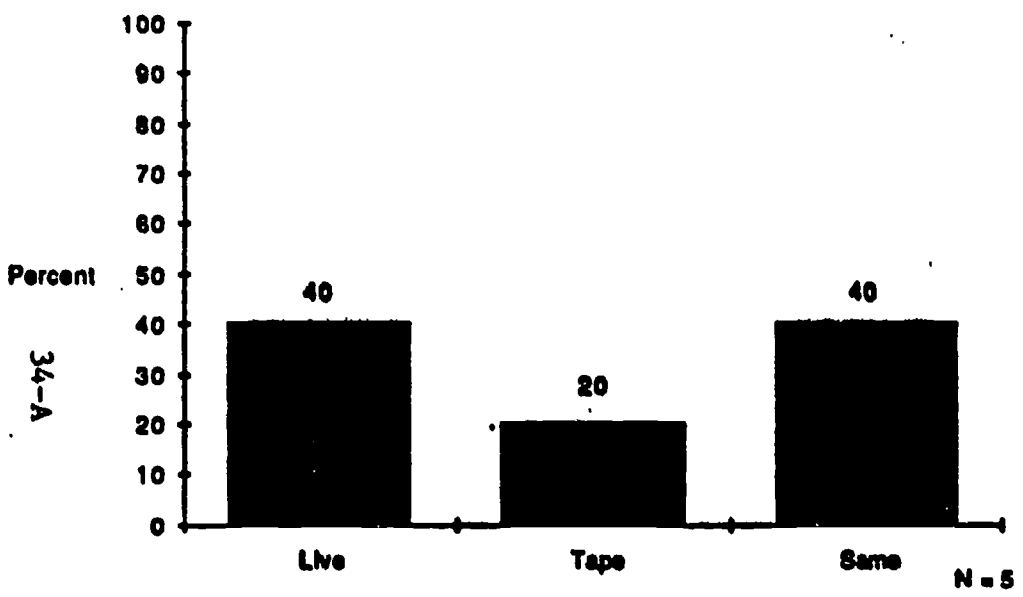


Figure 2.5B


5C. If you answered yes to both questions, in which of the two types of test did you feel more anxious or nervous? (UV)
Percent
100
90
80
70
60
50
40
30
20
10
0
40
20
40
Live
Tape
Same
N = 5
34-A

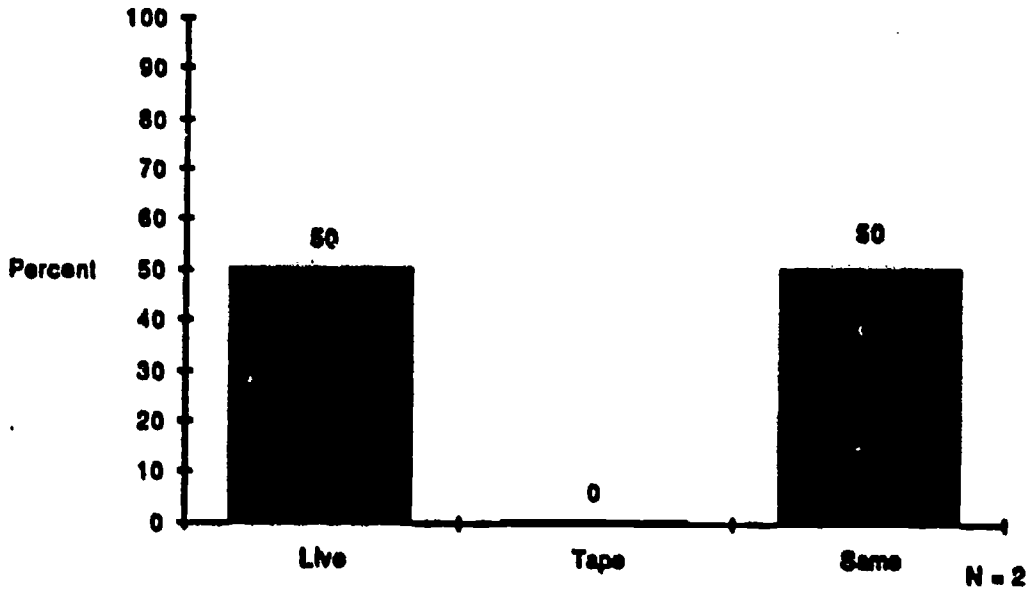
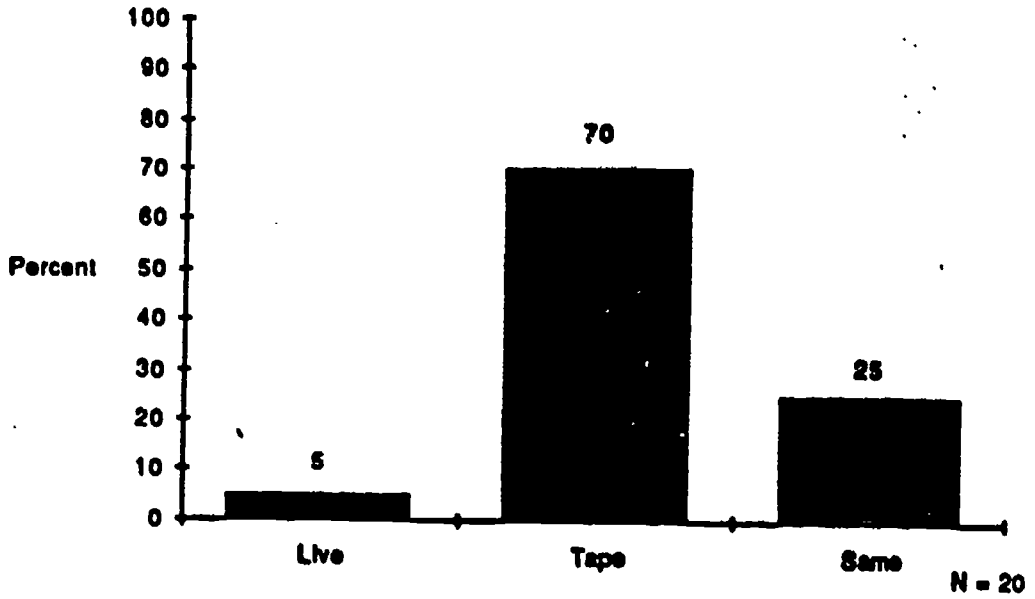

5C. If you answered yes to both questions, in which of the two types of test did you feel more anxious or nervous? (IV)
Percent
100
90
80
70
60
50
40
30
20
10
0
50
0
50
Live
Tape
Same
N = 2

Figure 2.5C

6. Which of the two types of test (live interview or taped test) did you feel was more difficult? (UV)



6. Which of the two typer of test (live interview or taped test) did you feel was more difficult? (IV)

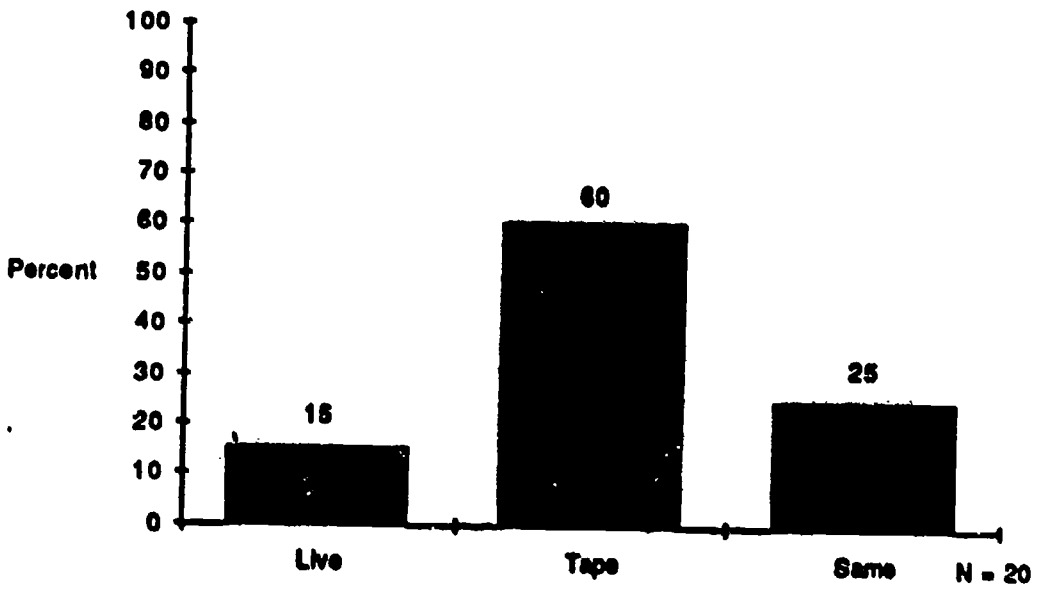


Figure 2.6

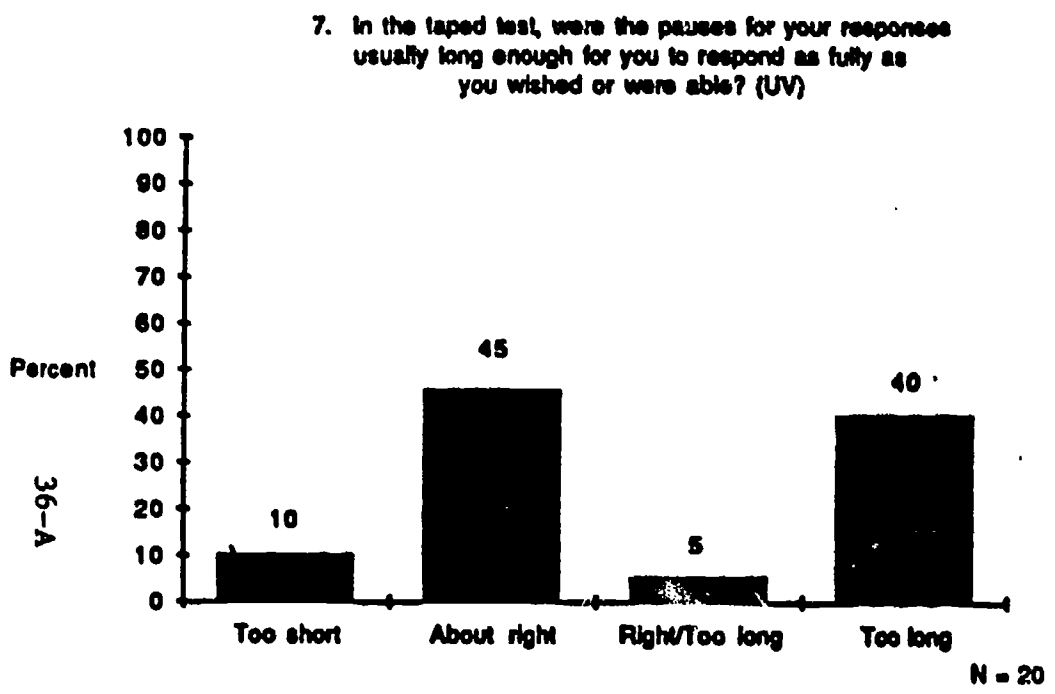
Questions 7 and 8 focused on technical qualities of the taped test.

(7) In the taped test, were the pauses for your responses usually long enough for you to respond as fully as you wished or were able?

(8) Where the directions on the taped test clear?

The majority of the subjects had no problem with the timed pauses in general. In both studies, only 2 examinees reported that the pauses were generally too short (see Figure 2.7). This means that 90% of the examinees felt they were able to respond as fully as they wished.

In both studies, 100% of the subjects felt the taped test directions were clear (Figure 2.8). This is a very positive reflection on the technical quality of the test. Because there is no possibility in the taped-test mode for examinees to ask questions once Part One of the test is begun, it is important that the directions be clear.


7. In the taped test, were the pauses for your responses usually long enough for you to respond as fully as you wished or were able? (UV)
Percent
100
90
80
70
60
50
40
30
20
10
0
10
45
5
40
Too short
About right
Right/Too long
Too long
N = 20

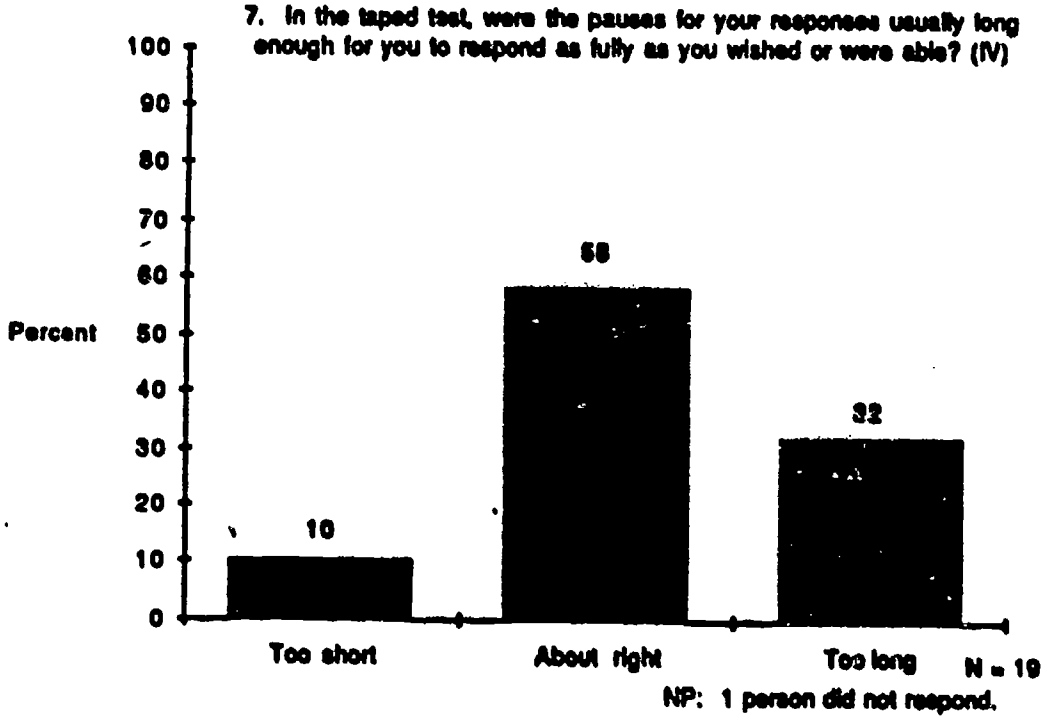
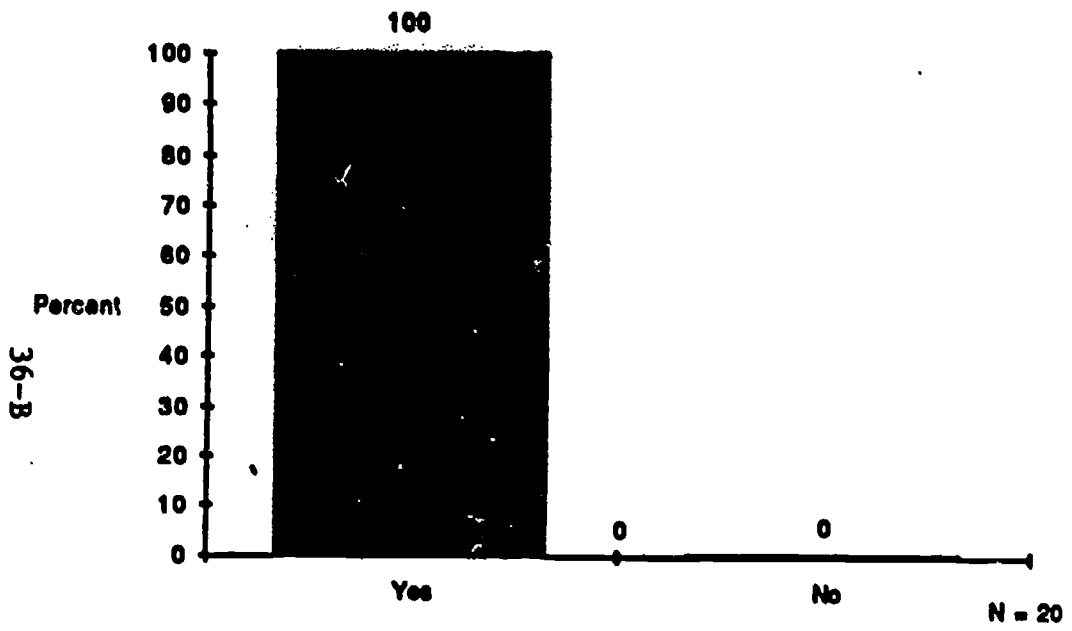

7. In the taped test, were the pauses for your responses usually long enough for you to respond as fully as you wished or were able? (IV)
Percent
100
90
80
70
60
50
40
30
20
10
0
10
58
32
Too short
About right
Too long
N = 19
NP: 1 person did not respond.

Figure 2.7

8. Were the directions on the taped test clear? (UV)



36-B

8. Were the directions on the taped test clear? (IV)

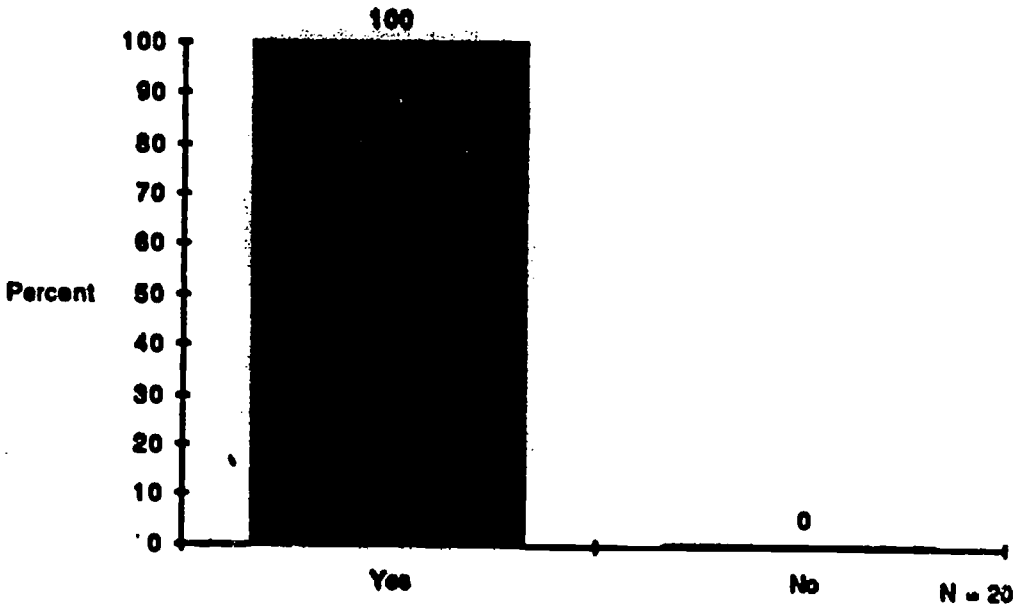


Figure 2.8

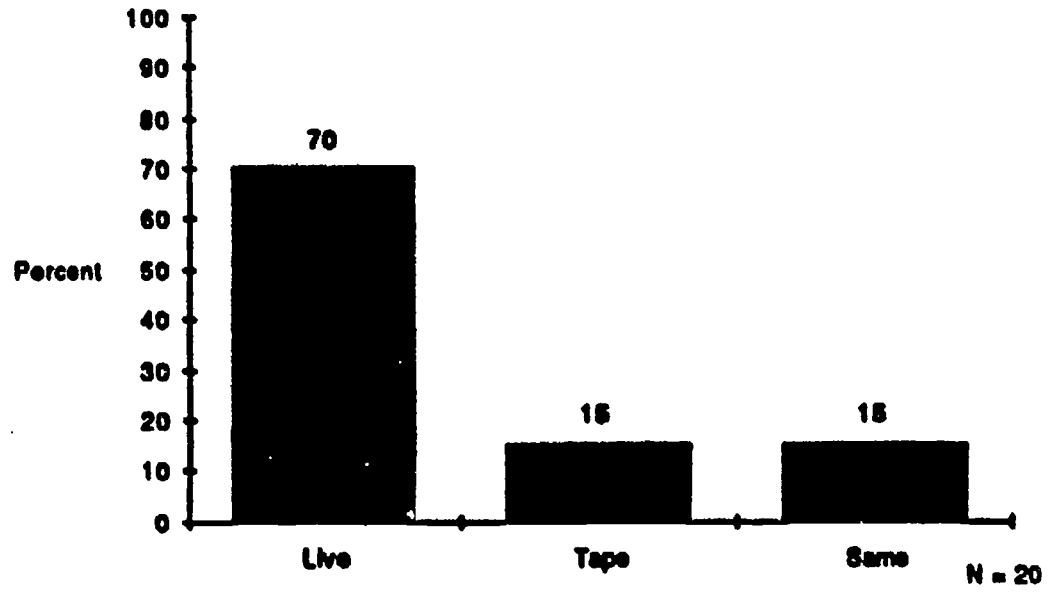
Question 9 is the 'catch-all' summary question.

(9) Which of the two types of tests did you prefer--the live interview or the taped test?

The majority (70% in each study) choose the live interview (Figure 2.9). From the comments in Appendix A-5, we can see that this is probably a reflection on the live interview testing mode, which seemed more natural, rather than a reflection on the technical quality of the taped test. However it is interesting to note that 30% of the subjects in each study either preferred the taped test or had no preference.

37-B

9. Which of the two types of test did you prefer--the live interview or the taped test? (UV)



9. Which of the two types of test did you prefer--the live interview or the taped test? (IV)

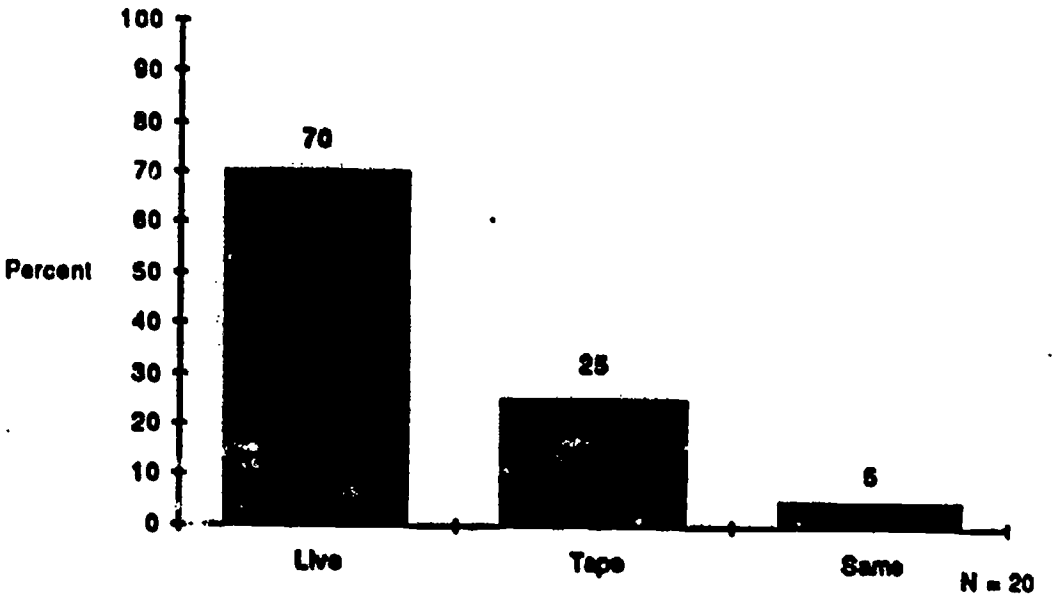


Figure 2.9

In summary, it appears that though the subjects were very positive about the content, technical quality and ability of the taped test to probe their speaking ability, its unfamiliar mode of testing and the perceived 'unnaturalness' of speaking to a machine caused a greater perceived difficulty and nervousness than the live interview. Thus, the majority of the subjects said they preferred the live interview to the taped test. Nevertheless, given the extremely high correlations between the two types of tests and the positive response to the taped test quality it appears that the taped test may confidently be used as an alternative, albeit "second choice" in the examinee's eyes, to the live interview. Moreover, it is expected that examinees who are more prepared for the test through the Examinee Handbook (see next section) may find the testing mode less threatening than the subjects participating in the validation study who went to both the live interview and the taped tests without any special advance preparation in order to avoid any biasing.

2.7 OPERATIONALIZATION OF THE TESTS

To operationalize the test, a supply of tests were professionally printed: 250 copies of each test form for each version were printed. In addition, 50 copies of each format of each test form for each version of the Master Test Tape were copied.

A Test Manual, giving comprehensive information on the development, uses, and administration of the test as well as the interpretation of examinee scores was prepared and is included as Appendix A-1. An Examinee Handbook was also prepared to be distributed to HeST examinees before taking the test and is found in Appendix A-2. The two booklets above establish and explain in detail the procedures for ordering and handling the test in-house. They also contain registration and order forms that are used in the operationalization of the test.

Announcements of the availability of the test are being produced to be sent to Hebrew Language Departments and other

interested parties throughout the country. In addition, an article on the test will appear in the Bulletin of Higher Hebrew Education in the Fall of 1989, together with the provisional ACTFL Hebrew Guidelines. Presentations on the test have been given by Elana Shohamy at a special seminar for language instructors at Brandeis University, and additional presentations are planned.

3. INDONESIAN SPEAKING TEST

3.1 MAJOR PROJECT ACTIVITIES

The day-to-day work of the project was conducted at the Center for Applied Linguistics (CAL) in Washington, DC. Charles W. Stansfield served as Project Director and Dorry Kenyon as Test Development Coordinator. A Test Development Committee was formed which included, in addition to the above, Mr. Daniel Kennedy, an experienced language test item writer and two experienced instructors of Indonesian with training in using the ILR oral proficiency testing procedures and rating scales: Ms. Jijis Chadran (Foreign Service Institute Language School) and Mr. Kadir Noor (United States Government Language School). Ms. Ruth Ephraim completed the local test development team as the artist for the test.

Three leading professors of Indonesian in U.S. academic institutions served as members of an External Review Board: James T. Collins (University of Hawaii at Manoa), Ellen Rafferty (University of Wisconsin, Madison) and John Wolff (Cornell University). Input from these individuals, received as they reviewed draft forms of the test items, played an important part in shaping the format and content of the final versions of the test.

The local test development committee met on a regular basis from November, 1988, to February, 1989, to develop the specific items for the two forms of the pilot version of the test. These items were based on the question types used in the semi-direct tests of Chinese and Portuguese, described above.

3.2 TRIALING OF THE TEST FORMS

The two forms of the IST were trialed on six individuals from the Washington, DC, area who had learned Indonesian in a variety of ways, some with experience in the country. The purpose of the trial was to ensure that the questions were clear, understandable and working as intended, and to check the appropriateness of the pause times allotted on the tape for examinee responses. The

basis using two tape-recorders. In each case, an Indonesian speaking member of the local test development committee observed the examinee taking the test and made notes on his or her performance on a specially prepared questionnaire (see Appendix B-3). In addition, upon completion of the test, examinees responded to a detailed questionnaire about it (Appendix B-3). In most cases, they were also debriefed on their testing experience in person.

The project coordinator took the tapes made during the trialing to the Spring, 1989, meeting of Indonesian instructors working on developing the ACTFL guidelines for Indonesian, at which the three members of the External Review Board were present. At one session of the meeting, the IST was discussed and one of the trialing tapes was listened to and commented on by the entire group. Many comments were offered for improving the pilot version of the test, most notably the need to contextualize the test to an even greater degree, giving specific information on the age and social status of every interlocutor presented in the test. This was deemed more crucial in the IST than in the PST, as one of the major characteristics separating Indonesian from Western languages is the importance of using correct terms of address and modifying speech depending on the social status and relationships of the speakers.

On the basis of data collected during the trialing and of comments from the External Review Committee members, the final format of the test was modified from the description of the proto-typical test format given above in three major ways. First, the Personal Conversation section of the IST is completely contextualized into a single role-play. In one form the examinee is being interviewed by a member of a scholarship selection committee; in the other, the examinee is talking before dinner to an Indonesian friend's aunt. Questions continued to be of a "warm-up" nature, focussing on the examinee's personal background, education, interests in the Indonesian language, etc. There are 11 questions in each form. Second, for the rest of the test, more

information on the one being spoken to is given in the IST than in the CST or PST. In narrative form, information on the person's sex, age, social status and name (when applicable) is given with each question. Third, it was decided to leave out picture item number 3 (detailed description), as the External Review Committee perceived it to be more of a vocabulary exercise and not helpful in rating examinees above the Novice level in Indonesian. In its place, an extra topic item was included.

Examples of the test questions on the final form of the test are available in the IST Examinee Handbook, located in Appendix B-2.

Once the local test development committee completed revising the two forms of the IST, the forms were again reviewed by the members of the External Review Board. After final revisions were made, test booklets and tapes for the validation study were prepared.

3.3 VALIDATION STUDY

Similar to the study conducted for the CST and PST, a research study was designed and carried out to validate the two forms of the IST. 16 subjects were involved in this study; eight were students in the intensive and regular Indonesian programs at Cornell University and eight were available locally in Washington, DC, having learned Indonesian through a variety of means. Each subject was first administered the OPI by Ms. Jijis Chadran, a certified tester from the Foreign Service Institute. The subjects at Cornell took the taped tests at the language lab at the University within two weeks after the live interview was administered. Subjects in Washington took the two taped tests at the Center for Applied Linguistics normally directly following the live interview. In one case, the subject returned a week later to take the taped tests. The design controlled for order of administration, with half of the subjects in each group (Cornell and Washington) receiving form A first and form B second, and the other half in reverse order. The design also attempted to select subjects

representing a variety of proficiency levels. Thus, participants were selected on the basis of the amount of their exposure to Indonesian. The responses to the OPI were recorded on tape for later scoring.

Ms. Chadran served as one of the raters for the study. Mr. Andang Poeraatmadja, an Indonesian examiner certified by the Foreign Service Institute, served as the second rater. The ratings of all the tapes were done independently and in random order, and subjects were rated anonymously; however, raters scored all the Oral Interviews before proceeding to rate the IST tapes. After all the ratings were completed, subjects were sent their test results in the mail: the scores of the two raters on the live interview and on each of the IST versions.

To proceed with the empirical analysis of the ratings, scores on both the live interview and the tape-based semi-direct tests converted to a scale combining both ACTFL and ILR rating scales with weights assigned as follows:

ACTFL/ILR Level	Coded as:
Novice-Low	0.2
Novice-Mid	0.5
Novice-High	0.8
Intermediate-Low	1.0
Intermediate-Mid	1.5
Intermediate-High	1.8
Advanced	2.0
Advanced-Plus	2.8
Superior/Level 3	3.0
High-Superior/Level 3+ and above	3.8

The system of score coding above is based on the ILR 0 to 5 rating scale and is intended to assign an appropriate numerical value to the proficiency level descriptions. For example, proficiency at an Advanced-Plus level is characterized by many of the same features as at the Superior/3 level, though the examinee cannot sustain the performance. Thus, the numerical interpretation falls closer to 3.0 than mid-way between the two, as may be expected.

The several tables below provide descriptive statistics,

interrater reliabilities and parallel-form reliability data obtained in the study. Rater 1 is Ms. Chadran and Rater 2 is Mr. Poeraatmadja.

Table 3.1 shows the mean score, standard deviation and other basic statistics for the ratings assigned by each of the two raters to subject performance on each of the semi-direct test forms and on the live interview.

Table 3.1
Descriptive Statistics for Scoring Levels Assigned Tape and Live Tests

Test Form	Minimum Score	Maximum Score	Mean	Standard Deviation
Form A (n=16)				
Rater 1	0.8	3.8	2.47	0.94
Rater 2	0.8	3.8	2.50	0.92
Form B (n=16)				
Rater 1	0.5	3.8	2.58	1.03
Rater 2	0.5	3.8	2.44	1.00
Live Interview (n=16)				
Rater 1	0.8	3.8	2.64	0.96
Rater 2	0.8	3.8	2.63	0.90

Table 3.2 shows the frequency of ratings given by the two raters on the live interviews (n=16) and on both forms (n=32).

Table 3.2
Frequency Distributions

Live Interview Ratings (Rater 1)

Rating	Frequency	Percent
0.8	1	6.25
1	1	6.25
1.5	1	6.25
2	2	12.50
2.8	6	37.50
3	1	6.25
3.8	4	25.00

Live Interview Ratings (Rater 2)

Rating	Frequency	Percent
0.8	1	6.25
1	1	6.25
1.5	1	6.25
1.8	1	6.25
2	3	18.75
2.8	5	31.25
3	1	6.25
3.8	3	18.75

Rating of IST Test Forms (Rater 1)

Rating	Frequency	Percent
0.5	1	3.13
0.8	1	3.13
1	2	6.25
1.5	2	6.25
1.8	2	6.25
2	5	15.63
2.8	9	28.13
3	3	9.38
3.8	7	21.88

Rating of IST Test Forms (Rater 2)

Rating	Frequency	Percent
0.5	1	3.13
0.8	1	3.13
1	2	6.25
1.5	2	6.25
1.8	2	6.25
2	5	15.63
2.8	8	25.00
3	4	12.50
3.8	6	18.75

These statistics indicate that each form of the test was taken by a group of examinees that was quite proficient in Indonesian. The average ratings assigned by both raters, between an Advanced and Advanced-Plus, reflect this. The mean scores for each rater were very similar, indicating that the raters were almost equal in their degree of severity.

The degree of agreement between the absolute ratings may be seen from the following three cross-tab diagrams. First, Table 3.3 presents the ratings of Rater 1 (down) against the ratings of Rater 2 (across) for the live interview.

Table 3.3
Crosstabulations of Live Interview Ratings (n=20)

Rater 1 (down)/ / Rater 2 (across) Frequency	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.8	1	0	0	0	0	0	0	0	1
1	0	1	0	0	0	0	0	0	1
1.5	0	0	0	1	0	0	0	0	1
1.8	0	0	0	0	0	0	0	0	0
2	0	0	0	0	2	0	0	0	2
2.8	0	0	0	0	0	5	1	0	6
3	0		0	0	0	0	1	0	1
3.8	0		0	0	0	0	1	3	4
Total	1	1	0	1	2	5	3	3	16

For the live interview, there was total agreement in 81.25% of the ratings. In the three cases were there was disagreement, Rater 2 was more generous in two cases and Rater 1 was more generous in one. None of the disagreements was more than one step away on the rating scale. Only one case crossed an ACTFL level boundary.

Table 3.4 presents the ratings of Rater 1 (down) against Rater 2 (across) for IST Form A.

Table 3.4

Crosstabulations of Form A Ratings (n=16)

Rater 1 (down) / Rater 2 (across)

Frequency	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.8	1	0	0	0	0	0	0	0	1
1	0	1	0	0	0	0	0	0	1
1.5	0	0	0	1	0	0	0	0	1
1.8	0	0	0	1	0	0	0	0	1
2	0	0	0	0	3	0	0	0	3
2.8	0	0	0	0	0	4	1	0	5
3	0	0	0	0	0	0	1	0	1
3.8	0	0	0	0	0	0	0	3	3
Total	1	1	0	2	2	4	2	3	16

From Table 3.4 we see that the agreement of the absolute ratings was again extremely high. There was total agreement in 87.5% of the 16 Form A ratings. For the two cases of disagreement, neither was more than one step away on the rating scale. Rater 2 was more generous in both cases, and one of the two disagreements crossed an ACTFI level boundary.

Table 3.5 presents the ratings of Rater 1 (down) against Rater 2 (across) for IST Form B.

Table 3.5

Crosstabulations of Form B Ratings (n=16)

Rater 1 (down)/ / Rater 2 (across)

Frequency	0.5	1	1.5	1.8	2	2.8	3	3.8	Total
0.5	1	0	0	0	0	0	0	0	1
1	0	1	0	0	0	0	0	0	1
1.5	0	0	1	0	0	0	0	0	1
1.8	0	0	1	0	0	0	0	0	1
2	0	0	0	0	2	0	0	0	2
2.8	0	0	0	0	1	2	1	0	4
3	0	0	0	0	0	2	0	0	2
3.8	0	0	0	0	0	0	1	3	4
Total	1	1	2	0	3	4	2	3	16

From Table 3.5 we see that the agreement of the absolute ratings was again relatively high (62.5%). Where there was disagreement, Rater 1 was more generous in five of the six cases; Rater 2 was more generous in only one. For none was the disagreement more than one step away on the rating scale. Three of the disagreements crossed an ACTFL level boundary.

The tables above show very high consistency between the two raters and no consistent trend apparent in either rater in terms of rater severity, though Rater 1 was more generous than Rater 2 on Form B.

Interrater reliabilities (Pearson product-moment correlations) between the ratings assigned by Rater 1 and those assigned by Rater 2 for the two semi-direct test forms and for the live interview are shown in Table 3.6 below.

Table 3.6
Interrater Reliabilities

Test Form	Correlation
A (n=16)	.99
B (n=16)	.96
Interview (n=16)	.97

These interrater reliabilities are all uniformly high across the test forms and the live interview. Interrater reliability was not adversely affected by the semi-direct test format. This suggests that the IST elicits a sample of speech as ratable as the live interview.

On performance-based tests such as the IST, there is an increased concern for test-retest reliability. This form of reliability measures the degree of inconsistency in examinee performance on two separate administrations of the same test. The amount of inconsistency reflects the degree to which the test score may be confounded by such inconsistency. Therefore, it is important to examine this factor. However, on a test with a limited number of questions such as the IST, it is not wise to administer the same test twice, since the first sitting will serve to instruct the examinee in the task at hand. (For a thorough discussion of this "reactivity effect," see Stansfield and Ross, 1988, p. 174.) Under such circumstances, it is preferable to administer different forms of the test while still using the same rater to score the performance. This type of reliability is known as parallel-form reliability, which is the degree of correlation between scores on two forms of the test.

Parallel-form reliabilities for the same subject taking two different test forms, with the same rater scoring both forms, are shown in Table 3.7.

Table 3.7
Parallel-Form Reliabilities (Same Rater)

	Rater 1	Rater 2
Forms A and B (n=16)	.92	.95

The statistics indicate that the parallel form reliability of the IST is very high. With the first rater, the parallel-form reliability was .92, while with Rater 2 it was even higher (.95). Such favorable statistics provide strong support for the proposition that each form of the IST elicits a sample of speech that is uniformly challenging to the examinee. The fact that the parallel-form reliability was high for two different raters supports the claim that the sample of speech elicited by different forms is equally ratable.

In summary, the evidence from Table 3.7 warrants the conclusion that natural variations in examinee oral language performance are adequately controlled for by the IST format.

Table 3.8 shows parallel-form reliabilities for subjects taking two different test forms, with each form scored by a different rater.

Table 3.8
Parallel Form Reliabilities (Different Forms and Raters)

Rater/Form Combination	Correlation
Rater 1/Form A - Rater 2/Form B (n=16)	.90
Rater 1/Form B - Rater 2/Form A (n=16)	.91

This type of parallel-form reliability involves error that can be attributed to natural variation in examinee speech, error that can be attributed to differences in test form, and error that can be attributed to differences in raters. Thus, it may be viewed as a lower-bound estimate of the reliability of an IST score. Again the reliabilities here are high, even under these severe conditions

(different forms and different raters).

Correlations of semi-direct test scores with the live face-to-face interview are given in Table 3.9 below. These correlations are evidence of the validity of the IST as a surrogate live interview.

Table 3.9
Correlations with Live Interview

Rater/Form	Rater 1/Interview	Rater 2/Interview
Rater 1/Form A (n=16)	.95	.96
Rater 1/Form B (n=16)	.93	.90
Rater 2/Form A (n=16)	.93	.96
Rater 2/Form B (n=16)	.94	.96
All Matched Interviews/Forms (64 pairs)	.95	

Again, the correlations are all high. The average correlation based on 64 pairs of ratings (16 subjects x 2 IST forms x 2 ratings, correlated with the score assigned for the live interview) was .95. Such results support the claim that the IST is a valid measure of oral language proficiency that can be substituted for a live interview.

The degree of agreement in absolute ratings given on the live interview with ratings given on the same examinee's IST may be seen from the following cross-tab diagram. In Table 3.10 all 64 pairs of interview ratings (down) with IST ratings (across) are presented.

Table 3.10

Crosstabulations of Interview ratings by IST ratings

Interview (down) / / IST (across)

Frequency	0.5	0.8	1	1.5	1.8	2	2.8	3	3.8	Total
0.5	0	0	0	0	0	0	0	0	0	0
0.8	2	2	0	0	0	0	0	0	0	4
1	0	0	4	0	0	0	0	0	0	4
1.5	0	0	0	2	0	0	0	0	0	2
1.8	0	0	0	1	1	0	0	0	0	2
2	0	0	0	1	3	3	1	0	0	8
2.8	0	0	0	0	0	8	12	2	0	22
3	0	0	0	0	0	0	4	4	0	8
3.8	0	0	0	0	0	0	0	1	13	14
Total	2	2	4	4	4	11	17	7	13	64

From the table we see that in 64% of the cases there was an absolute agreement between the two ratings. For all of the remaining ratings except for one, the difference was only one step away on the rating scale. In one case, an examinee was awarded a 2.0 on an interview, but received a 1.5 by the same rater on one of the tape forms. Thus, for 98% of the ratings, the rating on the live interview and the rating on the IST was equal to or less than one step away on the rating scale. Thus, besides the high correlations documented above, the absolute values given to examinees on both the live interview and the IST were extremely close.

The above chart shows, however, that when there was a disagreement between the rating on the taped test and the rating on the interview, in 83% of the disagreements the score on the live interview was higher than the score on the taped test (although in only eight cases--44% of the total number of disagreements--did this mean crossing an ACTFL level boundary). In the three cases were a higher score was awarded on the IST, two cases crossed an

ACTFL level boundary.

One possible reason for the more generous ratings on the live interview may be the unfamiliarity of the raters with the taped test. It appears that when in doubt about a rating on the taped test, they erred on the side of being conservative, while they knew better what to look for in the live interview. Another explanation may be that the vast majority of the examinees were of very high oral proficiency level. 81% were at the Advanced level or above, and 38% were at the Superior level or above. 25% of the examinees (or 4 of the sixteen) was rated at the "High-Superior" level by at least one rater on one test. While the OPI (as given by the Foreign Service) can accommodate higher levels, the taped tests were designed for the range beginning at Intermediate and going up to Superior (as a ceiling). Thus, a quarter of the sample population in this study can be considered out of range for the taped test. These high level examinees may not at times have had the opportunity to fully show what they could do. In fact (see below), 6 examinees felt that the pause times in general were too short, while none of the examinees felt that the pause times were too long.

As a general summary of the statistical information above, it may be stated that both forms of the semi-direct test reveal high interrater reliabilities, with Pearson product-moment correlations for Form A at .99 and Form B at .96. Parallel form reliabilities are also very high, even under the most "severe" conditions (i.e., different raters rating two different forms), where correlations are at .91 (Form A) and .92 (Form B); with the same rater, correlations range from .92 to .95. The correlations with the live interview are also very high; with the same rater they range from .93 to .96 and with different raters from .90 to .96.

3.4 SUBJECT RESPONSE TO THE TEST

As part of the validation study, feedback information from the participants on various aspects of their experience with and opinions about both types of testing procedures were elicited by

means of a short questionnaire (Appendix B-4). The questionnaire was given to the subjects directly after completing the semi-direct tests and completed and returned before leaving the testing site. All subjects completed the questionnaire for a 100% participation rate.

The answers to the examinee questionnaires are given in graphic summary form below. Written comments in response to the questionnaire are presented in Appendix B-5.

The first two questions sought to elicit from the subjects the extent to which they felt their Indonesian speaking ability had been probed by the two types of test: the live interview and the IST.

(1) Over the course of the live interview, do you feel that your maximum level of speaking ability in Indonesian was adequately probed by the tester?

(2) Over the course of the taped test, do you feel that the descriptions, narratives, situations, and other types of questions in the test were adequate to probe your maximum level of speaking ability in Indonesian?

There was a difference in how the examinees felt the two testing modes were able to probe their maximum level of Indonesian speaking ability. Figure 3.1 reveals that 69% answered affirmatively as regards the live interview, but, as Figure 3.2 shows, only 38% answered affirmatively as regards the taped test. This result is very different to the results obtained for the Chinese, Portuguese, and Hebrew tests, and is most likely due to the fact that the proficiency level of the sample in this study was so high (as discussed above). Even 31% of the examinees felt that the OPI, which can test at levels above Superior, failed to probe their maximum level of Indonesian proficiency.

1. During the live interview, do you feel that your maximum level of speaking ability in Indonesian was adequately probed by the tester?

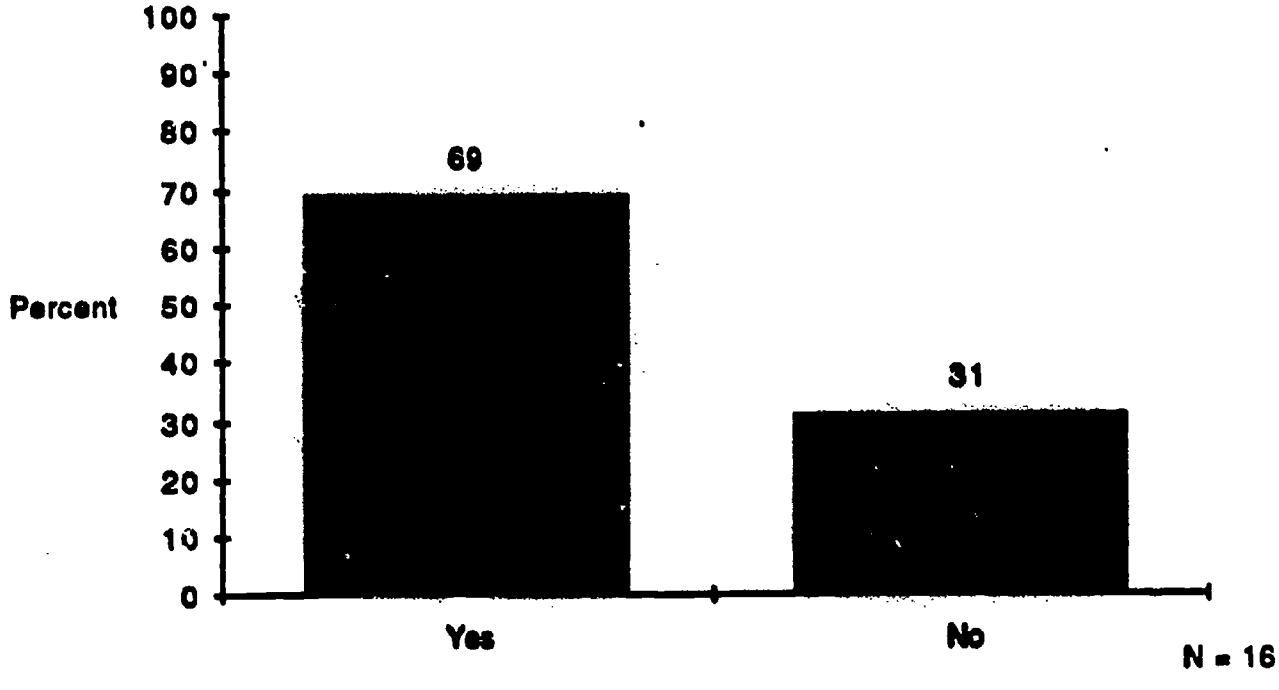


Figure 3.1

2. During the taped test, do you feel that the descriptions, narratives, situations and other types of questions were adequate to probe your maximum level of speaking ability in Indonesian?

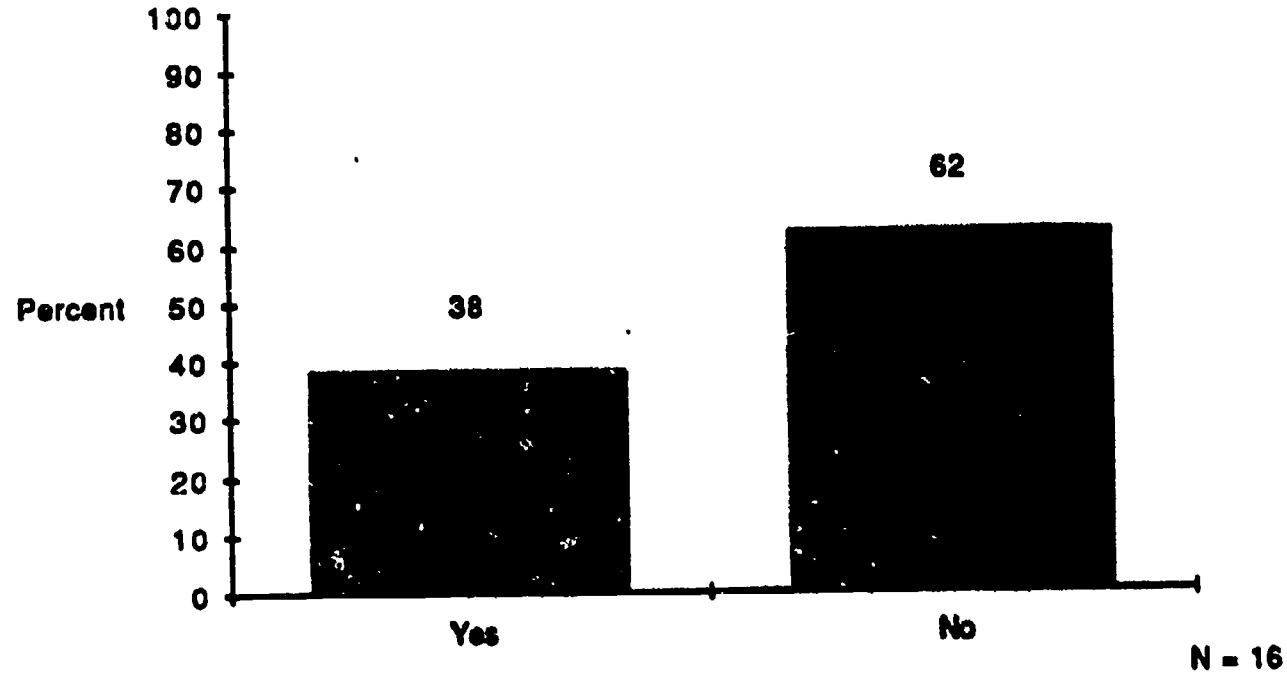


Figure 3.2

The next two questions focused on whether the subjects perceived any unfair questions on either test format.

(3) In the live interview, were there any questions asked or speaking situations required which you felt were in any way 'unfair'?

(4) In the taped tests, were there any picture/descriptions, narratives, situations, or other questions that you felt were in any way 'unfair'?

As shown in Figures 3.3 and 3.4, only a small minority felt there were unfair questions in either of the testing modes (1 individual for the live interview, 3 for the taped tests). This small number again most likely reflects the fact that the IST was at a level that was not challenging enough for many of the examinees, since a taped test which cannot adapt itself to the level or circumstances surrounding the testing of a particular examinees. In the taped test, the examinee is asked every question, whether it is challenging or not.

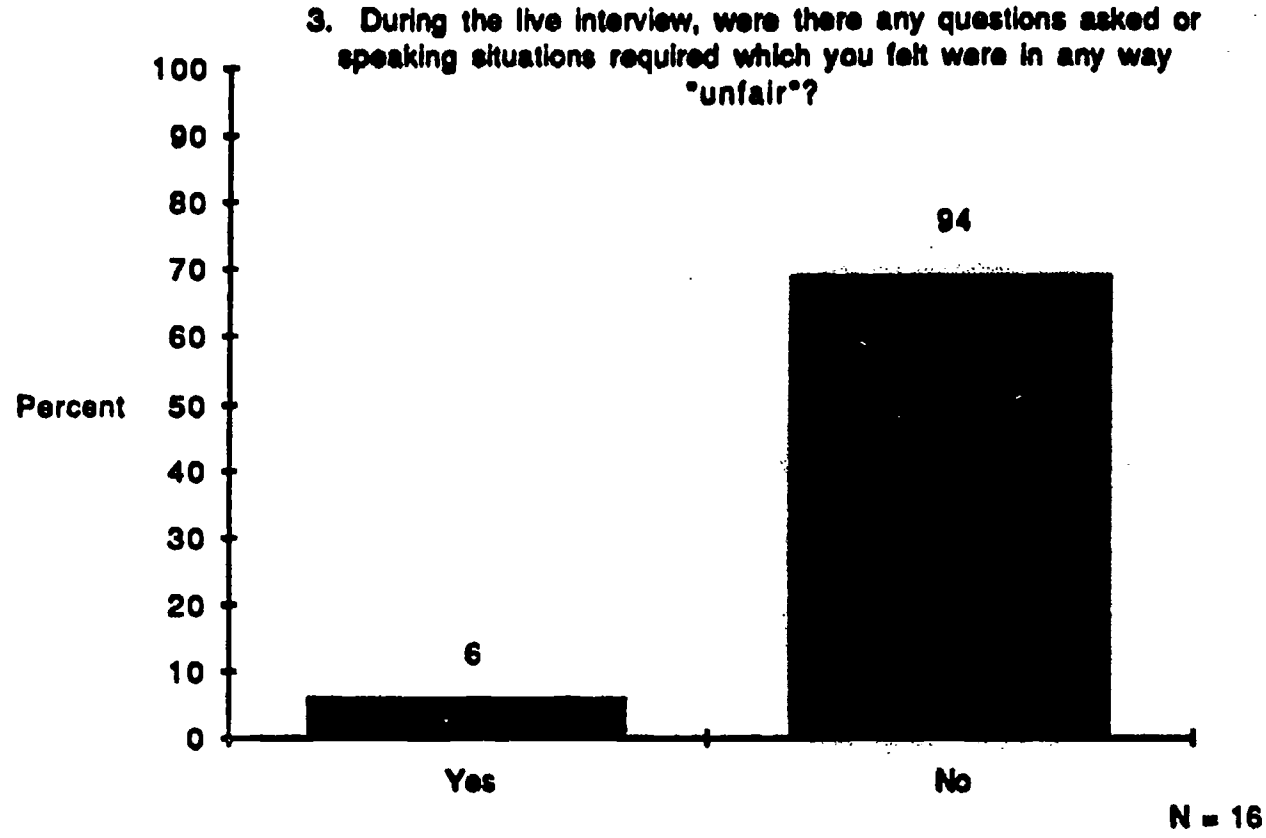


Figure 3.3

4. In the taped tests, were there any picture/descriptions, narratives, situations or other questions that you felt were in any way "unfair"?

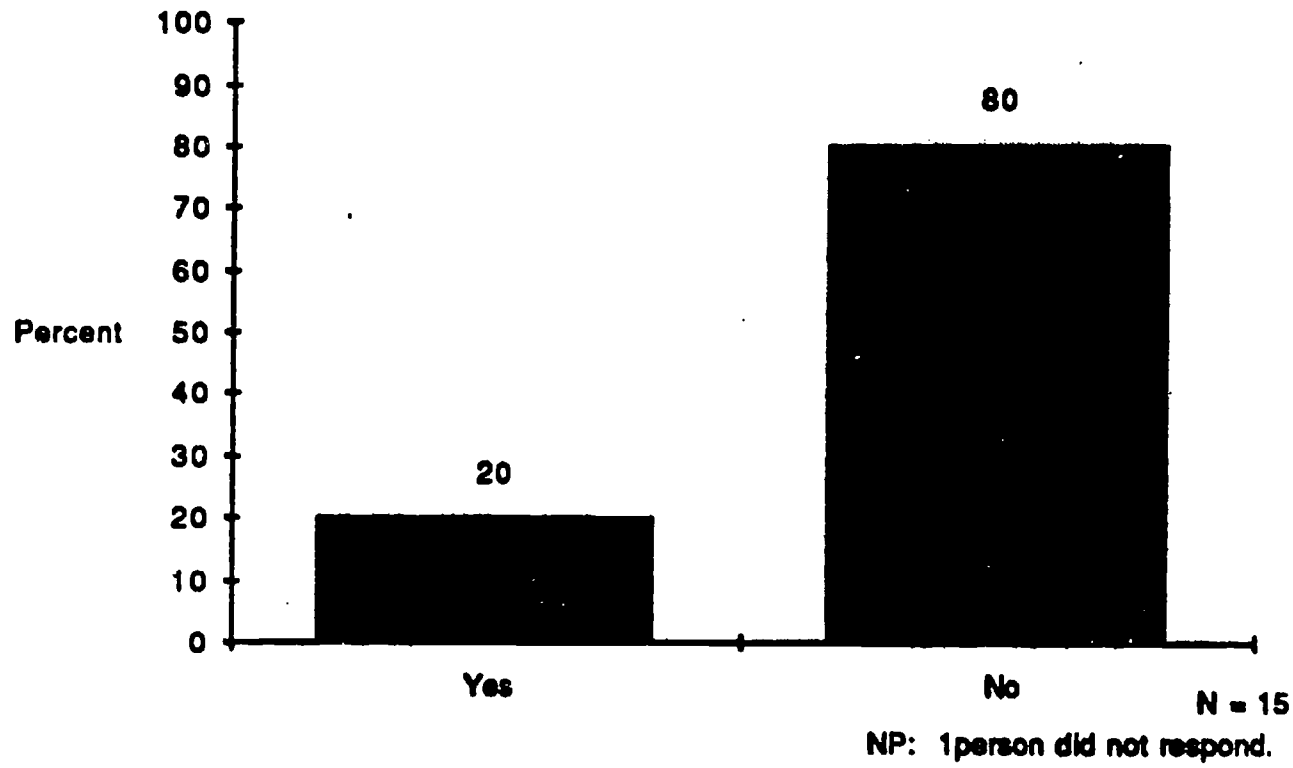


Figure 3.4

The next two questions focused on the subject's affective perceptions of the test.

(5) (A) Did you feel unduly nervous in the live interview?
(B) Did you feel unduly nervous in the taped tests?
(C) If you answered yes to both questions, in which of the two types of test did you feel more anxious or nervous?

(6) Which of the two types of tests (live interview or taped test) did you feel was more difficult?

Because the semi-direct mode of testing may be unfamiliar and perhaps 'unnatural' to students in general, it would not be unusual for a large percentage of the students in this study to feel more nervous in the taped test than in the live interview. Five subjects (31%) answered they felt unduly nervous in the live interview, while 8 subjects (50%) answered affirmatively for the taped test. However, of the four who answered yes to both questions (25% of the entire group), only one felt more nervous taking the taped test, while two felt more nervous taking the live interview (see Figures 3.5A, 3.5B and 3.5C).

Question 6 focused on perceived difficulty. Despite the fact that subjects did approximately the same on both tests (see correlations above), a small majority (56%) of the subjects perceived the taped test as more difficult, while 31% felt the live interview was more difficult (Figure 3.6). This, again, most likely reflects the high level of proficiency in the group. Of the five semi-direct tests developed by CAL, this is the smallest percentage that felt the taped tests were more difficult. It appears the more one is proficient in the language, the less "threatening" the taped mode is, even when it is unfamiliar.

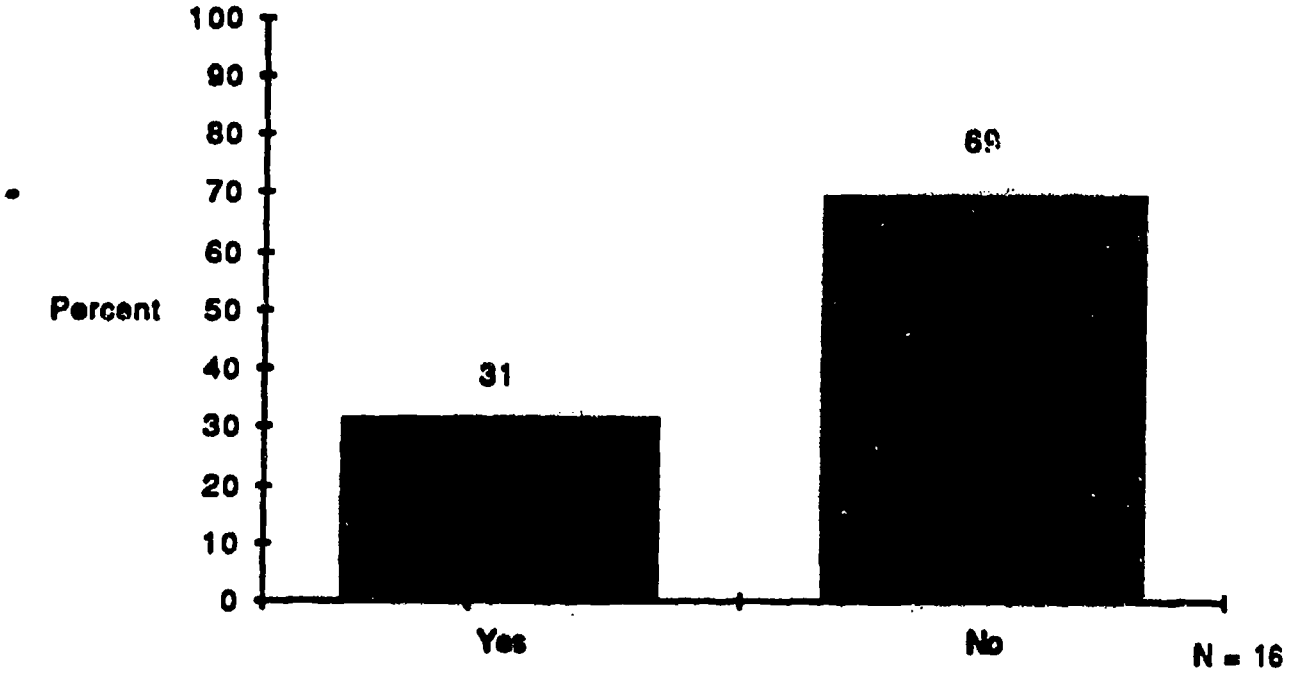


Figure 3.5A

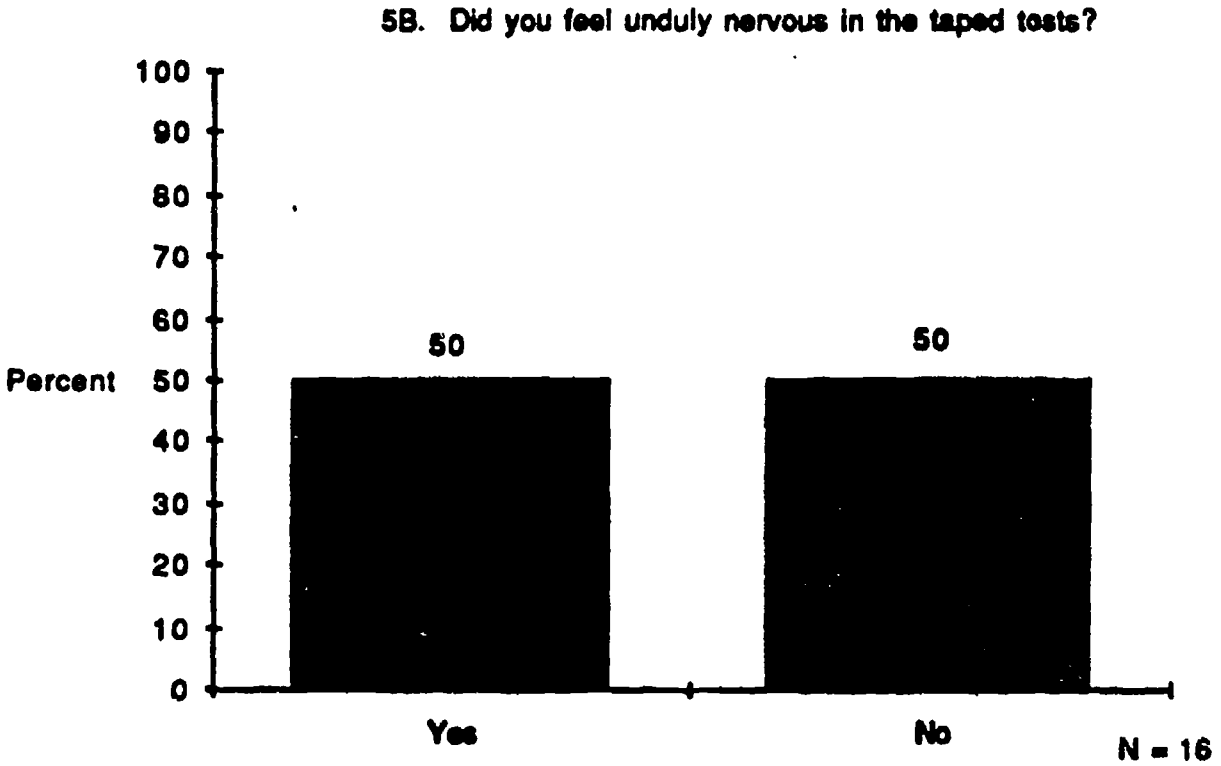


Figure 3.5B

5C. If you answered yes to both questions, in which of the two types of test did you feel more anxious or nervous?

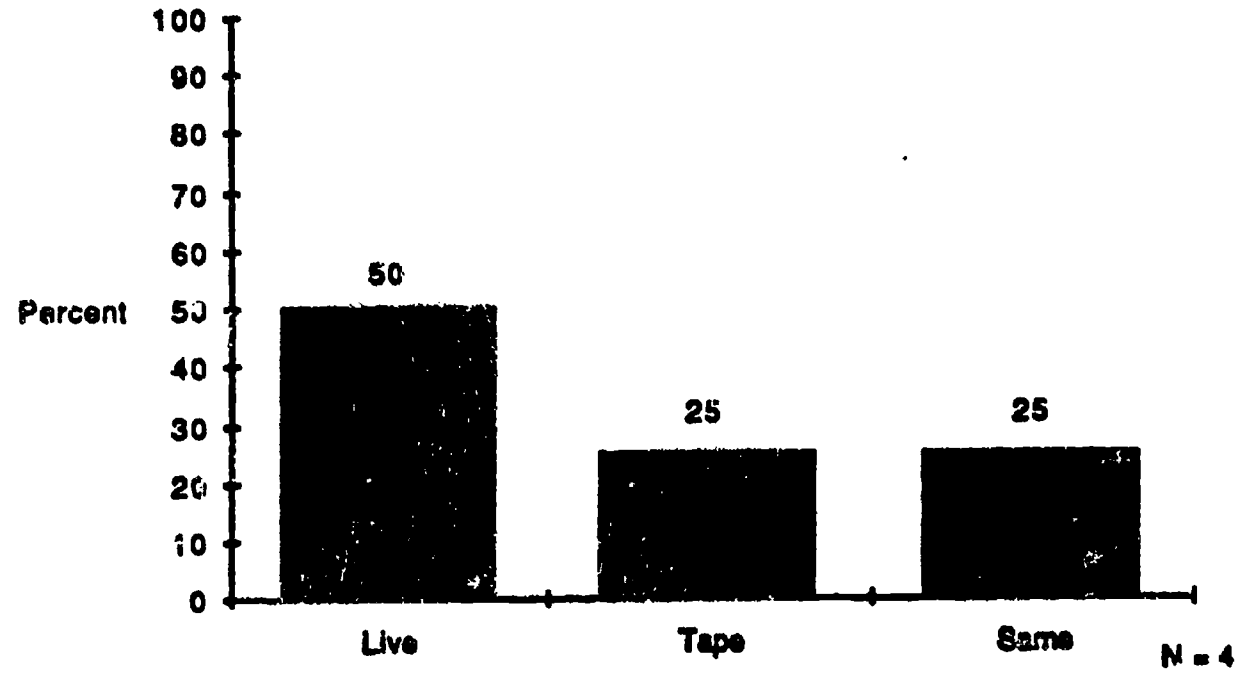


Figure 3.5C

6. Which of the two types of test (live interview or taped test) did you feel was more difficult?

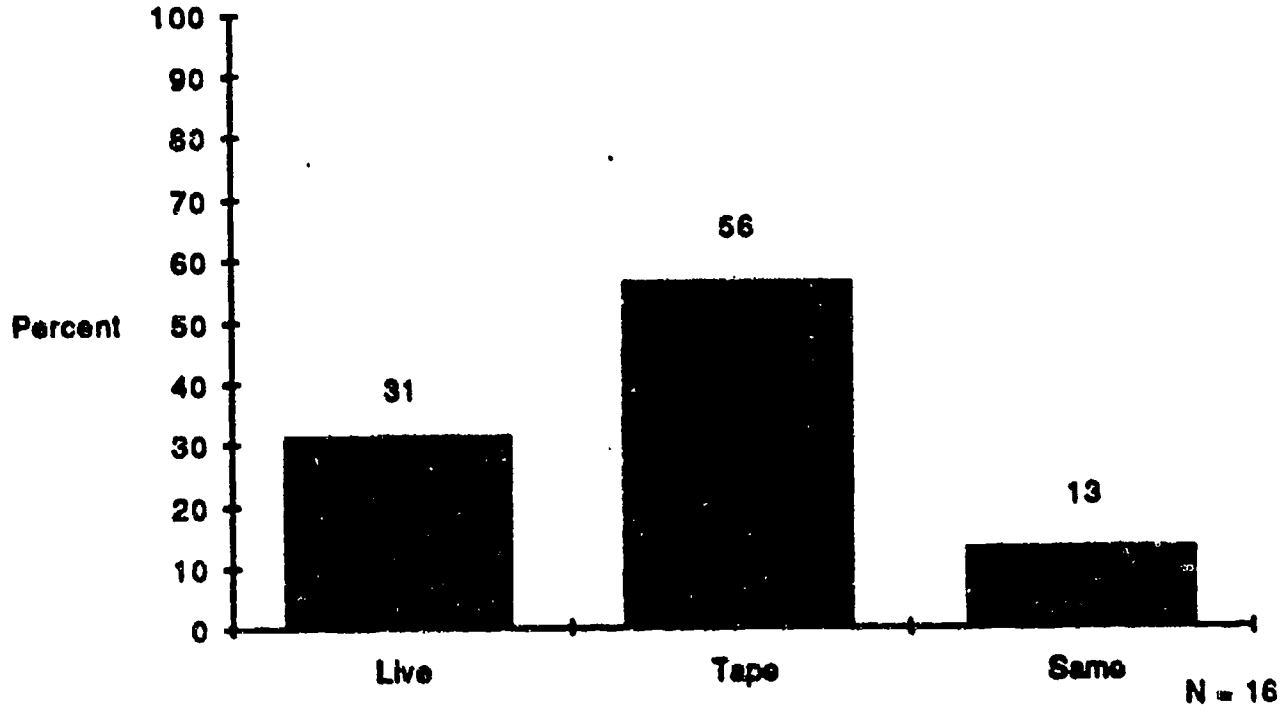


Figure 3.6

Questions 7 and 8 focused on technical qualities of the taped test.

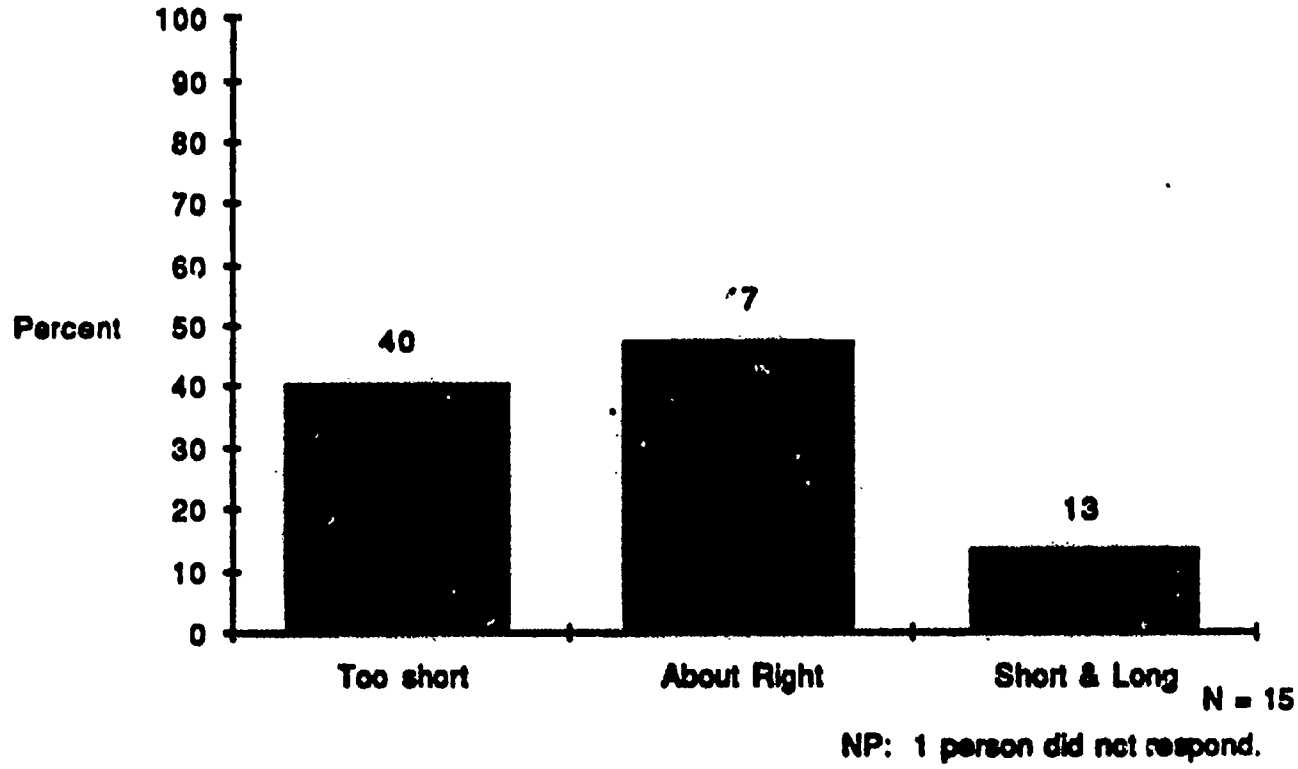
(7) In the taped test, were the pauses for your responses usually long enough for you to respond as fully as you wished or were able?

(8) Where the directions on the taped test clear?

As mentioned above, a large number of examinees (40%) felt the timed pauses were in general too short, while 47% felt the pauses were about right (Figure 3.7). Again, this is most likely a reflection on the high level of proficiency demonstrated by the sample in this study compared to the level for which the test is intended.

100% of the subjects felt the taped test directions were clear, which is a very positive reflection on the technical quality of the test (Figure 3.8). Because there is no possibility in the taped-test mode for examinees to ask questions once Part One of the test is begun, it is important that the directions be clear.

7. In the taped test, were the pauses for your responses usually long enough for you to respond as fully as you wished or were able?



NP: 1 person did not respond.

Figure 3.7

8. Were the directions on the taped test clear?

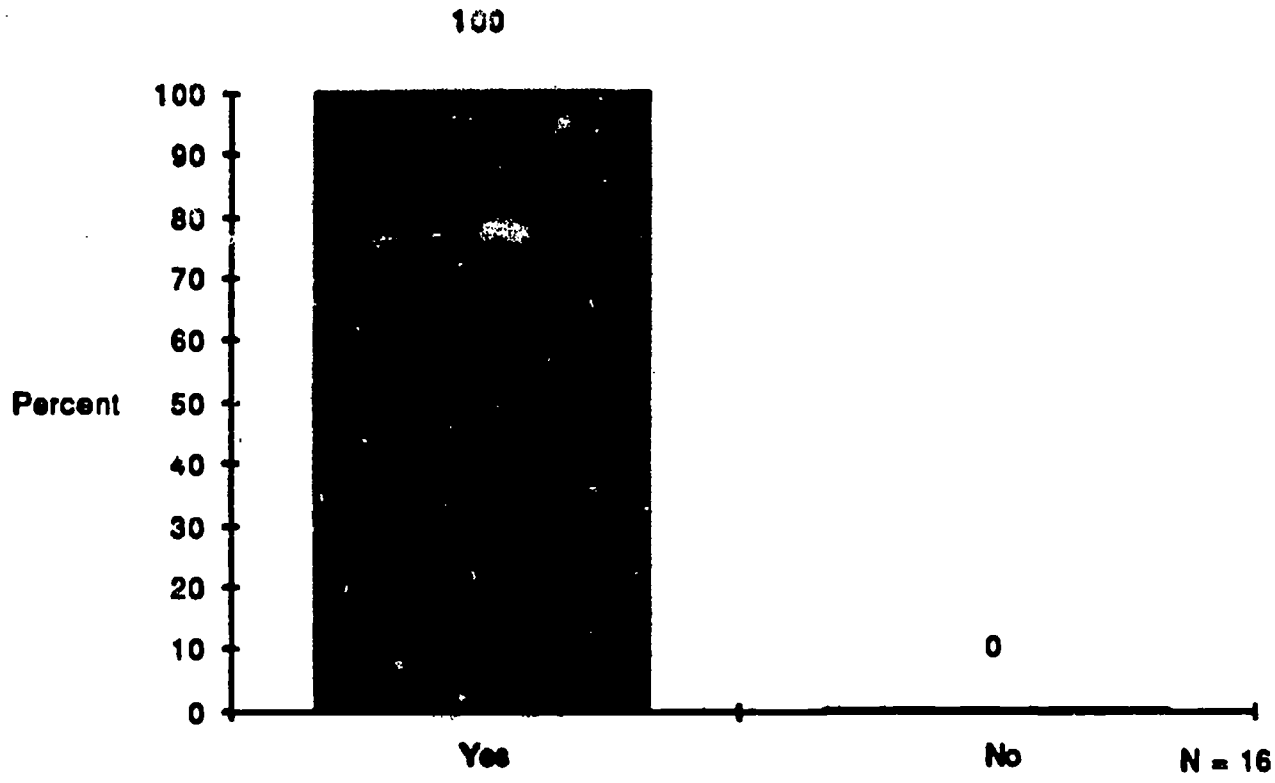


Figure 3.8

Question 9 is the 'catch-all' summary question.

(9) Which of the two types of tests did you prefer--the live interview or the taped test?

The majority (88%) choose the live interview, while two of the examinees had no preference (Figure 3.9). Again, besides the unfamiliarity of the taped-testing mode, this result was most probably due to the mismatch between the levels for which the test is intended and the proficiency levels of the subjects in the sample.

9. Which of the two types of test did you prefer--the live interview or the taped test?

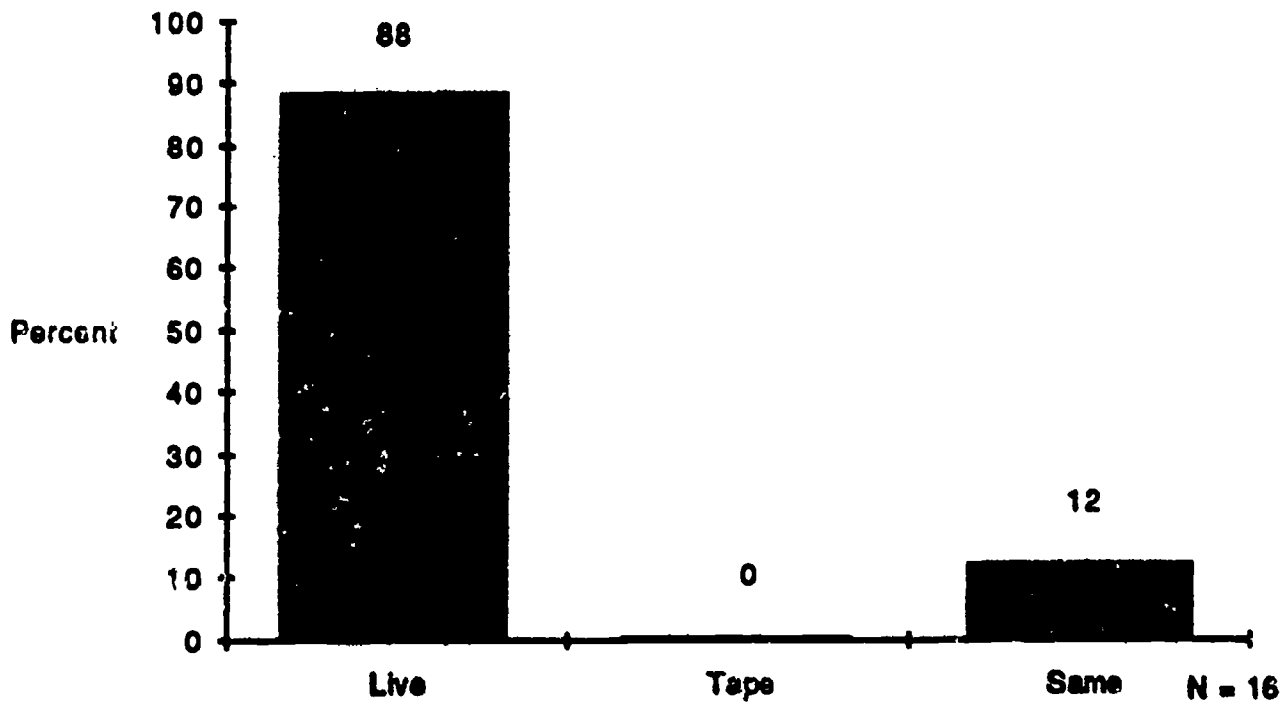


Figure 3.9

In summary, it appears that there was a mismatch between the proficiency levels for which the test was intended and the proficiency levels of those who took the test as part of the sample for the validation study. The sample subjects tended to have a very high ability in Indonesian (only 19% were below the Advanced level while 25% scored above Superior on at least one test), while the IST is designed to cover the range from Intermediate to Superior (ILR level 3) with Superior as a ceiling. Much of the findings in the validation study, and especially in the student questionnaire, may be explained by this mismatch. If more Intermediate level and fewer "High-Superior" examinees had been included in the sample, it is likely that the results of the student questionnaire would be more similar to those obtained for the studies on the Chinese, Portuguese and Hebrew Speaking Tests, all of which were very similar. It should be noted, however, that this mismatch did not negatively affect the raters' ability to score the IST reliably.

3.5 OPERATIONALIZATION OF THE TESTS

To operationalize the test, a supply of tests were professionally printed: 100 copies of each test form were printed. In addition, 50 copies of each form of the Master Test Tape were copied.

A Test Manual, giving complete information on the development, uses, and administration of the IST as well as the interpretation of examinee scores was prepared and is included as Appendix B-1. An Examinee Handbook was also prepared to be distributed to IST examinees before taking the test and is found in Appendix B-2. The two booklets above establish and explain in detail the procedures for ordering and handling the test in-house. They also contain registration and order forms that are used in the operationalization of the test.

Announcements of the availability of the test are being produced to be sent to Indonesian Language Departments and other interested parties throughout the country.

4. HAUSA SPEAKING TEST

4.1 MAJOR PROJECT ACTIVITIES

The day-to-day work of the project was conducted at the Center for Applied Linguistics (CAL) in Washington, DC. Charles W. Stansfield served as Project Director and Dorry Kenyon as Test Development Coordinator. A local test development committee was formed which included, in addition to the above, Mr. Daniel Kennedy, an experienced language test item writer and two experienced Hausa linguists: Dr. Beverly Mack (George Mason University) and Mr. Steven Lucas (USIA, Voice of America). Ms. Ruth Ephraim completed the local test development team as the artist for the test.

Three leading professors of Hausa in U.S. academic institutions served as members of an External Review Board: William R. Leben (Stanford University), Roxanna Ma Newman (Indiana University) and Russell G. Schuh (University of California, Los Angeles). These individuals reviewed draft test forms and provided feedback during the test development process by listening to examinee tapes of the trial version of the test forms.

The local test development committee met on a regular basis from November, 1988, to February, 1989, to develop the specific items for the two forms of the trial version of the test. These items were based on the question types used in the semi-direct tests of Chinese and Portuguese described above.

4.2 TRIALING OF THE TEST FORMS

The two preliminary forms of the HaST, after being reviewed by the members of the External Review Board and subsequently revised, were trialed on six individuals from the Washington, DC, area who had learned Hausa in a variety of ways, all with at least some experience in Hausaland (the area of Africa where Hausa is spoken). The purpose of the trial was to ensure that the questions were clear, understandable and working as intended, and to check the appropriateness of the pause times allotted on the tape for examinee responses. The subjects in the trialing took the test at

the Center for Applied Linguistics on an individual basis using two tape-recorders. When possible, Beverly Mack observed the examinee taking the test and made notes on his or her performance on a specially prepared questionnaire (see Appendix C-3). When this was not possible, she listened to the tapes and completed the form at a later date. Copies of the tapes were made and listened to by Steven Lucas, who also completed a form on each subject. In addition, each of the three members of the External Review Board listened to at least four of the tapes of examinee responses to the pilot version. Finally, upon completion of the test, examinees themselves responded to a detailed questionnaire about it (Appendix C-3). In most cases, they were also debriefed by Dorry Kenyon on their testing experience in person.

On the basis of data collected during the trialing and comments from the External Review Committee members, the final format of the test was slightly modified from the description of the test format given above. The questions were further focused on the Hausaland culture and "de-urbanized" as much as possible. Instead of a picture sequence question focussing on future narration, the third picture sequence focused on giving commands. As in the Indonesian Speaking Test, it was decided to leave out picture item number 3 (detailed description), as it was perceived to be more of a vocabulary exercise. In general, the level of the test was "toned down," i.e., more Intermediate level questions and fewer Advanced and Superior level questions were used, as it became apparent that few, if any, of the students for whom the test was intended would reach the Advanced, much less Superior, level. The opening conversation was recorded in two versions, one addressing a male examinee and one addressing a female examinee. Three questions in Hausa were included at the end of the tape as a type of "wind-down," a suggestion that arose from experience with the Portuguese Speaking Test.

Examples of the test questions on the final form of the test are available in the HaST Examinee Handbook, located in Appendix C-2.

Once the local test development committee completed revising the two forms of the HaST, the forms were again reviewed by the members of the External Review Board. After final revisions were made, test booklets and tapes for the validation study were prepared.

4.3 VALIDATION STUDY

It was hoped that a research study similar to the studies conducted for the CST and PST could be carried out to validate the two forms of the HaST. However, there were several obstacles preventing a replication of those studies for Hausa. First, there was no one trained and able to administer an Oral Proficiency Interview in Hausa up to the ACTFL standards. Russell Schuh, as a preliminary step towards certification, was working on certification in English-as-a-Second-Language, but had not yet completed all the steps in that process. None of the other Hausa linguists had yet begun the certification process. (Note: Hausa is not currently taught in any of the government agencies, so government trained and certified testers were not available.) Therefore, it was impossible to administer an OPI to the subjects involved. The second obstacle was a dearth of students qualified to take the exam, i.e., students at the Intermediate level. There were eight students scheduled to take the HaST at universities not affiliated with any of the members on the local or external test development committees. They had all completed at least two years of Hausa. In the end, several of them never finished the test, finding it too difficult for them. Only three of these university students completed both forms of the HaST and could be included in the sample. Thus, it was impossible to get a mix of ability levels for the study. The final sample included two of the subjects who had participated in the trialing of the test and several subjects who had learned Hausa through experience in the Peace Corps (though had no formal academic training in it). The total number of subjects participating in the validation study was thus 13. Because of the field of Hausa is so small, it was unavoidable that

some of the subjects were personally known to the raters.

Beverly Mack and Russell Schuh served as raters for the study. Because neither is certified, the results of the study should be seen as provisional. However, much was learned through the experience.

Most of the subjects were administered the HaST at the Center for Applied Linguistics using two tape recorders. Some of the subjects were administered the test at the language labs at their respective universities or by their Hausa instructors. Two of the subjects administered the taped tests to themselves at home using two tape recorders. The design controlled for order of administration, with half of the subjects receiving form A first and form B second, and the other half in reverse order.

Once all the tapes were collected, they were copied and sent to the raters for scoring. They were scored independently and in different order in sets of five. After each set of five tapes was scored, however, the two raters, without changing their original rating, compared their ratings and discussed divergent ratings. This was an aspect of self-training that was built into the design.

In order to answer the question of whether the test was doing what it was intended to do without having an OPI score with which to compare results, a special form was designed in order to get feedback from the raters on how well the test was eliciting a ratable speech sample. This form (see Appendix C-6) asked the raters not only to award a holistic rating on the entire performance, but to rate each examinee's performance on each individual item, and to award a score for the usefulness of the speech sample elicited by that particular item in making that examinee's holistic rating. In this way the test could be evaluated item by item as to its usefulness in eliciting a ratable speech sample, and each item could be evaluated as to its ability to draw out speech ratable at an examinee's proficiency level.

In the empirical analysis of the ratings, scores on the tape-based semi-direct tests were converted to a scale combining both ACTFL and ILR rating scales with weights assigned as follows:

ACTFL/ILR Level	Coded as:
Novice-Low	0.2
Novice-Mid	0.5
Novice-High	0.8
Intermediate-Low	1.0
Intermediate-Mid	1.5
Intermediate-High	1.8
Advanced	2.0
Advanced-Plus	2.8
Superior/Level 3	3.0
High-Superior/Levels 3+-5	3.8

The system of score coding above is based on the ILR 0 to 5 rating scale and is intended to assign an appropriate numerical value to the proficiency level descriptions. For example, proficiency at an Advanced-Plus level is characterized by many of the same features as at the Superior/3 level, though the examinee cannot sustain the performance. Thus, the numerical interpretation falls closer to 3.0 than mid-way between the two, as may be expected.

The several tables below provide descriptive statistics, interrater reliabilities and parallel-form reliability data obtained in the study. Rater 1 is Beverly Mack and Rater 2 is Russell Schuh.

Table 4.1 shows the mean score, standard deviation and other basic statistics for the ratings assigned by each of the two raters to subject performances on each of the semi-direct test forms and on the live interview.

Table 4.1

Descriptive Statistics for Scoring Levels Assigned

Test Form	Minimum Score	Maximum Score	Mean	Standard Deviation
Form A (n=13)				
Rater 1	0.2	2.8	1.54	0.75
Rater 2	0.5	2.8	1.53	0.65
Form B (n=13)				
Rater 1	0.2	3.0	1.61*	0.66
Rater 2	0.5	2.8	1.42	0.65

* The difference between these paired means is significant at the $p<.05$ level.

Table 4.2 shows the frequency of ratings given by the two raters on the two test forms (n=13).

Table 4.2
Frequency Distributions

Form A Ratings (Rater 1)

Rating	Frequency	Percent
0.2	1	7.69
0.8	1	7.69
1	3	23.08
1.5	2	15.38
1.8	3	23.08
2	1	7.69
2.8	2	15.38

Form A Ratings (Rater 2)

Rating	Frequency	Percent
0.5	1	7.69
0.8	3	23.08
1.5	2	15.38
1.8	4	30.77
2	2	15.38
2.8	1	7.69

Form B Ratings (Rater 1)

Rating	Frequency	Percent
0.2	1	7.69
1	2	15.38
1.5	3	23.08
1.8	4	30.77
2	2	15.38
3	1	7.69

Form B Ratings (Rater 2)

Rating	Frequency	Percent
0.5	2	15.38
0.8	1	7.69
1	2	15.38
1.5	3	23.08
1.8	3	23.08
2	1	7.69
2.8	1	7.69

Among other things, these statistics again show the difficulty of getting examinees at an appropriate level to take the Hausa

exam. Some examinees at times scored well below the suggested Intermediate Low level. There were also few examinees at a level above Intermediate. However, given the small size of the sample, there was quite a range in performances.

The mean scores for each rater on Form A were very similar, indicating that the raters were almost equal in their degree of severity. However, on Form B, Rater 1 appeared slightly more generous than Rater 2, as shown by their average mean ratings.

The degree of agreement between the absolute ratings may be seen from the following cross-tab diagrams. First, Table 4.3 presents the ratings of Rater 1 (down) against the ratings of Rater 2 (across) on Form A.

Table 4.3

Crosstabulations for Form A (n=13)

Rater 1 (down)/ / Rater 2 (across) Frequency	0.2	0.5	0.8	1	1.5	1.8	2	2.8	Total
0.2	0	0	1	0	0	0	0	0	1
0.5	0	0	0	0	0	0	0	0	0
0.8	0	1	0	0	0	0	0	0	1
1	0	0	2	0	1	0	0	0	3
1.5	0	0	0	0	1	1	0	0	2
1.8	0	0	0	0	0	3	0	0	3
2	0	0	0	0	0	0	1	0	1
2.8	0	0	0	0	0	0	1	1	2
Total	0	1	3	0	2	4	2	3	13

On Form A, there was total agreement in 46% of the ratings. Where there was disagreement, no consistent differences in rater severity are readily apparent; in four cases Rater 1 awarded the higher score while in three cases Rater 2 did. Only in one case was the disagreement greater than one step on the rating scale. One examinee was awarded a Novice-Low by Rater 1 and a Novice-High

by Rater 2. Thus, in 92% of the cases there was either complete agreement or a difference of one step on the scale. In only two cases did the disagreement cross an ACTFL level boundary.

Table 4.4 presents the ratings of Rater 1 (down) against Rater 2 (across) for HaST Form B.

Table 4.4

Crosstabulations of Form A Ratings (n=16)

Rater 1 (down)/ / Rater 2 (across) Frequency	0.2	0.5	0.8	1.0	1.5	1.8	2	2.8	3	Total
0.2	0	1	0	0	0	0	0	0	0	1
0.5	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	2
1.5	0	0	0	2	1	0	0	0	0	3
1.8	0	0	0	0	2	2	0	0	0	4
2	0	0	0	0	0	1	1	0	0	2
2.8	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	1
Total	0	2	1	2	3	3	1	2	3	13

From Table 4.4 we see that there was agreement in absolute ratings awarded in 31% of the Form B ratings. Again, only one of the ratings were more than a step away in disagreement; Rater 1 awarded one examinee a 1.0, while Rater 2 gave the performance a 0.5. Rater 1 appears to be more generous in this table. In all the 9 disagreements except one, Rater 1 gave the examinee the higher rating.

Interrater reliabilities (correlations) between the ratings assigned by Rater 1 and those assigned by Rater 2 for the two semi-direct test forms are shown in Table 4.5 below. Given the provisional nature of this study, it is useful to examine the correlations in two ways: first, in terms of their agreement on the

score awarded and second, in terms of their agreement in the relative ranking of the 13 examinees. The first correlation is given by the Pearson product-moment correlation coefficient. The second is given by the Spearman rank order correlation coefficient and is presented in the following tables in parentheses. The rank order correlations are not affected by disagreements in score, only by disagreements in rank.

Table 4.5
Interrater Reliabilities

Test Form	Correlation
A (n=13)	.88 (.95)
B (n=13)	.93 (.95)

These interrater reliabilities, both on the absolute scale and in terms of rank ordering, are quite high across both test forms. This suggests that the HaST does elicit a ratable sample of speech.

On performance-based tests such as the HaST, there is an increased concern for test-retest reliability. This form of reliability measures the degree of inconsistency in examinee performance on two separate administrations of the same test. The amount of inconsistency reflects the degree to which the test score may be confounded by such inconsistency. Therefore, it is important to examine this factor. However, on a test with a limited number of questions such as the HaST, it is not wise to administer the same test twice, since the first sitting will serve to instruct the examinee in the task at hand. (For a thorough discussion of this "reactivity effect" when conducting research, see Stansfield and Ross, 1988, p. 174.) Under such circumstances, it is preferable to administer different forms of the test while still using the same rater to score the performance. This type of reliability is known as parallel-form reliability, which is the degree of correlation between scores on two forms of the test.

both within and among themselves in severity), even under these severe conditions (different forms and different raters), the ability of the raters to place the examinees in very nearly the same rank order on the basis of the examinees' performance on the HaST is impressive.

As mentioned above, the HaST raters were asked to rate each item on each form as answered by each examinee in terms of its usefulness in making the holistic rating for that examinee. The rating scale for item usefulness ranged from 1 (lowest) to 5 (highest), with the midpoint (3) defined as adequate. There were 16 such ratings per form per examinee. The average rating given by the two raters for all the items on Form A was 3.27 and on Form B it was 3.15. These average ratings indicate that the items, in the opinion of the raters, were more than adequate in eliciting a ratable speech sample from the group of examinees in the validation study. On Form A, the highest average rating (3.5) was awarded to two items (Topic #2 and Topic #3). The lowest average rating on Form A (2.962) was awarded to Situation #4. This was the only item on Form A with an average rating below 3.0. On Form B, the highest average rating (3.35) was also awarded to Topic #2. The lowest average rating (2.89) was awarded to Situation #1. Of the 16 test items on Form B, only two received ratings below 3.0: Situation #1 (2.89) and Topic #4 (2.96). Thus, the vast majority of items on Form B were also rated as adequate or above to elicit a ratable speech sample.

Another way of examining whether the HaST as a measurement instrument is eliciting a ratable speech sample as intended is to examine how examinees at the different proficiency levels are functioning on the different items. As in the ACTFL oral proficiency interviewing procedure, each question on the Hausa test has an intended level of difficulty on the ACTFL scale. The HaST is composed of a variety of items at different levels of the ACTFL scale designed to probe the oral proficiency of the examinee: four of the items are at the Intermediate level, eight at the Advanced level, and two at the Superior level. (Note: the warm-up

conversation itself is a collection of items at various levels. However, in this study the raters were not asked to score performance on the items composing the conversation individually. Thus, the warm-up items are not included in the following analysis.)

- It can be hypothesized that if the HaST is functioning to probe the proficiency level of examinees in a way similar to the face-to-face interview, then lower level examinees will perform at a low level on all items, regardless of the item's placement on the scale, while higher level examinees should be distinguished from lower level examinees on all items, but particularly on items designed to probe the higher proficiency levels. On any item, examinees should not perform at a level higher than their holistic rating. However, higher level examinees may perform at a level lower than their holistic rating on items whose intended level is below their proficiency level. Thus, it was hypothesized that such easier items do not allow higher level examinees to demonstrate their full ability.

In order to test these hypotheses, the following post-hoc analysis was made on the HaST on the data collected from the raters. Ideally, it would have been best to have been able to divide the sample into groups of Intermediate, Advanced and Superior level subjects. However, as noted above, the sample who took the Hausa test was rather low in average proficiency. Thus, the thirteen examinees were divided into three groups on the basis of their holistic scores awarded as follows. Group 1 contained five individuals who across both HaST forms and across both raters had received holistic scores ranging between Novice-Low (0.2) and Intermediate-Mid (1.5). The mean score of group 1 members across raters and across forms was .87, nearest to a score of Novice-High on the ACTFL scale. Group 2 contained five individuals who had received holistic scores of Intermediate Mid (1.5) or Intermediate High (1.8). The mean score of this group across raters and forms was 1.70, nearest to a score of Intermediate-High on the ACTFL scale. Finally, group 3 contained three individuals who had

received holistic scores ranging from Intermediate High (1.8) to Superior (3.0). The mean score of this group across raters and forms was 2.42, about midway between Advanced and Advanced-Plus on the ACTFL scale.

To examine the hypothesis that higher level students would outperform lower level students on all types of items, it is necessary to examine the mean ratings by intended level of the item, by rater and by form. These mean ratings are given in Table 4.8.

Table 4.8
Mean Examinee Performance on Items
By Intended Item Level by Form and By Rater

Form A--Rater 1

Intended Item Level	Proficiency Group 1 (n=5)	2 (n=5)	3 (n=3)
Intermediate (n=4)	0.91	1.50	1.73
Advanced (n=8)	0.96	1.63	2.40
Superior (n=2)	0.88	1.49	2.60

Form A--Rater 2

Intended Item Level	Proficiency Group 1 (n=5)	2 (n=5)	3 (n=3)
Intermediate (n=4)	0.88	1.62	1.85
Advanced (n=8)	0.90	1.70	2.02
Superior (n=2)	0.88	1.69	2.40

Form B--Rater 1

Intended Item Level	Proficiency Group 1 (n=5)	2 (n=5)	3 (n=3)
Intermediate (n=4)	0.98	1.61	1.82
Advanced (n=8)	1.07	1.57	2.47
Superior (n=2)	0.90	1.62	2.67

Form B--Rater 2

Intended Item Level	Proficiency Group 1 (n=5)	2 (n=5)	3 (n=3)
Intermediate (n=4)	0.87	1.49	1.93
Advanced (n=8)	0.85	1.60	2.03
Superior (n=2)	0.83	1.68	2.50

Using Scheffe's test for the pairwise comparison of means in group performance by each intended item level (analyzed separately by rater and form), it was shown that only two of the 24 possible comparisons were NOT significant at the $p<.05$ level: Rater 1, Form A, Groups 2 and 3 on the Intermediate item level and Rater 1, Form B, Groups 2 and 3 again on the Intermediate item Level. This, however, could be expected because both groups are capable of functioning well at the Intermediate level and the nature of these easier items may not allow examinees at the two different levels enough chance to distinguish themselves in their oral performance. These results indicate that the individual items of the HaST are able to elicit representative speech samples in which more proficient examinees can distinguish themselves from less proficient examinees at the different item levels. In this respect, the test items appear to be working as intended.

To examine whether or not examinees performed at a higher level on items at an intended level higher than their holistic rating would indicate and if higher level examinees performed at a level lower than their holistic rating would indicate on items that are below their proficiency level, again Scheffe pairwise comparisons were made contrasting performance at various intended item levels holding the proficiency level group constant (analyzed separately by rater and form). This analyses compared the means in Table 4.8 under each proficiency group column read DOWN. The results indicate that there were no significant differences in means for either group 1 or group 2 across intended item levels for any rater on any form. Examinees within these groups performed the same whether the item was placed at the Intermediate, Advanced or Superior level. This again confirms the hypotheses that the items are working as intended when it is remembered that the members of group one averaged at the Novice-High level and those in Group 2 at the Intermediate-High level. On any type of item they never scored above their performance level. However, for both of the raters and for both of the forms, there were significant differences between performance on the various item levels for

Group 3. For Rater 1 on both Form A and Form B, performance on the Intermediate Level Items was significantly different from that on the Advanced Level Items and on the Superior Level Items. For Rater 2 on both Form A and Form B, performance on the Superior Level Items was significantly different from that on the Intermediate and Advanced Level Items. Remembering that the three members of Group 3 were holistically rated at the Advanced level, these results confirm the hypotheses that the HaST items are working as intended. Although (for Rater 1) the performance of members of groups 2 and 3 did not serve to distinguish themselves from each other on the items at the Intermediate level, their performances on the Advanced and Superior level items were such that the raters were able to distinguish group 2 members from group 3 members. Moreover, the performance of the members of Group 3 on the items at different levels was such that both raters on both forms distinguished the average rating of Group 3 members on the Intermediate items from their average rating on the Superior items, while they didn't distinguish between the performance of the members of Groups 1 and 2 on the two different item levels. These results indicate that the HaST items are working as level probes, and that the variety of item difficulties on the test are working to probe the examinee's ability to speak Hausa.

These results also provide some initial support for the validity of the ACTFL Proficiency Guidelines and the ILR skill level descriptions as a hierarchy of performance descriptions of second language learners' speaking ability. The items were written according to the content and functions described in the Guidelines/skill level descriptions. The fact that examinees were able to handle the content and functions in a way that matched the items and examinees' proficiency level suggests that the hierarchy of tasks included in the descriptions is valid, at least for this limited sample. If the Guidelines were without validity, then the higher level group in this study (with a mean holistic rating at the Advanced level) would not have performed any better on Superior level tasks than they did on Advanced or Intermediate level tasks,

which was not the case. On the other hand, the middle group (with a mean holistic rating at Intermediate-High) did perform equally well on all items, as they were not expected to exceed their holistic rating even on Advanced or Superior level tasks. The lowest group in this study (with a mean holistic rating at Novice-High) also performed consistently across Intermediate, Advanced and Superior level tasks, which would be expected.

Although it was not possible to administer an OPI for Hausa, it should be noted the HaST is one of a family of five semi-direct tests developed by CAL that all have a similar format and were developed in a similar manner. Thus, if an OPI could have been administered, it would not have been surprising if its correlation to the HaST would have been similar to those obtained in the validation studies on the other tests in the family. These are given in Table 4.9, which presents the average correlations between the OPI as administered by trained and/or certified ACTFL/ILR testers and the corresponding semi-direct tests in each case.

Table 4.9
Average OPI/Semi-direct Test Correlation*

Chinese Speaking Test	.93	(32 examinees)
Portuguese Speaking Test	.93	(30 examinees)
Hebrew Speaking Test (USA Study)	.93	(20 examinees)
Hebrew Speaking Test (Israeli Study)	.89	(20 examinees)
Indonesian Speaking Test	.95	(16 examinees)

*across all raters, examinees and forms

The figures above take into account both same rater and different rater scoring. Their magnitude and consistency across languages and testing conditions strongly support the criterion-related and construct validity of this family of semi-direct tests. It may be argued that if it had been possible to undertake a study correlating examinee scores on the OPI with their scores on the HaST, a similar high correlation between them would have resulted. The consistent evidence from other similar CAL tests supports the use of the HaST as a measure of oral proficiency and as an

alternative to the OPI.

4.4 SUBJECT RESPONSE TO THE TEST

As part of the validation study, feedback information from the participants on various aspects of their experience with and opinions about the HaST were elicited by means of a short questionnaire (Appendix C-4). The questionnaire was given to the subjects directly after completing the semi-direct tests and completed and returned to CAL. All subjects completed the questionnaire for a 100% participation rate.

The answers to the examinee questionnaires are given in graphic summary form below. Written comments in response to the questionnaire are presented in Appendix C-5.

The first two questions sought to elicit from the subjects the extent to which they felt their Hausa speaking ability had been adequately and fairly probed by the HaST.

(1) Over the course of the taped test, do you feel that the descriptions, narratives, situations, and other types of questions in the test were adequate to probe your maximum level of speaking ability in Hausa?

(2) In the taped tests, were there any picture/descriptions, narratives, situations, or other questions that you felt were in any way 'unfair'?

11 of the 13 subjects (85%) responded affirmatively to the first question and negatively to the second (see Figures 4.1 and 4.2), which indicates that they were satisfied with the ability of the HaST to test their level of speaking ability in Hausa in an adequate and fair manner.

1. Did you feel the questions adequately probed your ability to speak Hausa?

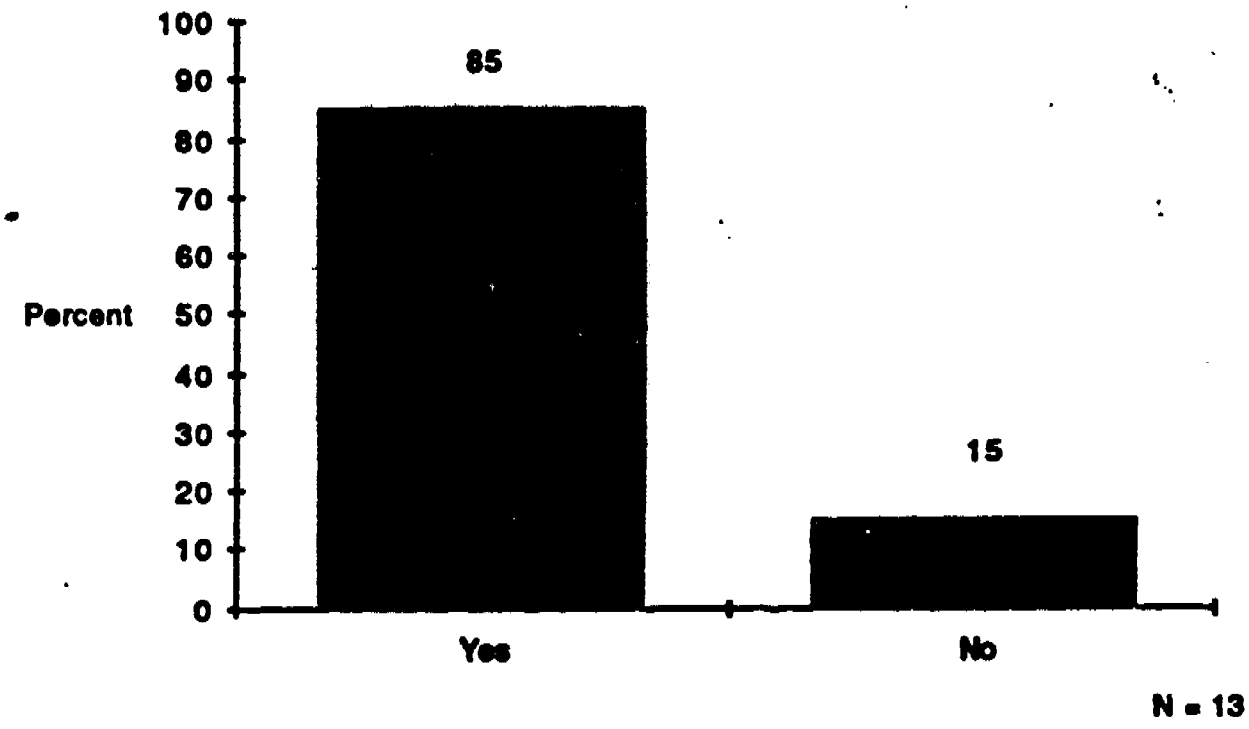


Figure 4.1

2. In the taped tests, were there any picture/descriptions, narratives, situations or other questions you felt were in any way "unfair"?

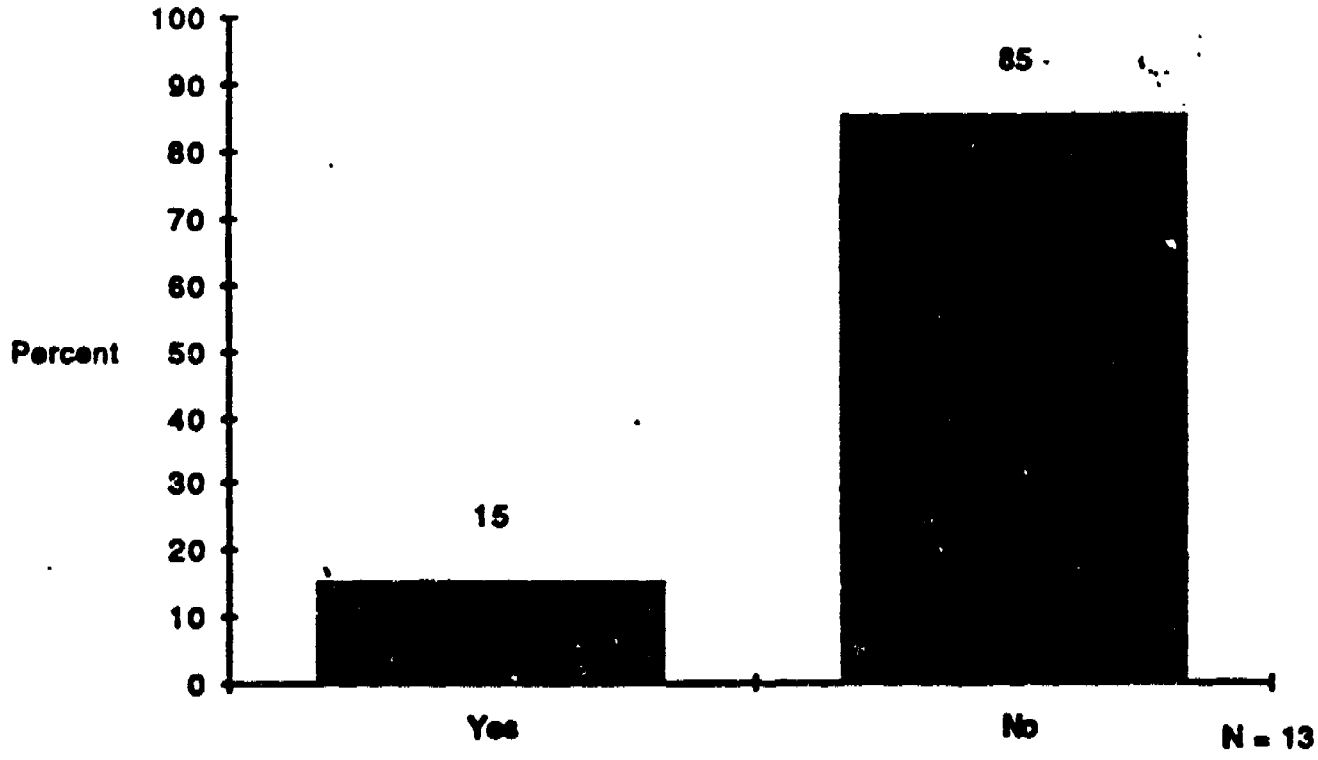


Figure 4.2

The next question focused on their affective feelings towards the testing experience.

(3) Did you feel unduly nervous in the taped tests?

A small majority (54%) reported feeling unduly nervous during the testing (Figure 4.3). This is not surprising for two reasons: one, the test was above the actual proficiency level of some of the subjects in the groups and two, the semi-direct mode of testing may be unfamiliar and perhaps 'unnatural' to students in general.

Questions 4 and 5 focused on technical qualities of the taped test.

(4) In the taped test, were the pauses for your responses usually long enough for you to respond as fully as you wished or were able?

(5) Where the directions on the taped test clear?

12 of the 13 subjects (92%) felt the timed pauses were in general usually about right (Figure 4.4) and 100% felt the directions were clear (Figure 4.5). These positive results are indicative that the technical quality of the test is satisfactory. Since the testing situation is controlled by the taped instructions and timed pauses, i.e., there is no possibility for examinees to ask questions or stop the tape once Part One of the test is begun, it is important that the directions be clear and timed pauses be appropriate.

The last question asked whether the examinees felt the two tests were of equal difficulty.

(6) Do you feel that the two taped tests were of the same difficulty?

A large majority (77%) answered that the two tests were equally difficult (Figure 4.6). This is important as the tests were designed to be alternate forms. The one comment written by an examinee who answered negatively (see Appendix C-5) reveals that the individual held no strong feeling about them being not equal in difficulty.

3. Did you feel unduly nervous in the taped tests?

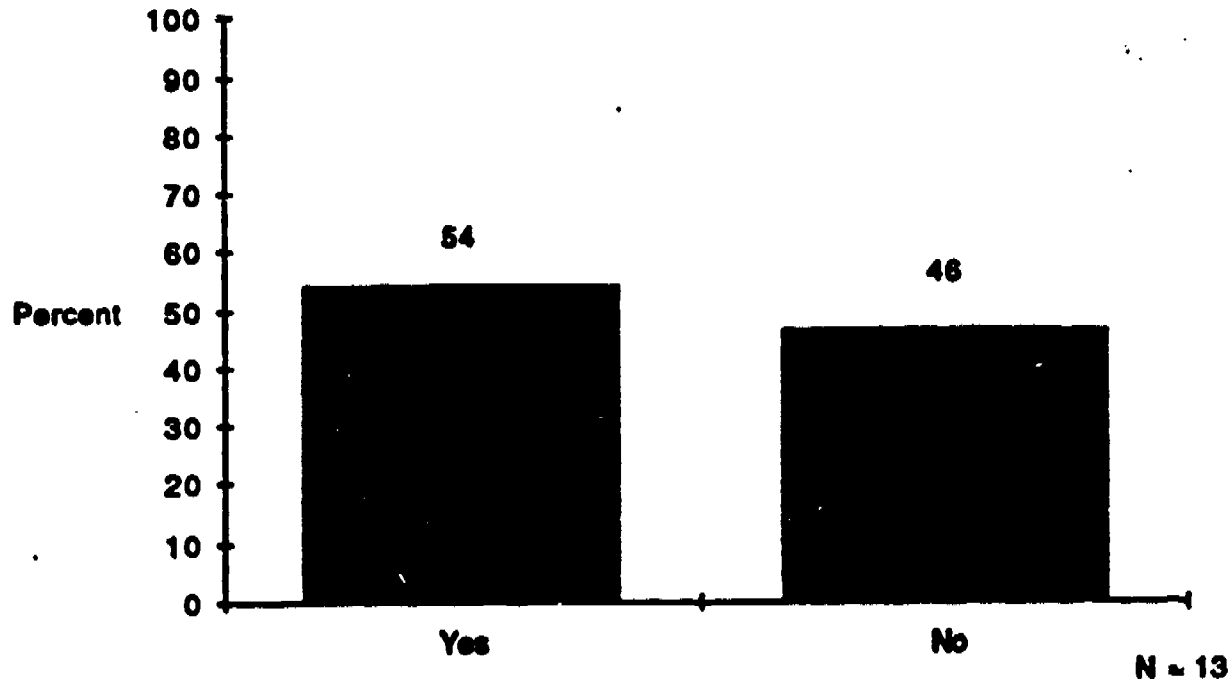


Figure 4.3

4. Were the pauses for your responses usually long enough for you to respond as fully as you wished or were able?

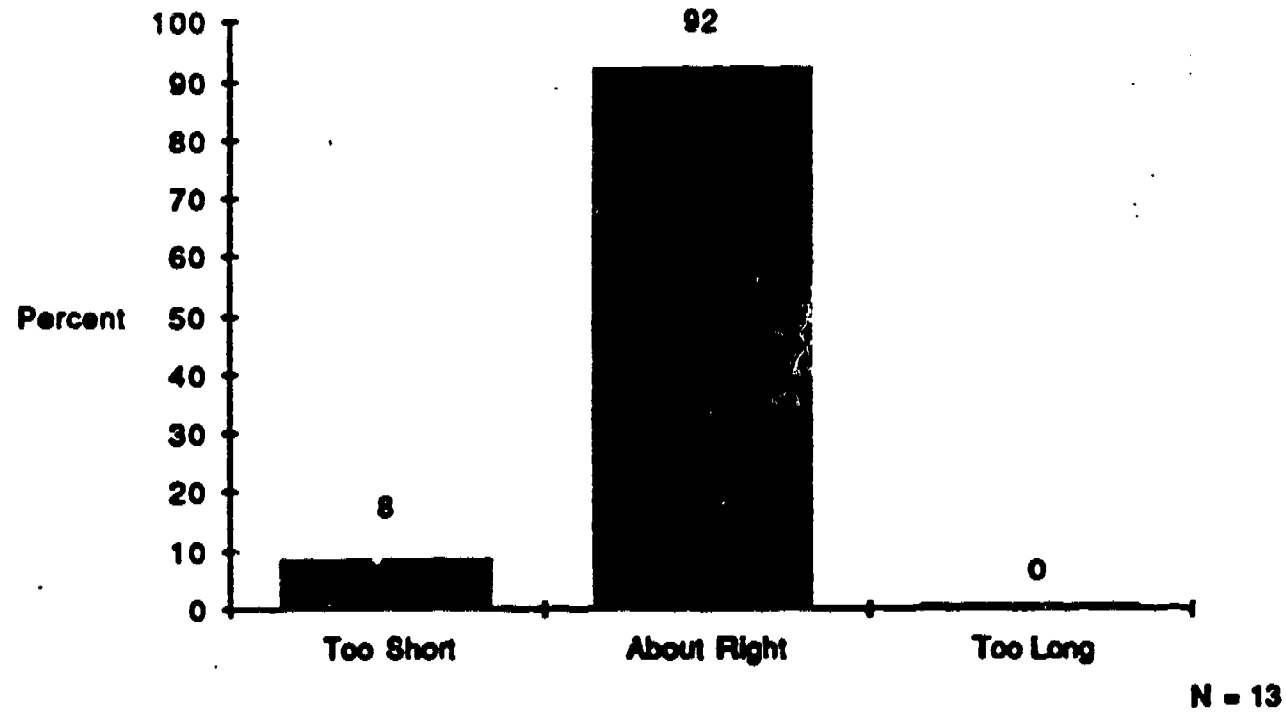


Figure 4.4

5. Were the directions on the taped test clear?

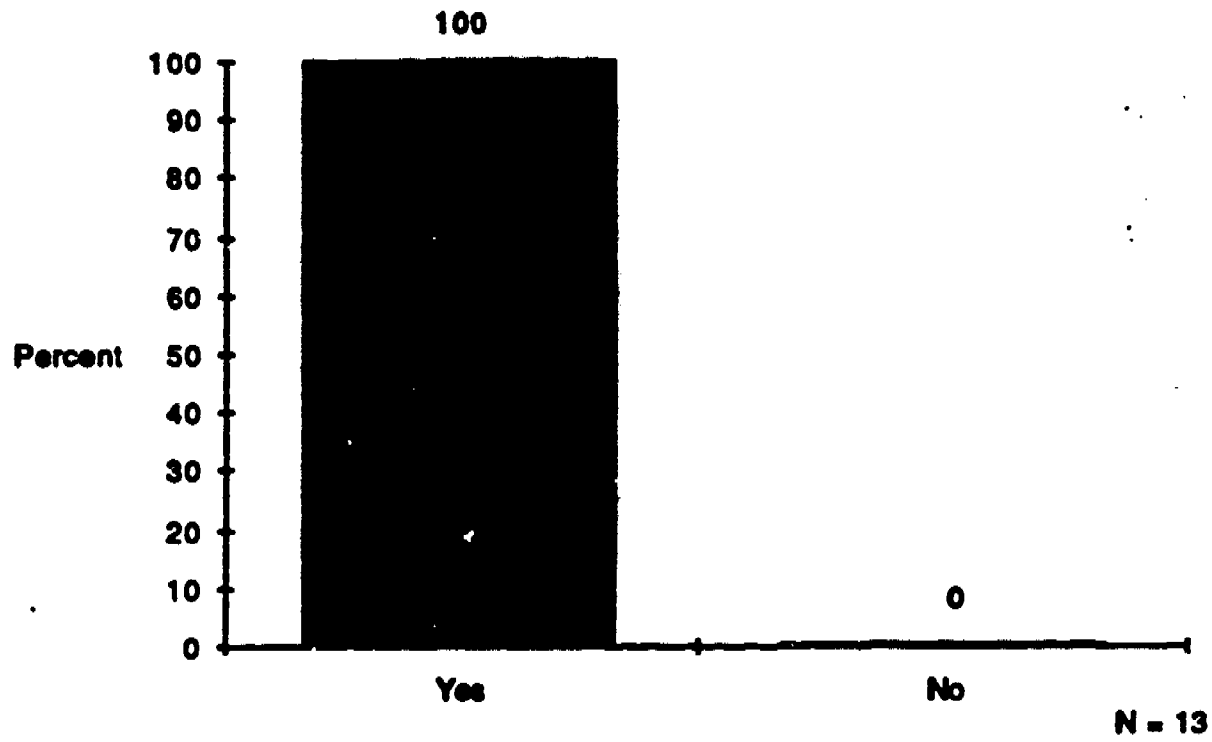


Figure 4.5

6. Do you feel that the two taped tests were of the same difficulty?

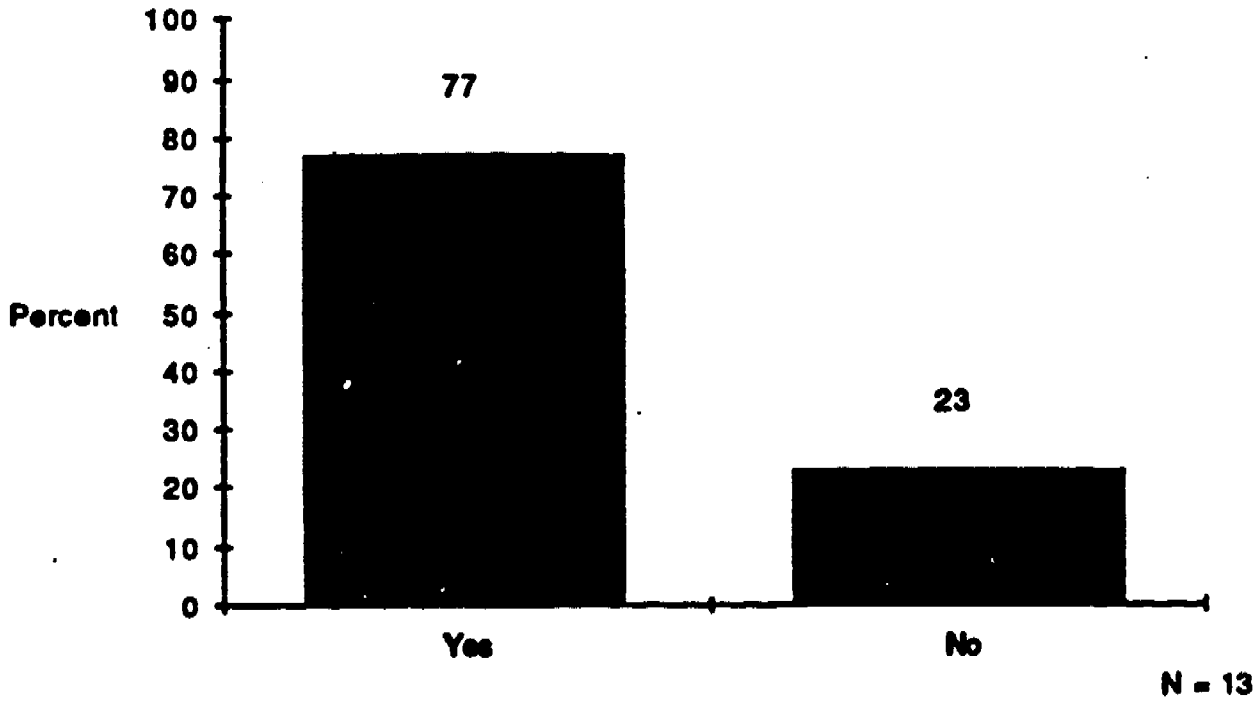


Figure 4.6

In summary, examinee reaction to the HaST was very positive. From the examinee's point of view the HaST appears to be probing Hausa speaking ability fairly and adequately and is technically sound.

4.5 OPERATIONALIZATION OF THE TESTS

To operationalize the test, a supply of tests were professionally printed: 100 copies of each test form were printed. In addition, 50 copies of each form of the Master Test Tape were copied, in both the male and the female versions.

A Test Manual, giving complete information on the development, uses, and administration of the HaST as well as the interpretation of examinee scores was prepared and is included as Appendix C-1. An Examinee Handbook was also prepared to be distributed to HaST examinees before taking the test and is found in Appendix C-2. The two booklets above establish and explain in detail the procedures for ordering and handling the test in-house. They also contain registration and order forms that are used in the operationalization of the test.

Announcements of the availability of the test are being produced to be sent to Hausa and African Language Departments and other interested parties throughout the country.

References

Clark, J.L.D. (1986). Handbook for the development of tape-mediated, ACTFL/ILR scale-based tests of speaking proficiency in the less commonly taught languages. Washington, DC: Center for Applied Linguistics.

Clark, J.L.D. and Li, Y.-C. (1986). Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages. Washington, DC: Center for Applied Linguistics. (Alexandria, VA: ERIC Document Reproduction Service No. ED 278 264)

Stansfield, C.W. & Kenyon, D.M. (1988). Development of the Portuguese Speaking Test. Washington, DC: Center for Applied Linguistics. (Alexandria, VA: ERIC Document Reproduction Service No. ED 296 586)

Stansfield, C.W. and Ross, J. (1988). A long-term research agenda for the Test of Written English. Language Testing, 5(2), 160-186.