DOCUMENT RESUME

ED 243 934 TM 840 247

AUTHOR Nevo, David; Shohamy, Elana
TITLE Applying the Joint Committee's Evaluation Standards for the Assessment of Alternative Testing Methods.
PUB DATE Apr 84
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communicative Competence (Languages); Foreign Countries; Group Discussion; High Schools; Interviews; *Language Proficiency; *Language Tests; Measurement Techniques; *Oral Language; Role Playing; *Standards; Test Selection
IDENTIFIERS Israel (Tel Aviv); *Standards for Evaluation Educ Prog Proj Materials

ABSTRACT

This study is based on three sources: (1) an experimental try-out of four oral proficiency testing methods (oral interview, role play, reporting test, and group discussion); (2) an evaluation of the testing methods by a panel of experts; and (3) an analysis of the same testing methods by policy makers. The findings are reported in terms of the Joint Committee's "Standards for Evaluation of Educational Programs, Projects and Materials" (Utility, Accuracy, Feasibility, and Fairness), modified to apply to tests. Decision makers were interested in the information regarding the Utility and Feasibility of the various tests, and did not limit their interest to Accuracy. Comparing the results obtained from the experimental try-out study and the rankings provided by the panel of experts suggests that testing experts seem to be better in judging a test by some standards than by others. The findings also suggest that testing experts should not limit themselves to the technical aspects of Accuracy, but instead use the wide scope of all four standards to judge the merit of a test. (BW)

APPLYING THE JOINT COMMITTEE'S EVALUATION STANDARDS FOR THE ASSESSMENT OF ALTERNATIVE TESTING METHODS

DAVID NEVO
Tel-Aviv University

ELANA SHOHAMY
Tel-Aviv University

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, April 23-27, 1984.

The Standards for Evaluation of Educational Programs, Projects and Materials (Joint Committee, 1981) were developed during a period of four years and published in 1981 by a joint committee of 17 members representing 12 organizations associated with educational evaluation. These standards were developed in response to a recommendation appearing in the 1974 APA Standards for Educational and Psychological Tests (APA, 1974). They represent an extension not only from tests to program evaluations but also an extension from a narrow scope of concern for reliability and validity into a wide perspective on evaluation (Nevo, 1983), and evaluation standards. They focus on four major groups of standards: utility, feasibility, propriety and accuracy. It seemed reasonable to apply these four groups of standards also to testing methods, and not to limit their use to evaluations of projects and programs. Such an application could provide a wider basis for the development of a comprehensive set of standards for educational as well as psychological tests.

The Joint Committee's 30 standards for evaluation of programs, projects and materials were used to develop 23 standards for testing methods. Parallel to the Joint Committee's standards, they were organized in four groups of standards: Utility, Accuracy, Feasibility and Fairness. Following is a description of these newly-developed groups of standards:

STANDARDS FOR EDUCATIONAL TESTING METHODS

A. Utility Standards

The Utility Standards are intended to ensure that a testing method will serve the practical information needs of given audiences. These standards are:

A-1 Audience Identification

Audiences involved in or affected by the testing should be identified, so that their needs can be addressed.

A-2 Tester Credibility

The persons conducting the testing should be both trustworthy and competent to perform the testing, so that their findings achieve maximum credibility and acceptance.

A-3 Information Scope

Information collected by the test(s) should be of such scope as to address pertinent questions about students' achievements and be responsive to the information needs and interests of specified audiences.

A-4 Justified Criteria

Criteria used to determine test scores and marks are clearly described and justified.

A-5 Report Clarity

Testing results are presented in forms readily understood by identified audiences.

A-6 Report Dissemination

Testing results are disseminated to all relevant audiences, so that they can assess and use the findings.

A-7 Report Timeliness

Release of testing results should be timely, so that audiences can best use them.

A-8 Evaluation Impact

Testing has a positive impact on the teaching and learning process and on the decision making processes of all parties associated with the testing.

B. Accuracy Standards

The Accuracy Standards are intended to ensure that a testing method will reveal and convey technically adequate information

on the educational achievements of those that are being tested.

These standards are:

B-1 Valid Measurement

Testing is conducted by instruments and procedures providing valid information for a given use.

B-2 Reliable Measurement

Testing is conducted by instruments and procedures providing reliable information for a given use.

B-3 Testing Conditions

Testing conditions are described in enough detail, so that their adequacy can be assessed and considered when assessing the achievements of each student.

B-4 Test Security

Test materials and testing procedures are safeguarded to avoid fraud and cheating.

B-5 Data Analysis

Testing data are appropriately and systematically analyzed to ensure supportable interpretations of test scores.

B-6 Objective Reporting

Test results are reported objectively without distortion by personal feelings and biases of testers and scorers.

C. Feasibility Standards

The Feasibility Standards are intended to ensure that a testing method will be realistic, prudent and frugal. These standards are:

C-1 Practical Procedures

Testing is conducted with minimum disruption of educational and administrative processes at school and with consideration of existing constraints.

C-2 Political Viability

Testing is planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained.

C-3 Cost Effectiveness

Testing does produce information of sufficient value to justify the resources expended.

D. Fairness Standards

The Fairness Standards are intended to ensure that a testing method is conducted legally, ethically, and with due regard to the welfare of tested individuals as well as those affected by test results. These standards are:

D-1 Accepted Criteria

Tests are based on known and accepted subject matter and criteria.

D-2 Rights of Human Subjects

Testing is designed and conducted, so that rights and welfare of human subjects are respected and protected.

D-3 Public's Right to Know

The public's right to know the results of testing and its consequences is respected within the limits of other related principles such as those dealing with public safety and the right to privacy.

D-4 Conflict of Interest

Conflict of interest, frequently unavoidable, is dealt with openly and honestly, so that it does not compromise the testing process and results.

D-5 Social Values

Testing is conducted in accord with social values and does not stimulate violation of norms and values accepted at school or society.

D-6 Balanced Reporting

Test results are complete and fair in their presentation of strengths and weaknesses of the individual tested.

The purpose of this study was to test the validity and applicability of the newly developed standards. They were applied to assess four alternative testing methods of oral proficiency in English as a Foreign Language (EFL). The four testing methods were: an oral interview, a role play, a reporting task, and a group discussion test. These methods had to be assessed to develop a recommendation for the Ministry of Education in Israel regarding the adaption of an appropriate procedure to test oral proficiency in English as a Foreign Language within the matriculation exams administered at the end of High School to all students. It was apparent that such a decision could not be limited to validity and reliability, and a wide scope of decision criteria had to be used for this purpose. Thus the extensive scope of the Joint Committee's Standards seemed to be a plausible approach to this problem.

Before proceeding with the study design and its findings, a short discussion of testing methods of oral proficiency, on which this study focused, will be presented.

ALTERNATIVE TESTING METHODS OF EFL ORAL PORFICIENCY

The increased interest in the teaching of the communicative skills has brought about greater emphasis on both the teaching and the testing of oral proficiency. Yet, oral performance in communicative situations is one of the most difficult skills to assess. Although in the past decade several attempts have been made to

develop tests that would provide better measures of oral proficiency (Madsen & Jones, 1981), the research carried out on these tests is still very limited.

Currently in Israel EFL oral proficiency is tested within the framework of the high-school leaving examination ("The Matriculation Exam") administered nationally by the Ministry of Education. The testing procedure is a conversation in which a tester interviews each student individually. Students' performance on that test provides the basis for the oral proficiency score. Several deficiencies seem to be found with this procedure: (a) The oral interview test represents a narrow domain of oral performance, and it is therefore questionable whether it is a valid indication of students' overall oral proficiency. (b) Since rater reliability is not assessed, it is questionable whether the score obtained by the student is his "true score", especially since the testers are not trained in either interviewing techniques or rating oral proficiency. (c) The test has very low variance, and relatively high scores; literally nobody fails the test. This has caused some officials in the Ministry of Education to call for the abolishment of the oral test, since it provides little information compared to its cost.

Searching for an alternative to the existing system constitutes a problem, since among the tests available hardly any have been sufficiently researched to allow their implementation on a nation-wide scale. The only oral test that has been researched

extensively is the FSI (Foreign Service Institute) Oral Interview (Bachman & Palmer, 1981; Shohamy, 1982; Hinofotis, 1976; Clifford, 1977). However, one of the major shortcomings of that test is that it does not encompass a wide enough variety of speech styles (Shohamy, 1983)--it is limited to questions asked by the tester and answers supplied by the test-taker. It is obvious that such a test is not comprehensive enough to assess all the aspects of oral proficiency.

Thus, an attempt was made to develop a comprehensive battery of oral proficiency tests, representing four different speech styles. Following is a description of these four tests:

(a) The Oral Interview (OI):

The rationale underlying this test was to guide the test-taker into a dialogue with the tester which elicits answers to questions asked by the tester on different topics. The test encompassed a variety of topics and represented low-high role relationship between the participants. The test followed the model of the FSI Oral Interview (Lowe, 1981; Oller, 1981), where the test-taker is pushed to the highest level of his oral proficiency. The test consisted of four phases: (1) warm-up, where the test-taker was put at ease and the tester derived a preliminary indication of the test-taker's level of proficiency; (2) level-check, where the tester checked the functions and content which the test-taker could perform most accurately; (3) probing, where the tester assessed the highest level at which the test-taker could function accurately; and (4) wind-up, where the test-taker was returned to the level at which he could function most comfortably. The scoring of the oral interview was done on the basis of the same rating scale used for all the other tests.

(b) The Role Play (RPL)

The rationale behind this test was to stimulate the test-taker to produce spontaneous speech-behavior within the

limits of a pseudo-authentic situation. This test was a dialogue between two participants who represented various role-relationships between the speakers (equal, low-high, high-low), and the level of formality varied as required by the specific simulated situation. The test-taker was given a card on which he found the description of a situation and his expected role in it. The tester then engaged in the simulated conversation derived from the situation. The test lasted for about ten minutes, and a score was assigned by an assessor who was not involved in the RPI, on the basis of the same rating scale used for all the other tests.

(c) The Reporting Test (REP):

The rationale underlying this test was to stimulate the test-taker into a monologue in English based on authentic input in the mother tongue, Hebrew. This test required a unilateral skill of communication. The role relationship between the speaker and the listener was low to high, and the occasion was formal. The functions involved in the test were conveying facts, explaining and reporting. The student was given an article in Hebrew which he was asked to read silently, and then to report its general content in his own words. He was asked not to translate the text but rather to report freely, referring back to the text only if necessary. The test lasted about 10 minutes and was scored on the basis of the same rating scale used to rate oral proficiency for all the other tests.

(d) The Group Discussion (GD):

The rationale underlying this test was to stimulate the test-takers into a spontaneous discussion of a controversial issue, in which they could express views about topical matters, debate and argue over them, defend their opinions and try to persuade the other participants to accept them. This test required multilateral communication and the role relationship among the participants was equal. Four students were asked to discuss a topical subject or issue controversial enough to lend itself to a lively discussion. Members of the group picked a card randomly from among 20 cards which provided information regarding the topic of the discussion they were about to conduct. They were given a few minutes to read the card and plan the procedure of their discussion among themselves before starting the actual discussion (Reves, 1982). The tester listened to the discussion without interfering and scored the performance of each of the four test-takers on the basis of the rating scale used for all the other tests.

THE STUDY DESIGN

The study reported in this paper is based on three sources: (a) an experimental try-out of the four testing methods, (b) an evaluation of the testing methods by a panel of experts, and (c) an analysis of the same testing methods by policy makers. Following is a short description of each of the three sources.

(a) The experimental try-out

The four alternative testing methods of oral proficiency were tried out with a sample of 103 twelfth grade students of four classes in a comprehensive high school north of Tel Aviv*. The classes were randomly selected out of seven classes in that school. All students took all four tests. The tests were administered independently and lasted for about ten minutes each. To minimize the learning effect possibly created by the order of the tests, groups of students took the tests in various order, so that total rotation was ensured. The testers who were assigned to administer the tests were experienced EFL teachers who were trained in administering and rating the different tests. The rating of students' performance was done on the spot using a rating scale adapted from Clifford and Lowe (1981). It rated oral proficiency on a scale ranging from 4 to 10**, 10 being equivalent to native speaker's performance. The tests were all taped to allow for an additional rating in order to compute rater reliability.

---

*For a detailed description of this study see Shohamy, Reves, & Bejarano (1984).

**The scale of 4 to 10 is the conventional scale regularly used in the Israeli School System.

On completion of each of the four tests the students filled out a questionnaire which assessed their attitudes toward the four experimental tests.

Two weeks after the administration of the four tests 77 of the 103 students were tested by the existing conventional test (The Matriculation Exam) in their schools. It is on these 77 cases that the comparison between the experimental tests and the exsiting test was done.

(b) Evaluation by experts

A group of sixteen language testing experts, attending a convention on research on language testing, were presented with a detailed description of the four oral proficiency tests as well as some research findings. This group of experts had been previously exposed to the Standards for Educational Testing Methods presented in this paper. Following the discussion of the four testing methods, the experts were asked to rank each method according to its Accuracy, Utility, Feasibility and Fairness. The ranking was done individually using a four point scale from "1" (high) to "4" (low) for each standard. The ranking form also included a one-sentence definition of each standard, as a reminder to the experts. On the basis of those rankings an overall average rank was computed for each standard regarding each testing method.

(c) Analysis by policy makers

Since the four testing methods were considered by the Ministry of Education as a possible alternative for the existing oral test

of the Matriculation Examination, several discussions were held at the Ministry regarding this issue. Senior administrators, associated with EFL instruction and testing, and the developers of the four alternative testing methods, participated in those discussions. As a result of the discussions a decision was made to further experiment with an integrated version of the four testing methods with a sample of 1000 12th grade students to substitute for the existing Matriculation oral test. We use these discussions as a case study to demonstrate some interesting points related to the process in which policy makers use evaluative information to assess the merit of alternative testing methods.

RESULTS

We shall report our findings for each group of standards regarding the four testing methods on the basis of the relevant data obtained from the three sources of our study.

(a) Utility standards

The utility standards are intended to ensure that a testing method will serve the practical information needs of given audiences to have a positive impact on the teaching and learning process as well as on the decision making process of those associated with the testing and its results.

As can be seen in Table 1, the group of language testing experts ranked the Group Discussion (GD) test being the one with the highest utility value among the four testing methods. This high rank was justified by some of the experts with the positive back-

wash effect that the GD test might have on instruction, stimulating teachers to allocate time for discussion in their classes. The group of policy makers considered also the back-wash effect of the Group Discussion test as an important feature of this test and decided to support its possible use in the future in spite of some logistic difficulties associated with its administration and its relatively low accuracy qualities.

Insert Table 1 about here

It is interesting to note that while the GD test has been ranked highest on Utility, it has been ranked lowest for Accuracy. At the same time the experts ranked the Oral Interview (OI) test quite low (3) on Utility, in spite of the fact that it was ranked highest on all other three standards.

(b) Accuracy standards

The Accuracy standards are intended to ensure that a testing method will reveal and convey valid, reliable and otherwise technically adequate information on educational achievements. The experimental try-out as well as the evaluation by the language testing experts provided information on the accuracy of the four testing methods included in our study.

One of the concerns of the Ministry of Education regarding the existing Matriculation oral proficiency test was related to its relatively high scores and their low dispersion. Some of the opponents of those tests argued that "since almost every student gets

anyhow a high score on this test why waste on it so much time and effort." And indeed for the students, who participated in the experimental try-out and took also the existing Matriculation proficiency test, a mean score of 7.79 and a standard deviation of 1.03 were obtained on this test (see Table 2).

Insert Table 2 about here

Compared to the existing test, lower mean scores and higher standard deviations were obtained for all four experimental tests. Among them the Group Discussion test seemed to have the lowest mean score ($\bar{X}$=6.00) with the highest standard deviation (S.D.=1.93). The lowest standard deviation (S.D.=1.32) was obtained for the Reporting test. Considering the relationship between variability and reliability, the data on the standard deviations of the four tests could be some kind of reflection of the reliability of the tests. Mainly, reliability associated with errors of measurement that apply to the test content itself rather than the biases of the scorers.

More direct information on the tests' reliability can be obtained from the findings for the inter-rater reliability. As can be seen in Table 2, the highest inter-rater reliability was obtained for the Oral Interview (r=.91). The inter-rater reliability for the Reporting test was r=.81, and for the Role-Play it was r=.76.*

The ranking of the tests by the level of their inter-rater reliability seems to be in general agreement with the overall ranking

*The inter-rater reliability for the Group Discussion test has not been computed yet at the present time.

for accuracy provided by the panel of experts (see Table 1). They also ranked the Oral Interview highest on Accuracy; second highest they ranked the Reporting test; third--the Role Play, and the Group Discussion test was ranked lowest on this standard. If we consider the findings on inter-rater reliability as a valid indication of one aspect of test accuracy, it is apparent that there is agreement between the data obtained from the experimental field try-out and the judgments provided by a panel of experts.

(c) Feasibility standards

Administering the four tests within the framework of the experimental try-out, suggested that all of them can be implemented as feasible testing methods to test oral proficiency without any major difficultires. The testers, who were in most cases regular EFL high school teachers, went through a relatively short and simple training process, and succeeded in completing each test in ten minutes per student. Regarding the feasibility of implementing these tests, there seemed to be no apparent advantage for any single test, except for the Group Discussion test which did create some difficulties in reaching uniform procedures among testers and overcoming some logistic problems in coordinating group testing sessions for students who took all other tests on an individual basis.

The students participating in the experimental try-out seemed to be enjoying the testing experience, although in their questionnaires they showed some preference for the OI and the RPL tests.

The panel of experts (see Table 1) ranked the Oral Interview

highest on feasibility and the Role Play test as lowest. The second lowest on feasibility they ranked the Group Discussion test, as was also indicated by the experience gained from the experimental try-out.

The policy makers expressed concern regarding the feasibility of introducing the newly developed tests into the system in terms of cost, testing time and training of testers. They were especially concerned about the logistics of administering the Group Discussion test in conjunction with the other tests administered on an individual basis.

(d) Fairness Standards

Two major sources of information were available in this study regarding the fairness of the four testing methods; the student questionnaire from the experimental try-out and the rankings provided by the panel of lanugage testing experts (Table 1). At the end of each test students filled out a questionnaire in which they were asked to agree or disagree with a set of statements expressing their attitude toward the test. One of those statements was "This test reflected my true knowledge in speaking English." Students' responses to this statement are presented in Table 3.

Insert Table 3 about here

If we consider this statement as a possible expression of test fairness, we can see in Table 3 that the Oral Interview was perceived by students as the fairest opportunity to express their

knowledge in speaking English. Almost 85 percent of the students agreed (or strongly agreed) with the statement and its mean rating was $\bar{X}=3.00$. The second best for Fairness came out the Role Play test for which more than 70 percent of the students agreed that it reflected their true knowledge in speaking English. Students' opinions seemed to be balanced on the Group Discussion test but were somewhat negative regarding the Role Play test. More than 60 percent of them did not think that this test reflected their true knowledge in speaking English.

Using the mean level of students' agreement with the statement for each test, we could rank the four tests for Fairness from high to low as follows: Oral Interview, Role Play, Reporting, and Group Discussion. If we compare Table 3 with Table 1, we will find that students' perceptions on tests' level of fairness differ from those of the testing experts, except for the Oral Interview. Both groups ranked this test highest on Fairness but strongly disagreed on the Role Play test. This test was considered as second best by students but was ranked lowest by the experts (see Table 1). Unfortunately, testing experts do not consult students whenever they are asked to rate the Fairness of a test.

SUMMARY AND DISCUSSION

Our study demonstrated that the Joint Committee's Standards could be adopted for testing methods and used as a framework to analyze and assess the merit of alternative testing methods. Being conducted in a context of a real decision making process, this study

showed that such a framework provides a wide scope of information relevant to decision makers. Decision makers were interested in the information regarding the Utility and Feasibility of the various tests, and did not limit their interest to Accuracy, when they considered the introduction of the newly developed tests into the educational system.

The rankings of the testing experts provided a clear distinction between the qualities of the four tests according to the various Standards. One example was the OI test which was ranked highest on all Standards except Utility. Another example was the GD test, which was ranked highest on Utility but lowest on Accuracy. These findings suggest that testing experts should not limit themselves to the technical aspects of Accuracy, and use the wide scope of all four Standards to judge the merit of a test.

Comparing the results obtained from the experimental try-out study and the rankings provided by the panel of experts, suggests that testing experts seem to be better in judging a test by one standard than another. Their assessments of the accuracy of the four oral proficiency tests were in strong agreement with the findings obtained from the experimental try-out regarding inter-rater reliability of these tests. At the same time there was a lack of agreement between experts ranking on Fairness and students' perceptions of test fairness as expressed in their questionnaires.

Although the study provided some interesting observations regarding the applicability of the Standards to the assessment of

testing methods, it was based mainly on secondary sources of information and provided only a partial attempt to study the whole scope of the Standards. More systematic efforts in this direction should be encouraged in the future.

Table 1

Average Rankings of Tests by Experts according to the Four Standards

Test	Standard			
	Utility	Accuracy	Feasibility	Fairness
Oral Interview (OI)	3	1	1	1
Role Play (RLP)	2	3	4	4
Reporting (REP)	4	2	2	3
Group Discussion (GD)	1	4	3	2

1 = High 4 = Low

Table 2

Mean Scores, Standard Deviations and Inter-rater Reliability of Oral Proficiency Tests

Test	Mean*	S.D.	Inter-rater Reliability**
Oral Interview (OI)	6.49	1.39	.91
Role Play (RLP)	6.17	1.51	.76
Reporting (REP)	6.57	1.32	.81
Group Discussion (GD)	6.00	1.93	--
Existing Matriculation Test	7.79	1.03	--

* n = 103 ** n = 25

Table 3

Distribution of Students' Responses to Statement "This test reflected my true knowledge in speaking English" by Test (in percentage)

Test	Strongly agree (4)	Agree (3)	Disagree (2)	Strongly disagree (1)	$\bar{X}$	Overall rank
Oral Interview (OI)	17.5	66.0	14.6	1.9	3.00	1
Role Play (RLP)	9.8	62.7	26.5	1.0	2.81	2
Reporting (REP)	3.9	35.0	52.4	8.7	2.34	4
Group Discussion (GD)	5.8	45.6	41.7	6.8	2.50	3

BIBLIOGRAPHY

American Psychological Association, Standards for Educational and Psychological Tests. Washington, D.C.: APA, 1974.

Bachman, Lyle F. and Adrian S. Palmer. "The Construct Validation of the FSI Oral Interview". Language Learning, 31, (1981) : 67-86.

Clark, John L.D. "Toward A Common Measure of Speaking Proficiency". 15-26 in J.R. Frith (ed.), Measuring Spoken Language Proficiency.

Clifford, Ray T. Reliability and Validity of Language Aspects Contributing to Oral Proficiency of Prospective Teachers of German. In J.I.D. Clark (ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton: Educational Testing Service, 1977.

Hinofotis, Frances B. An Investigation of the Concurrent Validity of Cloze Testing as a Measure of Overall Proficiency in English as a Second Language. Unpublished Ph.D. dissertation, Southern Illinois University, 1976.

Joint Committee on Standards for Educational Evaluation, Standards for Evaluations of Educational Programs, Projects and Materials. New-York: McGraw-Hill, 1981.

Lowe, Pardee Jr. Manual for LS Interview Workshops. Washington, D.C. CIA Language School. 1980. 1982 (revised) (mimeo).

Madsen, Harold S. and Randall L. Jones, Classification of Oral Proficiency Tests," 15-30 in Adrian S. Palmer, Peter J.M. Groot, and George A. Trosper, eds. The Construct Validation of Tests of Communicative Competence. Washington, D.C. Teachers of English to Speakers of Other Languages. 1981.

Nevo, David. "The Conceptualization of Educational Evaluation: An Analytical Review of the Literature", Review of Educational Research, Vol. 53 (1), 1983. pp. 117-128.

Oller, John W. Jr. Language Tests at School. London: Longman Group Ltd., 1979.

Reves, Thea. "The Group Oral Examination: A Field Experiment", World Language English, Vol. 1, no. 4, 1982. pp. 259-262.

Shohamy, Elana. "Predicting Speaking Proficiency From Cloze Tests", Applied Linguistics, 3, 2, 1982. pp. 161-171.

Shohamy, Elana. "The Stability of Oral Proficiency Assessment on the Oral Interview Testing Procedure", Language Learning, 33, 3, 1983.

Shohamy, Elana, Reves, Thea and Bejarano, Yael. "Toward a Valid and Reliable Measure of Oral Proficiency: From Research Results To Educational Policy, School of Education, Tel Aviv University, 1984 (Mimeo).