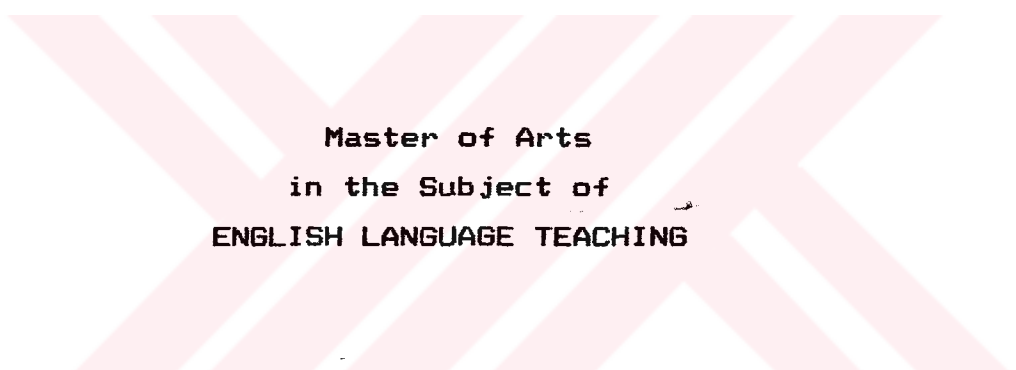


ÇUKUROVA UNIVERSITY
The Institute of Social Sciences

**ANALYSIS OF THE ORAL-PROFICIENCY EXAMINATION GIVEN TO
SCREEN STUDENTS ADMITTED TO THE ELT DEPARTMENT IN THE
1989-1990 ACADEMIC YEAR**



**Master of Arts
in the Subject of
ENGLISH LANGUAGE TEACHING**

Abdulkadir TURGAY

ADANA

1991

21599

**ANALYSIS OF THE ORAL-PROFICIENCY EXAMINATION
GIVEN TO SCREEN STUDENTS ADMITTED TO THE ELT DEPARTMENT
IN THE 1989-1990 ACADEMIC YEAR**

A Thesis Presented

by

Abdulkadir TURGAY

to

**The Institute of Social Sciences in Partial
Fulfillment of the Requirements for the Degree of**

**MASTER OF ARTS
in the Subject of
ENGLISH LANGUAGE TEACHING
Çukurova University**

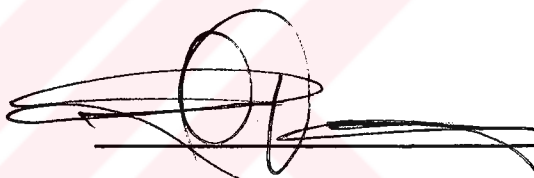
I certify that this thesis is satisfactory for the
award of the degree of Master of Arts.



Supervisor

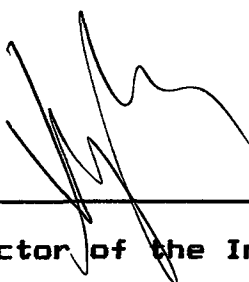


Member of Examining
Committee



Member of Examining
Committee

Certified that this thesis conforms to the formal
standards of the Institute of Social Sciences.



(Director of the Institute)

ABSTRACT

The study is aimed at analyzing the oral-proficiency examination given to applicants to the English Language Teaching (ELT) Department in the 1989-1990 academic year. The primary concern was to investigate the inter-rater reliability. The student interviews were evaluated by a four-member jury who conducted the actual test. They were also recorded on audio-cassette, and then re-rated by a second four-member jury. Statistical analysis of the ratings showed that there was very little agreement among the raters. Interview assessment was apparently impressionistic and subjective with minimal inter-rater reliability.

ACKNOWLEDGEMENTS

I have had the very good fortune during my teaching career to have had some teachers and colleagues whose influence on me has been incalculable. Among those are: Assoc. Prof. Özden EKMEKCI, my thesis advisor and our chairperson, and Assoc. Prof. John Terrance WEBB, teacher, colleague, friend, without whose guidance, encouragement, support, and contributions this thesis would not have been possible. They both receive my very special thanks.

Prof. Kyle PERKINS of Southern Illinois University must be thanked for his most helpful comments on, and insightful reactions to, the analysis and the interpretation of the statistical data involved in this study.

I am deeply indebted to Linda ETİZ, Hatice SOFU, and Fehmi Can SENDAN, my colleagues, for their invaluable contributions.

Then, my very special thanks are due to Leon AMADO, an ex-teacher and a colleague, for his invaluable feedback and suggestions at various levels of the study.

I also offer a word of special thanks to Peter Kenneth HOWELL, my office-mate, for his honest reactions, queries, and helpful advice.

Finally, my heartfelt thanks go to my naughty-but-cute daughter, and my wife, whose understanding, patience, and untiring assistance helped enormously to give the thesis its current shape.

TABLE OF CONTENTS

	<u>Page</u>
Abstract	iv
Acknowledgements	v
List of Tables	vii
 CHAPTER I INTRODUCTION	 1
1.1. Problem	1
1.2. Purpose of the Study	4
1.3. Hypotheses	5
1.4. Scope of the Study	6
 CHAPTER II REVIEW OF LITERATURE	 7
2.1. Introduction	7
2.2. Role of Testing in Language Teaching	7
2.3. Basic Considerations in Objective Evaluation of Oral Proficiency	 8
2.4. Oral Test Types (In Relation to Reliability, Validity, and Practicality).....	 25
2.5. Elicitation Techniques	28
2.6. Marking Systems	34
 CHAPTER III METHODOLOGY	 40
3.1. Research Design	40
3.2. Subjects	41
3.3. Method	41
3.4. Field Procedures	42
3.5. Data Collection and Recording	43
3.6. Data Analysis	44
3.6.1. Correlation Analysis	45
3.6.2. The Product-moment Correlation coefficient	46
3.6.3. Computation of the product-moment correlation coefficient	 48
3.6.4. Test of significance of correlation coefficients	 49

LEGEND	51
RESEARCH QUESTIONS	52-53
 CHAPTER IV PRESENTATION AND EVALUATION OF FINDINGS	54
4.1. Agreement in terms of the average grades.... given for overall proficiency by JJ	55
4.2. Agreement in terms of the average grades.... given for each of the areas by JJ.....	57
4.3. Agreement in terms of gender.....	59
4.4. Interrater (JJ) agreement in terms of C.....	61
4.5. Interrater (JJ) agreement in terms of S.....	63
4.6. Interrater (JJ) agreement in terms of B.....	65
4.7. Interrater (JJ) and interjury (JJ vs. J).... agreement in terms of the five areas	67
4.8. Role of gender in the rating of the students' oral proficiency	69
4.8.1. Agreement between male and female assessors (within JJ), in terms of their .. evaluation of all the students (male and .. female)	69
4.8.2. Agreement between male and female assessors (within JJ) in terms of their .. evaluation of the female students	71
4.8.3. Agreement between male and female assessors (within JJ) in terms of their .. measurements of the male students' oral .. proficiency	73
 CHAPTER V DISCUSSION, CONCLUSIONS, RECOMMENDATIONS	75
5.1. Discussion	75
5.2. Conclusions	76
5.3. Recommendations (especially in terms of implications for the training and monitoring of testers).....	78

APPENDICES	80
APPENDIX I	81
APPENDIX II	84
BIBLIOGRAPHY	94
GLOSSARY	104



CHAPTER I

INTRODUCTION

1.1. The Problem

"There could be no science as we know it without measurement", says Henning (1987: 1). In education, especially in the field of language teaching, different types of (language) tests are designed and administered to serve singly and jointly as measures of various skills.

Clark (1979: 35) states that in the area of language skills measurement, the testing of speaking ability is in all probability the most interesting and the most challenging to both classroom teacher and professional test developer. Interest in testing speaking ability, he continues, has grown considerably in recent years, in keeping with an increasing emphasis on the development of a usable level of English-speaking competence on the part of language students, both at the secondary school and college levels, together with a growing practical need to determine the speaking ability of many other groups of individuals outside the classroom setting (among these are business people in international trade, prospective tourists abroad, foreign students in the United States and Britain, and so on).

Present interest in the attainment of speaking ability, as referred to by Clark, is bound to grow even further in the days to come, as there will be greater and ever-increasing demand for the acquisition of "the ability to communicate with people of different linguistic background (and, thus, for the measurement of that

particular skill), especially in an ever-changing world such as the one we are presently living in.

As is known to all language teachers, that functioning effectively in a language requires competence in several areas which are all linked together. Since so many components of language are involved in the testing procedure, it is most probably this aspect of speaking ability which makes the measurement of it so difficult and challenging.

Jones (1975) expresses his concern about the efficiency related to measurement of listening comprehension:

There is, unfortunately, no way to make a direct assessment of a person's ability to comprehend a foreign language short of gaining access to his language perception mechanism, whatever that is. (p. 5)

I would assume that this would not only apply to the assessment of receptive skills (listening, and reading) but to productive skills (speaking, and writing) as well.

Oral production skill is equally as important to success in university-level course work as other skills, so there is oftentimes a need to determine the degree to which this ability is present in the learner.

But, as language teachers (whose teaching activities are closely related to testing) know only too well, assessment of oral proficiency (or speaking ability), as with written production, is regarded as one of the least objectively measurable of students' skills. This is so,

partly because so many complex components of language are being evaluated simultaneously, and partly because such evaluation mostly has to depend on the judgement of raters. And as it, due to the reason Jones referred to, cannot be directly measured, and can only be estimated, we will have errors no matter how hard and well we try.

However, we can feel an obligation to review our methodology carefully, to attempt to make improvements where it is possible to do so, even if reasonable prediction says it may not be at all possible to surface and discuss all of the problems, let alone find adequate solutions for them. Thus, we are often concerned about reliability, validity and efficiency; and we, from time to time, try to analyze the tests we administer to see how well they are performing as measurement tools.

Thus, the present study aims to analyze the oral-proficiency examination given at the beginning of the 1989-1990 academic year (by a four-member jury, to be referred to as "first Jury", of which I was a member) to screen students admitted to the ELT (English Language Teaching) Department. The need for such a study arises from the wish to see whether this evaluation (of the students' oral proficiency) stands or falls according to the evaluation of the "second jury" (another four-member jury, consisting of four randomly chosen teaching staff--two native, and two non-native-- that will be, for the second time, evaluating the same students' oral proficiency, basing their judgement on the same criteria used by the first-jury members. Along with the criteria employed during the preceding evaluation (Rating Scale 1 in Appendix I), "self-developed" guidelines (developed by the writer of this thesis) will also be used in the evaluation of the students' recorded oral proficiency to be carried out by the second-jury members.)

1.2. Purpose of the Study

An aim of this thesis is to study the oral-proficiency examination given at the beginning of the 1989-1990 academic year to screen students admitted to the ELT Department for the purposes of diagnosis: to ascertain to what degree the assessors--within both the first and the second juries--agree in terms of their assessment of the students' oral proficiency, as well as to extract information relevant to the practice of our profession and important to the improvement of our testing procedures, and thus to the fuller satisfaction of the needs of our students.

This oral examination, given for the second time in 1989-1990, aims to serve as screening each year for all students admitted to ELT Department. On the basis of the results of this examination (and the results of the written examination administered prior to the oral-proficiency examination), the jury (the first jury) arrives at a decision as to which of students are eligible to be major students, and which are to attend the preparatory program, which enables them to bring their proficiency in English to the desired level.

First, the two sets of scores (awarded by the two juries separately) will be compared to see how much commonality there is between the two tests. Next, on the basis of the analysis of the statistical data to be obtained from comparison and contrast of the set of correlations between various items of students' performance, the main research question to be answered is:

to what degree the two juries agree in terms of their overall evaluation of the students' oral proficiency.

If the findings, however, indicate that the examination is not reliable, objective, and/or valid, sources of deficiency will then be investigated, and working hypotheses developed to modify the examination in question, in order to close the gap between the evaluation of the different assessors, thus contributing to its objectivity and its reliability.

1.3. Hypotheses

The primary intent of this study was to judge neither the individual students nor the teachers; but only to evaluate an examination and relevant school/department policy in an effort to diagnose our own efforts.

Accordingly, I am, by no means, claiming any "inadequacy" on the part of the jury-members. On the contrary, I would like to predict that high correlations will be obtained when the oral-proficiency examination is evaluated by a set of other instructors who would serve as second jury in the project that I propose to carry out, the description of which is a second aim of this thesis.

In the light of the above statements,

1. It is hypothesized that the two juries will agree (HA). This hypothesis is based on the assumption that (1) there were four judges/jury members involved in each examination jury, and such will most probably help erase individual differences, (2) averaging scores given by each member will also help reduce error, and increase objectivity, and (3) there is not a wide gap between proficiency levels of the students.

2. It is contrarily hypothesized that there is no agreement between the two juries (H_0). This hypothesis assumes that oral proficiency cannot be directly measured, but can only be estimated; and we will have errors no matter how hard and well we try. Existence of differences is considered to be natural in the measurement of speaking ability, even if the rating scale will be the same.

1.4. Scope of the Study

The study, as mentioned earlier, is primarily aimed at investigating the extent to which variations of opinions in the evaluation conducted by the second jury correspond with those in the assessment done by the first jury (inter-rater reliability). Thus, the project undertaken is descriptive in nature, as well as correlational.

So, the study, in order to keep on the right track and achieve its goal, is limited to, and concerned with touching on urgent problems connected with oral-proficiency testing which should be tackled prior to devising such tests, including the administration, and evaluation of the proficiency level of in-coming students.

On the basis of the analysis of the statistical data and the findings to be obtained, suggestions and attempts to create an 'ideal' testing situation (within the limits of the resources available) also lie within the scope of the study.

CHAPTER II

ROLE OF TESTING IN LANGUAGE TEACHING

2.1 Introduction

The section--review of literature--comprises the following:

- a) review of evaluation procedures in general, and assessment of oral proficiency, in particular, with a view to reliability, validity, and practicality.
- b) review of the theoretical and practical considerations related to preparation, administration, and scoring of tests (oral tests, in particular).

In order to establish a sound basis for the (understanding and) discussion of the purposes of the topic, magazines, articles, periodicals, and books of different sorts, including a variety of rating scales employed for the measurement of oral interaction, have been reviewed to enable information transfer to those situations and areas on which the study is focused.

2.2 Role of Testing in Language Teaching

We are exposed to many different types of tests measuring multifarious skills/qualities in us as we enter upon some a new phase of our life; chief among those phases are education, and career.

In an educational setting, especially in the field of language teaching/learning "four basic types of language

tests--aptitude, progress, achievement, and proficiency tests" (Valette, 1977: 5) are the ones that are widely utilized.

The most common use of language tests, and educational tests in general, as Henning puts, is to pinpoint strengths and weaknesses in the learned abilities of the student. (1982: 1)

Language testing today reflects current interest in teaching genuine communication, but it also reflects earlier concerns for scientifically sound tests. Testing during the last century, and the early decades of this one, was basically intuitive or subjective and dependent on the personal impressions of teachers (Madsen, 1983: 5).

Not a great deal of progress has been achieved since, it seems, in the field of measuring students' oral communication skills, in the sense that such measurement, to a certain degree, at least, still remains greatly subjective.

After the intuitive stage, testing entered a scientific stage, a time that stressed objective evaluation by specialists. We are now at the communicative stage, a time when we emphasize evaluation of language use rather than language form. (Madsen, 1983: 6-7)

2.3 Basic Considerations in Objective Evaluation of Oral Proficiency

The present study will not be dealing with language testing in all of its forms, but only the testing of oral proficiency.

In the light of such an approach, this section will cover the following objectives: (1) to discuss common problems relating to oral-proficiency testing, (2) to explore new ideas and techniques for measuring such skill, (3) to establish a future direction for (development and) improvement of measurement procedures being presently employed.

As mentioned earlier, contrary to multiple-choice or cloze-type tests, oral-proficiency tests allow room for great disagreement among scorers. The following section will present and discuss possible ways of making free-response tests (e.g. oral production or essay tests) fairly objective. (To the question "What changes would improve the objectivity of the test in question?", possible answers will be sought in section V.)

Valette (1977) refers to, and defines four requisites of a dependable test:

1. **Multiple Samples:** For students' mastery of a complex body of language to be reliably evaluated, they must be asked as many questions as possible. In other words, they must be given enough opportunities to demonstrate their familiarity with the various aspects of the subject. Consequently, a language test, in order to yield a reliable score, must be long enough to provide a generous sampling of the area or areas tested. It is also important that there be a wide spread in the level of difficulty of the items. Care must be taken so that the examination is brought politely to an end when the interlocutor/assessor is satisfied that he can

precisely determine the learner's level of proficiency, and when the learner is having extreme difficulty.

2. **Standard Tasks:** If the scores are to provide a dependable means of comparing student performances, then all students must be given the same items or items of equal difficulty. (Sufficiently detailed consideration should be given to the motivation and fatigue of the test-takers.)
3. **Standard Conditions:** Reliability of the test score can be assured only if the students take the examination under "identical" conditions to the extent possible.
4. **Standard Scoring:** All tests must be scored in an identical manner, to the extent possible. It may not be at all be possible to eradicate all problems of subjectivity in oral or written production tests, but some of the difficulties could be eliminated or, at least, minimized. (44-46)

Jones (1979: 50-51) points out the view of many experts in the field of testing: one of the most important considerations which must be kept in mind when designing or selecting a test of second language proficiency is the purpose for which the test is to be used (in order for our tests to discriminate well and provide consistent results about the performance level of students).

As our concern with oral-proficiency measurement is to

identify the right class for a particular learner, tests we use must provide predictive information about how well a candidate will perform in a given situation.

Underhill's arguments about placement tests are that "the test-maker must know which classes or levels are available; then her task is to decide in which of the pigeon-holes--the range of available classes--to put the learner" (12). "The fewer the options available, he further argues, the more consistent the decisions will be."

That the test-maker knows a lot about what goes on in those classes, and that she is a regular class teacher herself, are regarded, by Underhill, as advantages contributing to the robustness of the decision to be made. So she will be able to design a test which attempts to approximate, to the extent possible, the types of tasks the candidate is likely to face in typical situations of such classes or levels.

Apart from aims, the second most important consideration, as stated by Underhill (1987), is resources, which consists of three different sub-categories: people, time, and physical facilities.

The success of a testing program, to a large extent, depends on the utilization of resources available, to the utmost.

The three resources--people, time, and physical facilities--are specifically emphasized by Underhill:

- a) People: In a large testing programme, all these different tasks will be done by different people with different qualifications

and skills. Neither a big programme nor a small one is inherently better; the important point is to design a test that uses the people available in the most effective way to meet the aims of the programme. (This may involve people with oral testing and marking experience, teaching staff and administrative staff as well as native and non-native speakers.) (15-16)

What is vital (no matter what kind of people we can reach for help) is to make sure that we are using the best people available.

b) Time: At this point, care must be taken that sufficient amount of time is allotted to the test development, test operation, and test improvement stages. In doing so, one should not go to extra-ordinary lengths: just as learners' performances drop if a test goes on too long, so the consistency of interviewers' ratings fall quickly when they get tired (16)

c) Physical facilities: Some of the physical resources available for use may be: (1) rooms and furniture for preparation and testing, (2) sound or video recording equipment, (3) photocopying, printing or other duplication facilities.

Some careful thought in advance about the choice of room and the arrangement of furniture can make a big difference to the atmosphere in which the test is to be

conducted. Tests should be held somewhere that is quiet and free from interruption, in order to prevent the test situation from being affected by noise and other people. This will also make the test-taker get the feeling that his test has been given a very high priority. Otherwise, he may lose his train of thought or flow of speech, and so lose confidence.

Administering the test in an office (as opposed to a classroom), and arranging the furniture to be used away from the chalkboard or teacher's desk will create a natural, friendly atmosphere, and reduce the feeling of a teacher questioning a student.

If a third person is present to assess or observe, but not actually to take part, she should sit a bit further away, so that she is not physically in the area where the communication is taking place. (p.17)

In addition to the previous suggestions which were the shared view of many experts in testing, Berkoff (1985) quotes from Jones the four points which he considers to be important in the improvement of oral examinations:

- (1) Developing appropriate interviewing techniques: This can best be achieved by issuing clear instructions about the aims of the examination, both to the testees and to the examiners. Such instructions will also indicate the techniques that will be used.

This will add greatly to the validity of the

examination. These instructions will also have the effect of controlling the methods that the examiner uses.

- (2) Defining points on the rating scale and
- (3) Establishing criteria for rating: In addition to knowing the aims of the examination and how it will be conducted, of the examinee also needs to know how he is going to be evaluated.

Here, Berkoff (1985), finding such satisfactory, but still asking for refinement, emphasizes two points of the general criteria suggested by such as Wilds and Jones. Those are:

- (i) As foreigners are expected to have a "foreign" accent "accuracy of pronunciation" is not always one of relevant criteria (is not one of the areas evaluated).

- (ii) We all know people who, at least occasionally, have difficulty expressing themselves in their own language, so that "speaking proficiency equivalent to that of an educated native speaker" is not normally a criterion by which to judge a school leaver.

- (4) Training the examiners: In theory this is a good thing, but in practice it is extremely difficult. (1985: 96-97)

Some other commonly shared (especially by Underhill 1987, and by Henning 1982) problems to be tackled are:

Perhaps the most common cause of invalid tests is the failure of a test to measure adequately either instructional objectives or course content. The course must fit the instructional objectives, and the test as well should reflect instructional objectives by reference to vocabulary, structures, and skills actually taught. (Henning: 36)

Different people have different needs--at a personal level and at a professional level. If there is a wide variety of different individual needs, the overall aims of the program may be to (teach and) test general English, because there are not sufficient resources to develop materials to satisfy all such needs.

If the learner realizes that the interviewer is sufficiently interested in his personal needs to adapt the test accordingly, he will respond to that expression of interest. He will have more to say about topics that concern him personally. He will not necessarily perform better, but he will feel that the test is more relevant to him, and that the assessment will be based on a more representative sample of his language. (Underhill: 18-19)

The above conclusions are true for expectations on the part of the students. "How learners react to a test,

and therefore how well they do, depends on how the test compares with what they expected it to be like." (Underhill: 19).

With regard to a possible "discrepancy between the objectives of a testing program and the cultural expectations of the learners, Underhill asserts:

As oral tests allow the personal side of the learner to come through more than in written tests, such tests will particularly be sensitive to this discrepancy, and can produce distorted and inaccurate results. In such a case, the test will reflect the degree of familiarity with the culture with which the objectives are associated rather than just oral proficiency. (20)

Further, with the considerations mentioned so far, Henning (1982), in regard to subjectivity of scoring, touches on the following points:

...more than one judge should be consulted independently of marks assigned by other judges. The total of all judges' ratings should determine the student's mark. Second,

judges should make use of some precise rating schedule. A certain number of marks should be deducted for errors of specified type or number. In this way, judges will be giving equal weight to the same kinds of performance.

Finally, sufficient samples of language should be elicited from the students. In writing or

speaking tests, students should be given more than one topic, to ensure that a more comprehensive picture is taken of their language use in a variety of situations.

(36-37)

On matters of objective evaluation of oral proficiency, Cronbach (1984) deems, from a 'psychological' standpoint, the following points relevant and important. The following citations are taken from the first half of Cronbach's book, *Essentials of Psychological Testing*, have been given as a 'running commentary' on the importance of psychological factors in testee performance, slightly modified by changes made by the writer of the thesis (provided in brackets) to enhance continuity.

- By careful instructions to the observer or scorer, free-response tests and observations can be made fairly objective.
- The test-taker must be encouraged to earn the best score he can. If he does not push himself, the test does not display his maximum performance.
- The goal of the tester should be to bring out the [test-taker's] best possible performance (within the rules), and this means that the examinee must want to do well and must understand what is considered a good performance. If he is to show at his best, directions need to be clear and explicit, even to the extent of explaining how various sorts of errors will be penalized.

- Each learner/candidate should be placed in essentially the same situation.
- The tester's first responsibility is to choose a test that fits the test-taker. (Even when someone engages the tester to give a specific test, the tester ought to judge independently whether the test is suitable.)
- The tester is obliged to give the examinee a full chance to display his ability, and standard procedures may not fit. The tester will not necessarily realize in advance that a test is unsuitable, but she should remain alert to indications of invalid response.
- When a standard procedure will lead to invalid conclusions, she will often be wise to change the procedure.
- Some tests are sufficiently simple for any adult to give; a few are so subtle that months of special training are recommended.
- Considerable skill is required when the tester is to question the examinee orally, and to use follow-up questions when the first answer is unclear.
- To obtain meaningful results, the tester must become thoroughly familiar with the test. Even a simple test presents stumbling blocks; most will be avoided if the tester studies the manual in advance. But carelessness can produce spectacular errors.

- The physical setting for the test matters: It must put the [learner] at ease.
- Fatigue apparently affects effort rather than ability.
- Efficiency is to be maintained without interfering with the opportunity of individuals to ask questions.
- The tester who likes people will develop many techniques for establishing rapport. The person who proceeds coldly and "scientifically" to administer the test, without convincing the test taker that she regards him as important, will get only limited cooperation. (Evidence of poor rapport includes inattention during directions, giving up before time is called, restlessness, or complaining about the test items.)
- When the candidate is at his ease again (after some challenging questions), it is usually possible to return to the troublesome task with better success.
- To elicit the learner's best efforts is the sine qua non of good testing.
- The form of address which puts [a learner] at ease...may belittle or even antagonize another. Keeping the subject encouraged can be accomplished in many subtle, friendly ways; by an understanding smile, a spontaneous

exclamation of approval, an appreciative comment, or just quiet understanding, and so on.

- In general it is effective to praise frequently and generously, but if this is done in too lavish and stilted a fashion it is likely to defeat its own purpose.
- While [praising,] the examiner should remember that he is giving approval for effort rather than for success on a particular response.
- Under no circumstances should the examiner permit himself to show dissatisfaction with a response, though he may smilingly refuse to accept a flippant answer obviously intended to "test the limits."
- In praising poor responses of [adult] learners, the examiner should remember that the purpose of the commendation is to insure confidence and not to reconcile the learner to an inferior level of response.
- In the case of a failure that is embarrassingly evident to the learner himself, the examiner may take occasion to point out that he doesn't expect the learner to be able to do all those things, or he may interject, "That was a good try!"
- The examiner should not drag out responses by too much urging and cross-questioning. To do so robs the response of significance and

discourages spontaneous effort.

- While the examiner must be on his guard against mistaking exceptional timidity for inability to respond, he must also be able to recognize the silence of incapacity or the genuineness of an "I don't know" from the learner who knows when he [does not know the answer].
- If the learner is becoming confused and upset, to let him continue might leave him more discouraged. The tester should observe carefully and choose whatever course seems likely to have best effect.
- It is the duty of the tester to obtain the best record the learner can produce by his own efforts--without aid.
- The tester must learn to suppress direct and indirect hints. Unintended help can be given even by facial expression. The person taking a test is always concerned to know how well he is doing and watches the examiner for indications that he is [performing correctly].
- The tester has been accustomed to think of herself as an unemotional, impartial task setter. [and] traditional language suggests that the psychologist, like an engineer, is "measuring an object" with a technical tool. But the tester's "object" is a person. The traditional recommendations that the tester be encouraging and help the person to understand

the value of the test barely touch the socio-psychological complexities of testing.

- The learner [will] bring many hopes, fears, assumptions, demands and expectations into the test situation. He may...respond intensively to certain real as well fantasied attributes of that situation. Being human and having to make a living--facts often ignored--the tester too brings hopes, fears, assumptions, demands and expectations into the test situation. She too responds personally and often intensively to what goes on--in reality and in fantasy--in that situation, however well she may conceal her personal response[s] from the learner, from herself, and from her colleagues.
- The ideal of objectivity requires that we recognize as much as possible what is going on in the test situation we are studying. It requires in particular that we remember that the tester and the test taker are both human and alive, therefore inevitably interacting in the test situation.
- A test can be an hour-long (a ten-minute) stress situation, every moment of which involves emotional interaction between tester and test-taker (if not well attended to).
- The match of tester to test-taker in ethnicity or race is a serious question (due to the "pessimistic...hypothesis" that one cannot relate to another person unless [their culture and/or] skins [are similar]...

- The conventional style of testing may elicit little response from a test-taker who considers the tester an alien and a threat. The tester, of whatever social or family origins, must discover a style that reaches persons from a different background, or disqualify herself. (Holding oneself in is a defense many people use in alien situations.)
- To prevent learners (from) making procedural errors would mean to magnify a tester's superiority, and so result in reducing a test-taker's scores.
- In the weighing of a truckload of wheat, there is no problem of motivation. Likewise, in weighing a person, when we put him on the scale, we get a good measure no matter how he feels about the operation. In a psychological test, [however,] the subject places himself on the scale, and unless he cares about the result, he cannot be measured.
- Equally powerful and more universally available as a source of motivation [than in the case described above], is the desire to maintain self-respect and the respect of others.
- Not every learner wants to do well. Attitudes toward oneself and toward task performance have a profound influence on response to tests and to school assignments.
- The [learner's] desire to do well may

interfere with performance. A tense performer overlooks errors that he would correct otherwise.

- Tension leads to poor coordination and erratic movements. In a verbal test, the respondent may guard (himself) against criticisms of his answers by being overcritical of himself.
- For many students tests become stressful. Although much is said about detriment [caused by] anxiety, it is also true that mild anxiety can heighten performance.
- One cannot be confident as to the direction of causation here; poor learning may be the source of weak performance, from which anxiety follows.
- The tester enters into a contract with the person tested. The tester is a go-between between a college and an applicant, for example. The tester has responsibilities in both directions, and she should try to make sure that these opposite numbers have mutually consistent expectations.
- When a person's opportunity hinges on a score, he should have a way to question its accuracy, even [when] the likelihood of error is small. The principle of openness has its limits, though.
- ...variations in...scores given by two different groups/juries could perhaps be

tolerated because the [evaluation] of a composition [or of a student's] oral proficiency is inherently subjective.

- One way to systematize (increase agreement among scorers) scoring is to develop rules for judgement that all scorers will follow.
 - Make rules so clear that scorers would be able to agree in scoring new answers.
 - The direct count on measure of performance is called the raw score (for example, this can be a count of right answers minus penalties for errors). The most direct way to give meaning to a raw score is to display the content of the test.
 - The more the tester knows of her own personality, of her preferences for different types of subject, and of the biases she brings to test interpretation, the greater the chance that she can meet each situation properly.
- (1-536)

2.4. Oral Test Types

Different testing experts divide language tests (in general) into different categories, and the number of categories will depend on the definition of categories and sub-categories.

Before moving into the field of oral-test types, a general category of language test will be presented.

Valette's categorization of four basic types language tests (as briefly mentioned earlier) are:

1. Aptitude test: indicates whether a student is likely to learn a second language readily.
2. Progress test: measures how much the student has learned in a specific course of instruction.
3. Achievement test: measures how much the student has learned, but is usually not built around one set of (teaching) materials.
4. Proficiency test: measures what students have learned, and is intended to determine whether this language ability corresponds to one or more specific language requirements. (Placement tests are also included in this category.) (5-6)

In order to provide a basis for the better understanding and discussion of types of oral tests, Upshur's definition will serve as a point of departure (1975):

An oral English test is one in which, first, the examiner sets a task for the student to perform; the task demands English speech. Second, the student attempts to perform the task. Finally, the examiner notes the performance and decides whether or not it is correct or acceptable or intelligible English

speech. If he decides that it is, student has passed the test; if he decides that it is not, the student has failed. (p. 55)

Jones' (1975) view of oral testing (expressed in response to a question about the expense involved in such tests) is: "We have to have this kind of a test to be able to find out what a person's ability to speak really is."

This section will look at types of oral test with a view to their advantages and/or disadvantages (in terms of validity, practicality, and above all, reliability) so that one may come up with the 'best' oral-test type or technique that will serve a specific purpose. (The primary function of oral-proficiency testing, in our case, is to define our in-coming student's proficiency levels, in speaking, and thus, to help us with student placement.) For some testing people, such as Jones (1985) or Clark (1975), oral-proficiency tests will have two forms.

Clark's (1975) definition of such tests would include:

1. Direct: A direct test of oral proficiency, in the face-to-face communication sense, would involve a test setting in which the examinee and one or more human interlocutors do, in fact, engage in communicative dialogue.
2. In some cases, of course, an indirect test may involve certain quasi-realistic activities on the area, an indirect test may require the student to describe printed intelligible spoken responses. However, since such testing procedures are not truly reflective of a real-life dialogue situation, they are

considered indirect rather than direct measures of oral proficiency. (11)

In addition to these two, Clark (1979) proposes a third type of oral testing approach: semi-direct speaking test. This being characterized as "tests which, although eliciting active speech from the examinee, do so by means of tape recordings, printed test booklets, or other "nonhuman" elicitation procedures, rather than through face-to-face conversation with a live interlocutor." (38)

Harris (1969, 85) asserts that "interviews tend consciously or unconsciously to set unequal tasks for the candidates and to score performances on different bases because of the relatively limited nature of the guidelines provided. Harris therefore favors a more objective testing procedure, the "highly structured speaking tasks."

Types of activities/techniques that would appear in each type of the oral-proficiency tests will be dealt with in section 2.5.

2.5. Elicitation Techniques

This section will briefly touch on some common methods employed to "get a learner to say something in a speaking test.

Prior to listing elicitation techniques, one should note that each technique has some strengths and weaknesses, and not all of them may be used with confidence for the same purpose. The best test or technique is the one that provides us with the necessary information for making some decisions.

Underhill's (1987) list of elicitation techniques is long and varied:

1. Discussion/conversation: Two people (interviewer and learner) have a conversation on a topic of common interest.
2. Oral report: The learner prepares and gives an oral presentation.
3. Learner-learner joint discussion/decision making: A group of two or more learners is tested, without the participation of an interviewer.
4. Role-play: The learner is asked to take on a particular role, and to imagine himself in that role in a particular situation.
5. Interview: (For many people, this is the only kind of oral test.) It is a direct, face-to-face exchange between learner and interviewer. It follows a pre-determined structure, but still allows both people a degree of freedom to say what they genuinely think.
6. Learner-learner description and re-creation: One learner describes a design or construction of model-building materials to another learner who has to reconstruct the model from the description alone, without seeing the original.
7. Form-filling: The learner and interviewer

work together to fill in a form.

8. Making appropriate responses: The learner is given a number of short, unrelated situations that might occur in everyday life. He is asked to imagine himself in each situation in turn, and to provide what he thinks would be an appropriate spoken response.
9. Question and answer: This typically consists of a series of disconnected questions which are graded in order of increasing difficulty, starting with short simple questions, up to long and complex ones.
10. Reading blank dialogue: The learner is given a dialogue with only one role written in. He has a few minutes in which to read it through and prepare the missing lines.
11. Using a picture or picture story: The learner is given, before, the test starts, a picture or sequence of pictures to look at. Then, during the interview, he is asked to describe the picture or story, with the chance to speak freely.
12. Giving instructions/description/explanation: The learner is asked to describe, at some length, a well-known object, a system, or an everyday procedure.
13. Précis, or re-tell story, or text from aural stimulus: The learner is to immediately re-tell, or summarize a short passage or story

he hears on a recorded tape.

14. Re-telling a story from written stimulus: The learner is to immediately re-tell, in his own words, a passage or series of short passages he reads to himself.
15. Reading aloud: The learner reads aloud to the interviewer, either a passage of a text, or part of dialogue in which the interviewer or another learner reads the other part.
16. Translating/interpreting: The learner is asked to translate from his native language into the foreign language a short passage, or series of passages from a text he is familiar with.
17. Sentence completion from aural or written stimulus: The learner is asked to read through a series of sentences with some words missing, and to suggest a possible way of completing each sentence.
18. Sentence correction: The learner is to identify the error in a sentence he is presented with, to say what is wrong, and to correct it.
19. Sentence transformation: The learner is to demonstrate his knowledge of specific language structures by transforming stimulus sentences into a different grammatical pattern.
20. Sentence repetition: The learner hears a

series of sentences or utterances, and repeats them as accurately as possible.

On the basis of Clark's (1975: 23) definition of oral proficiency,

student's ability to communicate accurately and effectively in real-life language-use contexts, especially in the face-to-face conversations typical of the great majority of real-world speech activities, considerations of face validity appear to require human administration of a conversation-based test, which also must be evaluated by human raters

direct interview techniques (Elicitation techniques 1, 5, 9 would blend into this kind) deserve continuing close attention. (Advantages/disadvantages of human vs. mechanical administration and scoring processes will be presented in Section 2.6.)

Direct Interview Technique (the definition of which almost infallibly describes the Foreign Service Institute [FSI]-type interview) will continue to be used by our ELT (English Language Teaching) Department, at least until the argument that the Direct Interview Technique as an effective and efficient technique is discredited, or a better and more efficient measuring process is recommended.

As suggested by Liskin-Gasparro (1987) the four phases of the interview technique are:

1. Warm-Up, which consists of social amenities and simple conversation that are easy for the student. The purpose is to put the student at

ease; to reacquaint the student with the language necessary, and give him or her a

chance to speak easily; and to give the tester the first indication of the student's probable proficiency level.

2. Level Check comes when, after the student has been put at ease and has gained confidence during the warm-up, the tester chooses a level to begin the level check.
3. Probes follow, once a performance floor has been established. The purpose here is to find the linguistic "ceiling", the level at which the student can no longer speak consistently and accurately, and to make sure that the level the tester has been checking during the level check is indeed the highest level the student can sustain.
4. Wind-Down serves to leave student with a feeling of accomplishment after her/his speaking ability has been stretched to the limit. During this wind-down phase, the tester returns to a level the student can handle comfortably. (14-17)

1 For a more detailed study on the FSI-type interview, refer to Clark (1979), and Oller., Jr. (1979).

2.6. Marking Systems

The scoring of most free-response tests, such as tests or compositions, depends, to a large extent, on the personal judgement of the assessors. Such examinations, for this reason, do not always have such high reliability as objective test types (e.g. multiple-choice, cloze-type tests).

With respect to the problem of subjectivity involved in the administration and scoring of free-response tests, Underhill (1987) refers to two solutions:

...testing integrative skills such as speaking or writing, he has come to the conclusion that objective tests are of more questionable validity, and a lot of time and effort has been spent devising objective test systems with credible claims to validity. In situations such as formal examinations, it is essential that the results should be as reliable as possible, and in such cases validity may have to take second place. In other cases, subjective and objective tests are put together in a single examination in the hope that it will combine the validity of the former and the reliability of the latter.

The second solution, and the one on which this book is based, is to make a conscious decision that the person-to-person aspect is so important in testing oral proficiency that it cannot be traded away, and to face up to the consequent problems of involving human judgement. People are inconsistent; they

don't always agree, either with each other or with what they said or thought last week.
(88-89)

At this point, I would like to focus attention on the two major components of any testing process: administration and scoring. As will be found in the Glossary of this thesis, they both fall into the same two categories: mechanical and human, while "test scoring (both mechanical and human) can be further divided into simultaneous scoring and delayed scoring." (Clark, 1979: 35-49)

As our concern is, and will probably continue to be, with human administration and simultaneous human scoring, let us look at each marking category (from the standpoints of validity, reliability, and practicality) in terms of advantages/disadvantages.

Human Administration and Scoring

Clark (1975) says of the advantages/disadvantages of Human Administration and Scoring, that:

Advantages:

- it approaches real-life communication as closely as is possible in the test situation.
- a high level of reliability can be obtained.

Disadvantages:

- face-to-face conversation between examiner and

examinee on a variety of topics does, of course, differ to some extent from the contexts in which these communications take place in real life.

- also, specifically, the psychological and affective components of the formal interview differ from those of the real-life setting.
- it is more costly in terms of administration time and the associated manpower expense.
- it requires the presence/participation of a team of trained and experienced raters.
- the personalities of the interviewer and interviewee also constitute an extraneous factor.

Mechanical Administration and Scoring (Underhill, 1987: 35; Clark, 1975: 10-28)

Advantages:

- high level of reliability can be obtained.
- Mechanical Administration and Scoring may be cost-effective in that only one technician together with a rater would suffice to administer and score the test.
- it saves administration time. (A large number of students can be tested simultaneously in a

language laboratory setting.)

- Mechanical Administration and Scoring does not have to be done in real time (in the same place and at the same time as the test itself). It can be done where and when convenient, in comfort, with the facility to replay any part of the tape if desired.
- the marking time for each test can be considerably shorter than the test itself (due to "test compression", which enables the operator to turn the machines on to record each learner's response and to turn them off during instructions or longer stimuli).
- recorded tests can be useful where live interviews are difficult on account of difficulties in getting learners and assessors in the same place at the same time, for personal or cultural reasons.

Disadvantages:

- Mechanical Administration and Scoring is not very authentic.
- mechanical devices capable of evaluating speech in a conversational situation are not a viable possibility at the present time.
- computer-based equipment is costly and complex.
- programs used in mechanical devices which evaluate sounds or phrases have a very limited

repertoire in detecting human sounds.

- if only one rater is required to score all sections of a given test tape, gradual shifting of scoring standards in the course of the grading process may be observed.

Simultaneous Scoring

Advantages:

- Simultaneous Scoring is viewed as more favorable to high scoring-reliability.
- all relevant communicative stimuli are available to the scorer in Simultaneous Scoring.
- Simultaneous Scoring method may benefit from a "recency of exposure" factor (in that the rater has the conversation more clearly and more thoroughly in mind).
- the interview can be lengthened or modified in certain ways if deemed necessary.
- with Simultaneous Scoring, there is no possibility of encountering technical difficulties such as poorly recorded or otherwise distorted tapes that might hinder accurate scoring in Mechanical (delayed) Scoring.

- Simultaneous Scoring is preferable, as it is virtually free from expense, in that the examiner (already necessarily on hand to administer the interview) is required to spend only a few additional moments after the completion of each student's performance to assess the appropriate score level.



CHAPTER III

METHODOLOGY

3.1. Research Design

This study aims to be a descriptive and a correlational research project. Incoming students from various secondary schools in Turkey were interviewed individually by a four-member jury. They were asked questions on various topics. Their responses were tape recorded and an assessment of their oral skills was made as they presented themselves in the recordings. This project consists in providing the intercorrelations between various sets of assessments arrived at by the raters.

The students were required to have a certain level of competence in speaking (demonstrated by (1) the student's score on the 'Departmental-Prep-level-Exemption Examination' accounting for 90 % of the total mark and (2) his/her score on the 'Oral-proficiency Examination' accounting for 10 % of the total mark. It was then decided that students should be admitted (according to their scores) to either preparatory-level or first-year language courses at the ELT Department.

Departmental-Prep-level-Exemption Examination (DPPE) was designed (by the ELT staff-members) to measure students' discrete-point language skills, such as level of knowledge of grammar, vocabulary, reading, and listening comprehension.

A correlational study (measuring interrater reliability) was then conducted (after having had the students' utterances evaluated by another four-member jury) to see to what extent their scores covaried.

3.2. Subjects

The subjects of this study were 22 Turkish secondary-school graduates, of whom 6 were male and 16 were female. All of the students were newly graduated. About 40 % of them, however, were 'transfer' students from other universities wishing to study in their home town. These schools differed with respect to the grade level at which they began, the length of the program, and the proportion of the school day taught in English. In some private high schools, English was the medium of instruction for most of the school day for several years. Thus, they represented a rich sample of students of quite different cultural and educational backgrounds, and most probably with varied reasons for wishing to become second-language teachers.

3.3. Method

All the subjects were interviewed individually, in which direct face-to-face communication between learners and interviewers took place. This interview procedure required the interviewers to find out certain things about the learners, mainly through answers to spontaneous personal questions or predetermined questions on various topics or subjects raised during the interview, allowing the learner to develop and build on questions and topics and thus to show their proficiency level.

Below are some examples of such probing questions (asked about hobbies, professional interests, past experiences, future plans, current affairs, etc.) like the following:

- What do you do in your spare time?
- What kind of books do you like reading? and why?
- How did you learn/ start learning English?
- Why did you decide to study here?
- Which city would you choose to live in if you had to make a decision?
- Which country or countries in the world would you like to visit if you could? and why?, etc.

Each student was categorized later as "passed" or "failed" on the basis of whether his/her score had reached the cut-off score (70 % of the total) on the DPEE and the oral-proficiency examination.

3.4. Field Procedures

All the data were collected during the interviews, which were conducted in a colleague's office over a period of about three hours.

The researcher was a male native speaker of Turkish. As a member of the jury, he was present during all interview sessions with the subjects, to assess their oral proficiency. As the students were newcomers, the researcher had met them all for the first time during the

interview, after the DPÉE.

During the proceeding sessions, the researcher talked with each of the subjects in order to clarify a point, or to ask them further questions.

Each interview session lasted 3-5 minutes, depending upon the students' level of performance, or whether or not they were 'transfer' students.

The interview situation was so arranged that the subject would be seated close to the desk, holding the microphone to speak into. The tape recorder was placed in plain sight, on a nearby chair. A jurymember (rarely the researcher) started the talk or discussion.

3.5. Data Collection and Recording

The data which were the basis of this study consisted mainly of the students' utterances/responses elicited in response to the questions asked. These questions were related to topics raised during the interview. No pictures or other types of visual stimuli, but only aural stimuli were used.

The sessions of the interview conducted with (each of) the students were tape recorded from the moment communication began to take place. The tape recorder was turned on and off during the session in order to ensure that irrelevant material was not recorded.

Afterwards, the same recorded interview was evaluated by the second jury, listening to the students' recorded utterances. This procedure enabled the researcher to base

his/her conclusions on a comparison of the results of the first jury's evaluation of the 'live' interview and of the second jury's evaluation of the 'recorded' version of it.

Here one point must be noted. The second jury did the evaluation of the students' oral proficiency using the same criteria previously employed during the live interview procedures. There were some additional comments on and definitions of the rating scale (Compare rating scales I and II in Appendix I.), which were intended to produce a greater spread of marks, but which actually proved to be too analytic and so made the evaluation even more difficult for some members in the second jury.

3.6. Data Analysis

Data included comprise the scores obtained from a sample of 22 students, whose oral skills of pronunciation, fluency, comprehension, sticking to the subject, and grammatical accuracy have been rated by the two different juries described. (See p. 3.)

One problem in the analysis was that only overall average scores assigned by the members of the first jury were available to the researcher, as compared with the availability of individual scores assigned by each member of the second jury. (See Table 1 in Appendix II.)

With respect to the research questions, the relationship (correlational analysis) between the null (statistical) hypothesis and the research hypothesis was scrutinized. The criterion thereby arrived at for rejecting the null hypothesis was found to be that the Pearson product-moment correlations (r) must be significant at the .01 level for a one-tailed test, i.e. $p < .01$,

one-tailed test. (The statistical procedures followed in the computation of the Pearson product-moment correlations are described, in detail, below.)

3.6. Data Analysis

3.6.1. Correlation Analysis:

Correlation analysis is the statistical tool that we can use to describe the degree to which one variable is linearly related to another. (Levin, 1987: 535)

Let X and Y be a pair of random variables, with μ_X and μ_Y and variances σ_X^2 and σ_Y^2 . A measure of strength of their linear association is provided by the correlation coefficient, ρ , defined as (Newbold, 1984: 442)

$$\rho = \text{Corr } (X, Y) = r = \frac{\text{Cov } (X, Y)}{\sigma_X \sigma_Y}$$

It can be shown that the correlation must lie between -1 and 1, that is, with the following interpretations:

- (i) A correlation of -1 implies perfect negative linear association.
- (ii) A correlation of 1 implies perfect positive linear association.
- (iii) A correlation of 0 implies no linear association.

(iv) The larger in absolute value is the correlation, the stronger is the linear association between the random variables. (Levin, 1987: p.539)

One point that we must emphasize strongly is that r measures only the strength of a linear relationship between two variables.

There are many correlation coefficients in statistics. One of them is the product-moment correlation coefficient, which has been used in the computation of the correlation coefficients in this study.

3.6.2. The product-moment correlation coefficient :

Sokal and Rohlf's (1969) notion of the product-moment correlation coefficient and the derivation of its formula are described as follows: (Sokal and Rohlf, 1969: 494-523)

There are numerous correlation coefficients in statistics. The most common of these is called the product-moment correlation coefficient, which in its current formulation is due to Karl Pearson. We shall derive its formula through an intuitive approach similar to the one we explained above.

It can be concluded that the sum of products is a measure of covariation and it is, therefore, likely that this will be the basic quantity from which to obtain a formula for the correlation. We shall label the variables whose correlation is to be estimated as Y_1 and Y_2 . Their sum of products will, therefore, be $\sum y_1 y_2$ and their covariance $[1/(n - 1)]\sum y_1 y_2 = s_{12}$. The latter quantity is analogous to a variance, that is, a sum of squares divided by its de-

degrees of freedom.

A measure of association should be independent of the original scale of measurement so that we can compare the degree of association in one pair of variables with that in another. One way to accomplish this is to divide the covariance by the standard deviations of variables Y_1 and Y_2 . This results in dividing each deviation y_1 and y_2 by its proper standard deviation and making it into a standardized deviate. The expression now becomes the sum of the products of standardized deviates divided by $n - 1$

$$r_{Y_1 Y_2} = \frac{\sum y_1 y_2}{(n - 1) s_{Y_1} s_{Y_2}} \quad (3.1)$$

This is the formula for the product-moment correlation coefficient $r_{Y_1 Y_2}$ between variables Y_1 and Y_2 . We shall simplify the symbolism to

$$r_{12} = \frac{\sum y_1 y_2}{(n - 1) s_1 s_2} = \frac{s_{12}}{s_1 s_2} \quad (3.2)$$

This notation is adequate unless you have two-digit variable numbers, in which case commas are necessary. Thus, $r_{12, 14}$ is the correlation between variables Y_{12} and Y_{14} . Expression (3.2) can be written in another common form. Since,

$$s \sqrt{n - 1} = \sqrt{s^2 (n - 1)} = \sqrt{\frac{\sum y^2}{n - 1} (n - 1)} = \sqrt{\sum y^2}$$

Expression (3.2) can be written as

$$r_{12} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 \sum y_2^2}} \quad (3.3)$$

which is preferable for computation because it does not require you to compute two separate standard deviations, each requiring the finding of a square root, but employs the two sums of squares (obtained earlier than standard deviations) and requires the extraction of only a single square root. To state Expression (3.2) more generally for variables Y_j and Y_k , we can write it as

$$r_{jk} = \frac{\sum y_j y_k}{(n - 1) s_j s_k} \quad (3.4)$$

3.6.3. Computation of the product-moment correlation coefficient:

Assuming that we have two variables, e.g. Y and Y , for determining the correlation coefficient, then the computation will be as follows:

1. $\sum Y_1$ = the sum of observed values of the variable Y_1 .
2. $\sum Y_1^2$ = the sum square of observed values of the variable Y_1 .
3. $\sum Y_2$ = the sum of observed values of the variable Y_2 .
4. $\sum Y_2^2$ = the sum square of observed values of the variable Y_2 .

5. $\Sigma Y_1 Y_2$ = the sum of variables Y_1 and Y_2 multiplied.

$$6. \text{ Sum of squares of } Y_1 = \Sigma y_1^2 = \Sigma Y_1^2 - \frac{(\Sigma Y_1)^2}{n}$$

$$7. \text{ Sum of squares of } Y_2 = \Sigma y_2^2 = \Sigma Y_2^2 - \frac{(\Sigma Y_2)^2}{n}$$

$$8. \text{ Sum of products} = \Sigma y_1 y_2 = \Sigma Y_1 Y_2 - \frac{(\Sigma Y_1)(\Sigma Y_2)}{n}$$

9. Product-moment correlation coefficient by
Expression (3.3) =

$$r_{12} = \frac{\Sigma y_1 y_2}{\sqrt{\Sigma y_1^2 \Sigma y_2^2}}$$

3.6.4. Test of significance of correlation coefficients:

Test of the null hypothesis $H_0: \rho = 0$ versus $H: \rho \neq 0$

The simplest procedure is to consult the tabulated critical values of r (refer to Appendix I) for $df = n - 2$. If the absolute value of the observed r is greater than the tabulated value in the column for two variables, we reject the null hypothesis.

The 9 steps mentioned in 3.6.3. have been employed during the computation of the correlation coefficients.

The null hypothesis is tested by means of the t-distribution (with $n - 2$ df) by using the standard error of r . When $p = 0$,

$$s_r = \sqrt{(1 - r^2) / (n - 2)} \quad (3.4)$$

$$\text{Therefore, test statistics } t = \frac{r - p}{s}$$

$$t_s = \frac{(r - 0)}{\sqrt{(1 - r^2) / (n - 2)}} \quad (3.5)$$

For a one-tailed test the 0.10 and 0.02 values of t should be used for 5% and 1% significance tests, respectively. Such tests would apply if the alternative hypothesis were $H_1: p > 0$ or $H_1: p < 0$, rather than $H_1: p \neq 0$.

The expression (3.5) may be employed for testing the correlation coefficients, to see if they are significant or not. In the present study, however, the test of significance of correlation coefficients was done by comparing the respective correlation coefficient values against the tabulated critical values provided in Appendix I. (1)

(1) Sokal, Robert R., and F. James Rohlf. 1969. Biometry. San Fransisco: W. H. Freeman, p. 516.

LEGEND

J	: The first jury
JJ	: The second jury
1	: Assessor number one, native speaker of English, male.
2	: Assessor number two, native speaker of English, female.
3	: Assessor number three, non-native speaker of English, female.
4	: Assessor number four, non-native speaker of English, male.
P	: Pronunciation.
F	: Fluency.
C	: Comprehension.
S	: Sticking to the subject.
G	: Grammatical accuracy.
NAvg	: Native average.
Non-NAvg	: Non-native average.
Tot.Avg	: Total average.
Fem.	: Female.
Male	: Male.

NOTE : Numbers 1-4 refer only to each member in JJ, as, in fact, a record of the individual marks (assigned by each member of J) was not available to the researcher.

RESEARCH QUESTIONS

The statistical data obtained were analyzed to provide possible answers to the research questions cited below:

I- In terms of the average grades given by JJ and J for the five areas (Pronunciation, Fluency, Comprehension, Sticking to the Subject, and Grammatical Accuracy) on which the students were evaluated:

Is there significant agreement between the evaluation of the assessors of JJ (JJ1 vs. JJ2, JJ1 vs. JJ3, JJ1 vs. JJ4, JJ2 vs. JJ3, JJ2 vs. JJ4, JJ3 vs. JJ4, JJNAvg. vs. JJNon-NAvg., JJNAvg. vs. JTot.Avg., JJNon-NAvg. vs. JTot.Avg., JJTot.Avg. vs. JTot.Avg.)?

(See, Table 1 (Appendix I), Correlation Matrix 1, and Summary Table)

II- Is there significant agreement, in terms of the grade given (by both JJ and J) on each of the areas (Pronunciation, Fluency, Comprehension, Sticking to the Subject, Grammatical Accuracy) between :

The evaluation of the assessors in JJ (JJP1 vs. JJP2, JJP1 vs. JJ3, JJP1 vs. JJP4, JJP2 vs. JJP3, JJP2 vs. JJP4, JJP3 vs. JJP4, JJPNAvg. vs. JJPNon-NAvg., JJPTot.Avg. vs. JPTot.Avg.)?

(See, Table 2 (Appendix II) Correlation Matrices 2, 2.1, 2.2, 2.3, 2.4, 2.5 and Summary Tables 2, 2.1, 2.2, 2.3, 2.4, 2.5)

III- Is there significant agreement, in terms of gender between:

- a) The evaluation all of the students' oral proficiency carried out by the female assessors and that carried out by the male assessors within JJ?

(See (Table 3 in Appendix II and) Correlation Matrix 3 and Summary Table 3)

- b) The evaluation of the female students carried out by both the male and the female assessors within JJ?.

(See (Table 3.4 in Appendix II and) Correlation Matrix 3.4 and Summary Table 3.4.

- c) The evaluation of the male students carried out by both the male and the female assessors within JJ?.

(See (Table 3.5 in Appendix II and) Correlation Matrix and Summary Table 3.5.

CHAPTER IV

PRESENTATION AND EVALUATION OF FINDINGS

In this chapter, all the correlation coefficients for the variables (which are of immediate concern to the researcher, as already mentioned on pp. 52-53) are presented in tables, and they (correlation coefficients) are compared against the tabulated critical values (to be found in Appendix I). The significant ones are marked with a *, to help the reader easily notice them. On the basis of the statistical findings (as to whether or not the computed correlation coefficients are significant), the researcher will draw conclusions (and try to make some generalizations limited to the present study), and offer suggestions to help improve the conditions under which oral-proficiency examinations are administered at the ELT Department.

The reader must also note that the tables below contain data (variables) only part of which has been of immediate concern to the researcher, the rest being included for the benefit of possible future investigators, and dealt with only occasionally. It is only the former data (those of immediate concern to the present researcher) that have been dealt with in the notes under the tables.

CORRELATION MATRIX 1

	JJ1	JJ2	JJNAvg.	JJ3	JJ4	JJNon-NAvg.	JJTot.Avg.	JTot.Avg.
JJ1	—	-.084	.761*	.140	.003	.116	.564*	.218
JJ2		—	.582*	-.174	.488	.502	.647*	.392
JJNAvg.			—	.001	.320	.421	.881*	.432
JJ3				—	-.190	-.178	-.092	-.077
JJ4					—	.931*	.698*	.391
JJNon-NAvg.						—	.800*	.479
JJTot.Avg.							—	.536
JTot.Avg.								—

*p <.01, one-tailed test

4.1. Agreement in terms of the average grades given (for Overall Proficiency) by the two juries.

When Correlation Matrix 1 is examined, one comes to realize that the results of the Pearson product-moment correlations do not indicate (either interrater or interjury) significant agreement among the variables mentioned earlier. (See pp. 52-53.) This is so, although the correlation coefficient between JJ1 and JJNAvg. and the correlation coefficient between JJ2 and JJNAvg. are spuriously inflated because JJ1 and JJ2 both contribute to JJNAvg. independently.

A more detailed picture we can get is that assessors -- JJ1, JJ2, JJ3, and JJ4 -- do not agree significantly among themselves; no significant agreement exists between native (JJNAvg.) and non-native (JJNon-NAvg.) assessors; neither is there any significant agreement between JJTot.Avg. and JTot.Avg or between JJNAvg. and JJNon-NAvg.

SUMMARY TABLE 1

(SCORE PORTRAIT OF EACH OF THE ASSESSORS WITHIN J AND OF JJ)

Variable	Mean	Standard deviation
JJ1	52.364	15.133
JJ2	63.864	12.061
JJNAvg.	58.114	9.270
JJ3	83.227	19.535
JJ4	69.455	9.758
JJNon-NAvg.	78.614	7.319
JJTot.Avg.	68.364	7.012
JTot.Avg.	63.046	12.105

As can be seen in Summary Table 1, among the four assessors in the second jury, JJ3 shows the largest variation while JJ4 shows the least. As for the comparison of the amount of variation between JJNAvg. and JJNon-NAvg., the latter shows the smallest amount of variation (around the mean).

CORRELATION MATRIX 2

	P1	P2	PNAvg.	P3	P4	PNon-NAvg.	PTot.Avg.
P1	---	-.139	.891*	-.245	-.148	-.234	.415
P2		---	.326	.223	.390	.374	.509
PNAvg.			---	-.132	.037	-.052	.630*
P3				---	.376	.813*	.544*
P4					---	.846*	.683*
PNon-NAvg.						---	.743*
PTot.Avg.							---

*p <.01, one-tailed test

4.2. Interrater (JJ) agreement in terms of P. (Correlation Matrices and Summary Tables 2-2.5).

Correlation Matrix 2 shows that no 2 raters (whether native or non-native) correlate significantly in terms of their evaluation of the students' Pronunciation, even though P1 seems to correlate significantly with PNAvg., a score of which P1 is one of the two constituents. The same holds true for the correlation between PNAvg. and PTot.Avg., the one between P3 and PNon-NAvg., and PTot.Avg., the one between P4 and PNon-NAvg., PTot.Avg., the one between PNon-NAvg. and PTot.Avg.

Furthermore, no significant correlation exists between the native-speaking (PNAvg.) and the non-native-speaking (PNon-NAvg.) assessors. In short, it can be inferred that there is no significant interrater agreement among the assessors in the second jury in terms of their assessment of the students' Pronunciation.

SUMMARY TABLE 2.

(SCORE PORTRAIT OF EACH OF THE ASSESSORS
WITHIN JJ : IN TERMS OF PRONUNCIATION)

Variable	Mean	Standard deviation
P1	9.955	3.047
P2	13.955	1.463
PNAvg.	11.955	1.595
P3	16.591	2.130
P4	14.909	2.328
PNon-NAvg.	15.750	1.850
PTot.Avg.	13.852	1.189

In Summary Table 2, P2 shows the least variation, while P1 shows the most among the four. When the native-speaking (PNAvg.) and non-native-speaking (PNon-NAvg.) assessors are compared, they can be seen to show approximately the same amount of variation (around the mean).

CORRELATION MATRIX 2.1

	F1	F2	FNAvg.	F3	F4	FNon-NAvg.	FTot.Avg.
F1	---	-.038	.707*	.425	-.015	.223	.541*
F2		---	.680*	.312	.543*	.492	.676*
FNAvg.			---	.532	.373	.512	.875*
F3				---	.542*	.867*	.801*
F4					---	.889*	.720*
FNon-NAvg.						---	.863*
FTot.Avg.							---

*p <.01, one-tailed test

4.3. Interrater (JJ) agreement in terms of F.

In terms of interrater reliability among the four jurymembers, significant correlation exists only between the pairs F2-F4 and F3-F4. However, F2 and F3 do not significantly correlate themselves. Also, no significant correlation is observed between the native and non-native-speaking (FNAvg. vs. FNon-NAvg.) assessors, in terms of their evaluation of the students' Fluency. In this matrix (Matrix 2.1 above), correlation coefficients between F1 and FNAvg.; F1 and FTot.Avg.; F2 and FNAvg.; F2 and FTot.Avg.; FNAvg. and FTot.Avg.; F3 and FNon-NAvg.; F3 and FTot.Avg.; F4 and FNon-NAvg.; F4 and FTot.Avg.; and FNon-NAvg. and FTot.Avg. are, again, spuriously inflated, as the former variable in each pair contributes to the latter variable in the (same) pair.

SUMMARY TABLE 2.1

(SCORE PORTRAIT OF EACH OF THE ASSESSORS
WITHIN JJ: IN TERMS OF FLUENCY)

Variable	Mean	Standard deviation
F1	10.500	3.529
F2	11.682	3.400
FNAvg.	11.091	2.404
F3	16.273	2.511
F4	13.455	2.738
FNon-NAvg.	14.864	2.305
FTot.Avg.	12.977	2.047

In the table above, F1 shows the largest standard deviation (3.529) and F3 the smallest (2.511). The standard deviation for the native-speaking assessors (FNAvg.) is 2.404 and the one for the non-native-speaking assessors (FNon-NAvg.) is 2.305, showing approximately the same amount of variation (around the mean).

CORRELATION MATRIX 2.2

	C1	C2	CNAvg.	C3	C4	CNon-NAvg.	CTot.Avg.
C1	---	-.009	.678*	.257	.291	.318	.617*
C2		---	.729*	.398	.302	.388	.684*
CNAvg.			---	.468	.421	.503	.925*
C3				---	.499	.794*	.678*
C4					---	.923*	.702*
CNon-NAvg.						---	.793*
CTot.Avg.							---

*p <.01, one-tailed test

4.4. Interrater (JJ) agreement in terms of C.

No statistically significant interrater agreement can be observed in the table above. In other words, none of the four raters in the second jury correlate significantly among themselves in terms of their evaluation of the students' Comprehension. The same lack of significant agreement is seen between the CNAvg. and CNon-NAvg scores.

SUMMARY TABLE 2.2

(SCORE PORTRAIT OF EACH OF THE ASSESSORS
WITHIN JJ: IN TERMS OF COMPREHENSION)

Variable	Mean	Standard deviation
C1	10.500	3.143
C2	12.364	3.374
CNAvg.	11.432	2.295
C3	18.773	1.270
C4	13.136	2.007
CNon-NAvg.	15.955	1.430
CTot.Avg.	13.693	1.629

In this table, C2 shows the largest standard deviation (3.374) and C3 the smallest (1.270). CNAvg. shows a greater variation (2.295) than CNon-NAvg. (1.430) (around the mean).

CORRELATION MATRIX 2.3

	S1	S2	SNAvg.	S3	S4	SNon-NAvg.	STot.Avg.
S1	---	-.049	.755*	.249	.162	-.210	.651*
S2		---	.619*	.298	.471	.445	.642*
SNAvg.			---	.392	.436	.127	.934*
S3				---	.552*	.559*	.628*
S4					---	.474	.698*
SNon-NAvg.						---	.328
STot.Avg.							---

*p <.01, one-tailed test

4.5. Interrater (JJ) agreement in terms of S.

As seen in Correlation Matrix 2.3 above, no significant agreement exists among the assessors outside that between S3 and S4, in terms of their evaluation of Sticking to the Subject (in the students' oral proficiency). It is also seen that the expectation of significant agreement between native-speaking (SNAvg.) and non-native-speaking (SNon-NAvg.) assessors has not been verified.

SUMMARY TABLE 2.3

(SCORE PORTRAIT OF EACH OF THE ASSESSORS WITHIN
JJ: IN TERMS OF STICKING TO THE SUBJECT)

Variable	Mean	Standard deviation
S1	10.273	3.807
S2	11.727	3.180
SNAvg.	11.000	2.420
S3	17.909	1.231
S4	13.046	1.759
SNon-NAvg.	15.023	2.598
STot.Avg.	13.239	1.630

In this table, S1 shows the largest and S3 the smallest standard deviation among the four assessors. SNon-NAvg. shows a slightly greater standard deviation (2.598) than SNAvg. (2.420) (around the mean).

CORRELATION MATRIX 2.4

	G1	G2	GNAvg.	G3	G4	GNon-NAvg.	GTot.Avg.
G1	---	.024	.798*	.425	-.104	.158	.600*
G2		---	.621*	.343	.405	.485	.663*
GNAvg.			---	.540*	.162	.417	.871*
G3				---	.201	.698*	.726*
G4					---	.841*	.560*
GNon-NAvg.						---	.810*
GTot.Avg.							---

*p <.01, one-tailed test

4.6. Interrater (JJ) agreement in terms of G.

As can be clearly be observed in Correlation Matrix 2.4, neither of the two raters correlate significantly. In other words, none of the raters (G1, G2, G3, G4) correlate significantly among themselves. Likewise, GNAvg. does not correlate with GNon-NAvg. Therefore we can say that there is not any significant interrater or 'intergroup' (GNAvg. vs. GNon-NAvg.) agreement in terms of their evaluation of the students' Grammatical Accuracy.

SUMMARY TABLE 2.4

(SCORE PORTRAIT OF EACH OF THE ASSESSORS WITHIN
JJ: IN TERMS OF GRAMMATICAL ACCURACY)

Variable	Mean	Standard deviation
G1	11.136	2.660
G2	13.773	2.045
GNAvg.	12.455	1.697
G3	18.227	1.572
G4	15.046	2.081
GNon-NAvg.	16.636	1.424
GTot.Avg.	14.546	1.315

Summary Table 2.4 shows that G1 has the largest standard deviation and G3 the smallest, among the evaluations of the four assessors. Similarly, GNAvg. shows a relatively greater standard deviation (1.697) than GNon-NAvg. (1.424).

CORRELATION MATRIX 2.5

JJ

J

	JJ					J				
	PTot.Avg.	FTot.Avg.	CTot.Avg.	STot.Avg.	GTot.Avg.	PTot.Avg.	FTot.Avg.	CTot.Avg.	STot.Avg.	GTot.Avg.
PTot.Avg.	---	.821*	.762*	.676*	.754*	.254	.273	.330	.317	.356
FTot.Avg.		---	.878*	.749*	.870*	.369	.423	.527	.577*	.593*
JJ CTot.Avg.			---	.797*	.832*	.516	.490	.594*	.520	.558*
STot.Avg.				---	.742*	.409	.341	.378	.431	.486
GTot.Avg.					---	.395	.447	.512	.437	.491
PTot.Avg.						---	.700*	.814*	.557*	.682*
FTot.Avg.							---	.787*	.676*	.803*
J CTot.Avg.								---	.548*	.689*
STot.Avg.									---	.843*
GTot.Avg.										---

*p < .01, one-tailed test

4.7. Interrater (JJ) and Interjury (JJ vs. J) agreement in terms of the five areas.

When the two juries (the first and the second) are compared in terms of their evaluation of the five areas of the students' oral proficiency, the only significant correlation (agreement) is seen to exist between JJCTot.Avg. and JCTot.Avg.

SUMMARY TABLE 2.5

(SCORE PORTRAIT OF EACH OF THE ASSESSORS WITHIN J AND JJ: IN TERMS OF EACH OF THE FIVE AREAS RATED IN THE STUDENTS' ORAL PROFICIENCY)

Variable	Mean	Standard deviation
PTot.Avg.	13.852	1.189
FTot.Avg.	12.977	2.047
JJ CTot.Avg.	13.693	1.629
STot.Avg.	13.239	1.630
GTot.Avg.	14.546	1.315
PTot.Avg.	12.364	2.237
FTot.Avg.	12.409	3.673
J CTot.Avg.	12.455	3.277
STot.Avg.	13.364	2.105
GTot.Avg.	12.364	2.498

The second jury shows the largest standard deviation (2.047) in terms of their evaluation of the students' Fluency and the smallest standard deviation (1.315) in terms of their evaluation of the students' Grammatical Accuracy. The first jury, on the other hand, shows the largest standard deviation (3.673) in terms of their evaluation of the students' Fluency and smallest standard deviation (2.105) in terms of their evaluation of the students Sticking to the Subject.

CORRELATION MATRIX 3

	JJFem.	JJMale
JJFem.	---	.368
JJMale		---

*p <.01, one-tailed test

4.8. Role of gender in the rating of the students' oral proficiency.

4.8.1. Agreement between male and female assessors (within JJ), in terms of their evaluation of all the students (male and female).

Correlation Matrix 3 shows that male and female assessors have not correlated significantly in terms of their evaluation of the students' over-all proficiency.

SUMMARY TABLE 3

Variable	Mean	Standard deviation
JJFem.	75.818	7.928
JJMale	60.909	9.015

In terms of the evaluation of the students' overall oral proficiency, male assessors show a greater standard deviation than their female counterparts (9.015 and 9.728, respectively).

CORRELATION MATRIX 3.1

	JJFem.	JJMale
JJFem.	---	.446
JJMale		---

*p <.01, one-tailed test

4.8.2. Agreement between male and female assessors (within JJ)
in terms of their evaluation of the female students.

In terms of the evaluation of the overall oral proficiency of the female students, no significant agreement appears between male and female assessors.

SUMMARY TABLE 3.1

Variable	Mean	Standard deviation
JJFem.	74.406	8.003
JJMale	61.688	10.065

It is seen in this table that the male assessors show a greater standard deviation than their female counterparts (10.065 and 8.003, respectively).

CORRELATION MATRIX 3.2

	JJFem.	JJMale
JJFem.	---	.406
JJMale		---

*p <.01, one-tailed test

4.8.3. Agreement between male and female assessors (within JJ) in terms of their measurements of the male students' oral proficiency.

It is evident (from Correlation Matrix 3.2 and Summary Table 3.2) that male and female assessors do not show significant agreement in terms of their evaluation of the male students' overall proficiency.

SUMMARY TABLE 3.2

Variable	Mean	Standard deviation
JJFem.	79.583	6.960
JJMale	58.833	5.511

It is seen in Summary Table 3.2 that female assessors show a relatively greater standard deviation than their male counterparts (6.960 and 5.511, respectively).

CHAPTER V

DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Discussion

It has already been shown that by means of tabulated statistical data that:

- 1) JJ and J did not seem to approach the problem exactly in the same way; by and large, there was no significant correlation between the two juries; also none existed among the members within JJ (although they had been provided with a more detailed and controlled rating scale, which was expected to minimize differences).
- 2) Of the five areas evaluated in the students' oral proficiency, P, C and G were the most difficult to rate, as was seen in Correlation Matrices 2.2 and 2.4. The assessors also found it difficult to rate S. F did not present much difficulty as it was the area in the students' oral proficiency on which the assessors seemed to agree most. In terms of inter-jury agreement, all areas, except C, were too ambiguous and complicated to rate consistently.
- 3) Male (JJMale) and female (JJFem.) assessors did not produce similar evaluations, neither of the students of the same sex nor of those of the opposite sex.

5.2. Conclusion

Preliminary remark: The sample in this study was quite homogeneous by virtue of having taken the written examination and their cut-off scores in that examination. In view of this, one may have expected little variation in the results between the two groups in terms of proficiency.

My figures give evidence that:

- There is a significant difference between the evaluation of the native-speaking assessors and that of the non-native-speaking assessors. This would support the proposition that there is an "education-culture difference" between the two, which significantly influences their assessment. In other words, differences in their education and their cultural background lead them to value different aspects of the students' oral proficiency.
- There are also significant differences within/and between the two teams of judges; this, again, would seem to provide evidence that assessments of oral performance in interview techniques are intrinsically subjective, for the time being, at least.
- There is strong evidence that grades awarded to a student for each of the (five) areas assessed in his or her oral proficiency do not correlate significantly.
- Although there was little agreement among assessors in this study, the non-native-speaking assessors

tended to correlate slightly better than the native-speaking assessors in certain areas (e.g. F and S).

Interrater differences may be attributable to the facts that assessors:

- valued different aspects of the students' oral proficiency.
- were inconsistent in applying their standards (whatever they may have been).
- did not have enough training and/or practice in what they were doing (evaluation).
- (particularly those in JJ) did the evaluation under differing conditions. (The actual evaluation, on the other hand) was done simultaneously, while all the four members of the first jury were present.)
- Almost all the assessors have come to a different interpretation of the rating scale used. (This might be due to the fact that the rating scale did not provide a great spread of marks.)

What may be concluded in terms of the five areas (Pronunciation, Fluency, Comprehension, Sticking to the Subject, and Gramattical Accuracy) is that the five terms above:

- did not mean the same to the assessors in the two juries.
- were not defined well enough to be free from misinterpretations.

Of course, one trouble with the research findings is that we cannot adequately determine cause and effect. More detailed and careful studies, in this regard, would be most useful, as this study in itself is, by no means, exhaustive.

In addition to what the data have shown, I also received some comments--written and oral--from, at least, two of the assessors within JJ to the fact that "there were too many sections, which made the evaluation difficult." One could not be so precisely critical of each of the criteria for assessment, as it was difficult to keep, in one's mind, a precise record of all the details.

On the basis of the overall findings then, one may conclude that, at least indirectly, the null hypothesis that there would be no agreement between the two juries is retained, while the research hypothesis is rejected.

The results also indicate that the screening/placement procedures are not reliable, as the scores are not reproduceable on another similar occasions.

5.3 Recommendations (especially in terms of implications for the training and monitoring of testers):

As the research findings, and the empirical evidence at hand, confirm, the following suggestions would be timely and relevant:

- 1) Multiple raters should be involved in the actual evaluation procedures. This would, at least, help minimize differences, and increase the reliability.
- 2) A kind of training program could be set up to allow

present and future assessors to assess recorded tests and discuss scores.

- 3) Future assessors should observe live tests with experienced assessors.
- 4) Future assessors should gradually start taking part in live interview sessions and help with scoring.
- 5) They should keep practicing until difference is not significant.
- 6) An improved rating scale which would be analytic enough and provide a greater spread of marks would help further increase the reliability.
- 7- Similar research needs to be conducted at specified intervals, aimed at ascertaining inter- and intra-rater reliabilities, and at determining how well the rating scale is functioning.
- 8) A test manual could be developed which would familiarize assessors with the test itself, administration of the test, and its scoring procedures.
- 9) Persons of both sexes should assess the candidates, as gender may bring differences into the interpretation and scoring.



APPENDICES

APPENDIX I

(Used by the first jury)

[illegible]

TABLE D'

Values of r at the 5% and 1%
Levels of Significance

Degrees of Freedom (df) ²	Significance .050	Level .010	Degrees of Freedom (df)	Significance .050	Level .010
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081

1. Spence, Janet T., et al., *Elementary Statistics, Second Edition*. New York: Appleton-Century-Crofts, Inc., 1968, p. 236. In turn, a portion of Table D is abridged from Table VI of Fisher and Yates: *STATISTICAL TABLES FOR BIOLOGICAL, AGRICULTURAL, AND MEDICAL RESEARCH*, 4th. edition 1953, published by Oliver and Boyd, Limited, Edinburgh, by permission. The remainder of the table is from *STATISTICAL METHODS*, by George W. Snedecor, 1946 by the Iowa State College Press, Ames, Iowa by permission.

2. $df = N - 2$, where N is the number of pairs of scores.

Source: Isaac, Stephen 1971. *Handbook in Research and Evaluation*. San Diego: EdITS. P.182.



APPENDIX II

TABLE 1

(AVERAGE SCORES ASSIGNED BY EACH
MEMBER OF JJ, AND THOSE ASSIGNED BY J)

Ss	JJ		JJ			JJ		J
	JJ1	JJ2	NAvg.	JJ3	JJ4	Non-NAvg.	Tot.Avg.	Tot.Avg.
a.	79	67	73	84	58	71	72	58
b.	80	44	62	90	64	77	69.50	53
c.	81	82	81.50	95	94	94.50	88	75
d.	71	70	70.50	89	60	74.50	72.50	80
e.	73	50	61.50	89	68	78.50	70	81
f.	59	50	54.50	83	64	73.50	64	65
g.	43	56	49.50	91	71	81	65.25	46
h.	54	78	66	99	69	84	75	67
i.	55	71	63	92	79	85.50	74.25	63
j.	51	59	55	85	66	75.50	65.25	54
k.	47	68	57.50	82	71	76.50	67	53
l.	47	63	55	93	63	78	66.50	66
m.	45	76	60.50	90	66	78	69.25	62
n.	47	61	54	90	75	82.50	68.25	54
o.	37	63	50	80	55	67.50	58.75	56
p.	38	55	46.50	88	74	81	63.75	56
q.	34	76	55	95	85	90	72.50	96
r.	47	40	43.50	80	56	68	55.75	60
s.	44	57	50.50	79	72	75.50	63	60
t.	36	81	58.50	76	65	70.50	64.50	60
u.	35	59	47	81	68	74.50	60.75	47
v.	49	79	64	100	85	92.50	78.25	75

NOTE: JJ is made to refer to the second jury, and J to the first jury. (See p. 54.)

TABLE 2

(AVERAGE SCORES ASSIGNED - IN TERMS OF
PRONUNCIATION - BY EACH ASSESSOR WITHIN JJ)

Ss	P1	P2	PNAvg.	P3	P4	FNon-NAvg.	PTot.Avg.
a.	17	14	15,50	16	13	14,50	15
b.	15	12	13,50	12	14	13	13,25
c.	13	16	14,50	17	19	18	16,25
d.	13	14	13,50	16	12	14	13,75
e.	14	13	13,50	16	16	16	14,75
f.	12	12	12	14	13	13,50	12,75
g.	9	14	11,50	18	16	17	14,25
h.	10	16	13	20	15	17,50	15,25
i.	9	15	12	18	17	17,50	14,75
j.	8	13	10,50	17	13	15	12,75
k.	11	15	13	16	16	16	14,50
l.	9	14	11,50	19	14	16,50	14
m.	10	16	13	18	14	16	14,50
n.	9	14	11,50	19	16	17,50	14,50
o.	6	12	9	16	11	13,50	11,25
p.	7	13	10	19	15	17	13,50
q.	5	16	10,50	16	18	17	13,75
r.	11	12	11,50	15	11	13	12,25
s.	7	14	10,50	14	15	14,50	12,50
t.	8	16	12	14	15	14,50	13,25
u.	7	14	10,50	15	15	15	12,75
v.	9	12	10,50	20	20	20	15,25

TABLE 2.1

(AVERAGE SCORES ASSIGNED - IN TERMS OF
FLUENCY - BY EACH ASSESSOR WITHIN JJ)

Ss	F1	F2	FNAvg.	F3	F4	FNon-NAvg.	FTot.Avg.
a.	15	12	13,50	14	9	11,50	12,50
b.	16	7	11,50	19	11	15	13,25
c.	18	16	17	18	19	18,50	17,75
d.	14	12	13	16	11	13,50	13,25
e.	15	8	11,50	19	12	15,50	13,50
f.	11	8	9,50	15	12	13,50	11,50
g.	8	9	8,50	17	14	15,50	12
h.	14	16	15	20	14	17	16
i.	12	13	12,50	18	17	17,50	15
j.	9	11	10	15	14	14,50	12,25
k.	10	12	11	14	14	14	12,50
l.	10	12	11	17	12	14,50	12,75
m.	8	16	12	16	12	14	13
n.	10	12	11	18	15	16,50	13,75
o.	7	11	9	12	10	11	10
p.	8	9	8,50	15	15	15	11,75
q.	6	16	11	20	18	19	15
r.	9	5	7	13	9	11	9
s.	10	10	10	15	15	15	12,50
t.	5	16	10,50	12	13	12,50	11,50
u.	6	9	7,50	15	13	14	10,75
v.	10	17	13,50	20	17	18,50	16

TABLE 2.2

(AVERAGE SCORES ASSIGNED -IN TERMS OF
COMPREHENSION - BY EACH ASSESSOR WITHIN JJ)

Ss	C1	C2	CNAvg.	C3	C4	CNon-NAvg.	CTot.Avg.
a.	16	14	15	18	13	15,50	15,25
b.	16	7	11,50	20	14	17	14,25
c.	17	17	17	20	18	19	18
d.	14	14	14	20	12	16	15
e.	15	9	12	19	12	15,50	13,75
f.	12	8	10	16	12	14	12
g.	8	10	9	18	12	15	12
h.	11	17	14	20	12	16	15
i.	11	14	12,50	19	15	17	14,75
j.	11	11	11	19	13	16	13,50
k.	8	13	10,50	17	12	14,50	12,50
l.	9	13	11	20	11	15,50	13,25
m.	9	15	12	19	12	15,50	13,75
n.	9	11	10	20	15	17,50	13,75
o.	7	10	8,50	17	10	13,50	11
p.	8	10	9	20	15	17,50	13,25
q.	9	17	13	20	17	18,50	15,75
r.	9	7	8	18	12	15	11,50
s.	9	10	9,50	17	14	15,50	12,50
t.	8	17	12,50	18	11	14,50	13,50
u.	7	11	9	18	12	15	12
v.	8	17	12,50	20	15	17,50	15

TABLE 2.3

(AVERAGE SCORES ASSIGNED - IN TERMS OF
STICKING TO THE SUBJECT - BY EACH ASSESSOR WITHIN JJ)

Ss	S1	S2	SNAvg.	S3	S4	SNon-NAvg.	STot.Avg.
a.	14	14	14	19	11	15	14,50
b.	18	7	12,50	19	13	16	14,25
c.	17	17	17	20	19	19,50	18,25
d.	16	15	15,50	17	13	15	15,25
e.	17	8	12,50	17	13	5	13,75
f.	13	10	11,50	19	12	15,50	13,50
g.	7	11	9	19	13	16	12,50
h.	9	13	11	19	13	16	13,50
i.	10	15	12,50	19	13	16	14,25
j.	11	8	9,50	16	11	13,50	11,50
k.	8	14	11	17	13	15	13
l.	9	8	8,50	18	12	15	11,75
m.	8	13	10,50	17	13	15	12,75
n.	8	10	9	17	13	15	12
o.	8	10	9	17	11	14	11,50
p.	8	9	8,50	18	13	15,50	12
q.	5	12	8,50	19	16	17,50	13
r.	8	7	7,50	17	12	14,50	11
s.	9	12	10,50	17	13	15	12,75
t.	6	16	11	16	12	14	12,50
u.	8	12	10	17	13	15	12,50
v.	9	17	13	20	15	17,50	15,25

TABLE 2.4

(AVERAGE SCORES ASSIGNED - IN TERMS OF
GRAMMATICAL ACCURACY - BY EACH ASSESSOR WITHIN JJ)

Ss	G1	G2	GNAvg.	G3	G4	GNon-NAvg.	GTot.Avg.
a.	17	13	15	17	12	14,50	14,75
b.	15	11	13	20	12	16	14,50
c.	16	16	16	20	19	19,50	17,75
d.	14	15	14,50	20	12	16	15,25
e.	12	12	12	18	15	16,50	14,25
f.	11	12	11,50	19	15	17	14,25
g.	11	12	11,50	19	16	17,50	14,50
h.	10	16	13	20	15	17,50	15,25
i.	13	13	13	18	17	17,50	15,25
j.	12	16	14	18	15	16,50	15,25
k.	10	14	12	18	16	17	14,50
l.	10	16	13	19	14	16,50	14,75
m.	10	16	13	20	15	17,50	15,25
n.	11	14	12,50	16	16	16	14,25
o.	9	13	11	18	13	15,50	13,25
p.	7	14	10,50	16	16	16	13,25
q.	9	15	12	20	19	19,50	15,75
r.	10	9	9,50	17	12	14,50	12
s.	9	11	10	16	15	15,50	12,75
t.	9	16	12,50	16	14	15	13,75
u.	7	13	10	16	15	15,50	12,75
v.	13	16	14,50	20	18	19	16,75

TABLE 3

(TOTAL AVERAGE SCORES ASSIGNED (TO EACH OF THE
SUBJECTS -MALE AND FEMALE) BY JJ)

Ss	JJFem. (Assessor)	JJMale (Assessor)
a	75,50	68,50
b	67	72
c	88,50	87,50
d	79,50	65,50
e	69,50	70,50
f	66,50	61,50
g	73,50	57
h	88,50	61,50
i	81,50	67
j	72	58,50
k	75	59
l	78	55
m	83	55,50
n	75,50	61
o	71,50	46
p	71,50	56
q	85,50	59,50
r	60	51,50
s	68	58
t	78,50	50,50
u	70	51,50
v	89,50	67

TABLE 3.1

(TOTAL AVERAGE SCORES ASSIGNED (TO EACH OF THE
SUBJECTS - FEMALE ONLY) BY JJ)

Ss	JJFem. (Assessor)	JJMale (Assessor)
a	75,50	68,50
b	67	72
c	88,50	87,50
d	79,50	65,50
e	69,50	70,50
f	66,50	61,50
g	73,50	57
k	75	59
l	78	55
m	83	55,50
n	75,50	61
o	71,50	46
r	60	51,50
s	68	58
u	70	51,50
v	89,50	67

TABLE 3.2

(TOTAL AVERAGE SCORES ASSIGNED (TO EACH OF THE
SUBJECTS - MALE ONLY) BY JJ)

Ss	JJFem. (Assessor)	JJMale (Assessor)
----	----------------------	----------------------

h	88,50	61,50
i	81,50	67
j	72	58,50
p	71,50	56
q	85,50	59,50
t	78,50	50,50

BIBLIOGRAPHY

- Allen, Harold B., and Russell N. Campbell. 1972. Teaching English As A Second Language. New York: McGraw-Hill.
- Allen, Harold B., and Russell N. Campbell (eds). 1972. Teaching English as a Second Language: A Book of Readings. New York: McGraw-Hill
- Allen, J.P.B., and Alan Davies (eds). 1977. Testing and Experimental Methods. Oxford: Oxford UP.
- Allen, Virginia French. 1975. "Toward a Thumbnail Test of English Competence" in Newton 1975, pp. 320-321.
- Bachman, Lyle F., and Adrian S. Palmer. 1983. "The construct validity of the FSI Oral Interview" in Oller 1983, pp. 154-169.
- Berkoff, Nelson A. 1985. "Testing Oral Proficiency: A New Approach" in Lee et al 1985, pp. 93-99.
- Bott, Donald E. 1982. "Communicative Competence" in English Teaching Forum (20:2). pp. 47-48.
- Boyle, Joseph. 1984. "Being Philosophical about the Discrete-Point versus Global-Proficiency Debate" in English Teaching Forum. (22:1). pp. 20-23.
- Briere, Eugene J., and Richard H. Brown. 1975. "Norming Tests of ESL Among Amerindian Children" in Palmer and Spolsky 1975, pp. 121-127.

- Brière, Eugène J., and F.B. Hinofotis (eds). 1979. Concepts in Language Testing: Some Recent Studies. Washington, D. C.: U.S. Information Agency.
- Brière, Eugène J. 1972. "Are We Really Measuring Proficiency With Our Foreign Language Tests?" in Allen and Campbell 1972, pp. 321-329.
- Brown, Gillian. 1990. "Cultural Values: The Interpretation of Discourse" in English Language Teaching Journal (44:1). pp. 66-69.
- Brumfit, Christopher. 1984. Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy. Cambridge: Cambridge University Press.
- Canale, Michael. 1983. "On some dimensions of language proficiency" in Oller 1983, pp. 333-342.
- Candlin, Christopher N. (ed). 1983. The Communicative Teaching of English: Principles and an Exercise Typology. Harlow, Essex: Longman.
- Carroll, Brendan J. 1980. Testing Communicative Performance: An Interim Study. Oxford: Pergamon Press.
- Carroll, Brendan J., and Patrick J. Hall. 1985. Make Your Own Language Tests: A Practical Guide to Writing Language Performance Tests. Oxford: Pergamon Press.
- Carroll, John B. 1972. "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students" in Allen and Campbell 1972, pp. 313-320.
- Carroll, John B. 1983. "Psychometric theory and language testing" in Oller, Jr. 198, pp. 80-107.

- Cartier, Francis A. 1975. "Criterion-Referenced Testing of Language Skills" in Palmer and Spolsky 1975, pp. 19-24.
- Catford, J.C. 1975. "Some Aspects of Linguistics in Language Testing" in Newton 1975, pp. 316-319.
- Clark, John L. D. 1975. "Theoretical and Technical Considerations in Oral Proficiency Testing" in Jones and Spolsky 1975, pp. 10-28.
- Clark, John L. D. 1979. "Direct vs. Semi-Direct Tests of Speaking Ability" in Brière and Hinofotis 1979, pp. 35-49.
- Cohen, Andrew D. 1980. Testing Language Ability in the Classroom. Rowley, Massachusetts: Newbury House.
- Condon, Elaine C. 1975. "The Cultural Context of Language Testing" in Palmer and Spolsky 1975, pp. 205-217.
- Cooper, Robert L. 1972. "Testing" in Allen and Campbell 1972, pp. 330-345.
- Cronbach, Lee J. 1984. Essentials of Psychological Testing. New York: Harper & Row.
- Cummins, Jim. 1983. "Language proficiency and academic achievement" in Oller 1983, pp. 108-129.
- Cziko, Gary A. 1983. "Psychometric and edumetric approaches to language testing" in Oller 1983, pp. 289-307.
- Dobson, Julia M. 1988. Effective Techniques for English Conversation Groups. Washington, D.C.: Information Agency.

Farhady, Hossein. 1983. "New directions for ESL proficiency testing" in Oller 1983, pp. 253-269.

Farhady, Hossein, 1983. "On the plausibility of the unitary language proficiency factor" in Oller 1983, pp. 11-28.

Finocchiaro, Mary., and S. Sako. 1983. Foreign Language Testing: A Practical Approach. New York: Regents

Freed, Barbara F. 1984. "Proficiency in Context: The Pennsylvania Experience" in Savignon and Berns 1984, pp. 211-226.

Fulcher, Glenn. 1990. "Testing Spoken Language" in English Language Teaching Journal (44:1), pp. 80-81.

Gefen, Rafael. 1983. "Testing Listening Comprehension by Radio" in English Teaching Forum (21:4), p. 37.

Griffin, Patrick E. 1985. "The Use of Latent Trait Models in the Calibration of Tests of Spoken Language in Large-scale Selection-placement Programs" in Lee et al 1985, pp. 149-161.

Harris, David P. 1969. Testing English as a Second Language. New York: McGraw-Hill.

Heaton, J. Brian. 1982. Language Testing. Kettering, Shire: Modern English Publications.

Henning, Grant. 1982. "Twenty Common Testing Mistakes for EFL Teachers to Avoid" in English Teaching Forum (20:3). pp. 33-37.

- Henning, Grant. 1987. A Guide to Language Testing: Development-Evaluation-Research. Cambridge, Massachusetts: Newbury House.
- Hubbard, Peter., et al. 1984. A Training Course for TEFL. Oxford: Oxford UP.
- Isaac, Stephen. 1971. Handbook in Research and Evaluation. San Diego: EdITS.
- James, Charles J. (ed). 1985. Foreign Language Proficiency in the Classroom and Beyond. Lincolnwood, Illinois, U.S.A.: National Textbook Company.
- Jones, Randall L., and B. Spolsky (eds). 1975 Testing Language Proficiency. Arlington, Virginia: Center for Applied Linguistics.
- Jones, Randall L., 1975. "Testing Language Proficiency in the United States Government" in Jones and Spolsky 1975, pp. 1-9.
- Jones, Randall L. 1979. "Performance Testing of Second Language Proficiency" in Brière and Hinofotis 1979, pp. 50-57.
- Jones, Randall L., 1985. "Some Basic Considerations in Testing Oral Proficiency" in Lee et al 1985, pp. 77-84.
- Johnson, Keith., and Keith Morrow (eds). 1981. Communication in the Classroom: Applications and Methods For a Communicative Approach. Harlow, Essex: Longman.
- Karasar, Niyazi. 1984. Bilimsel Araştırma Yöntemi: Kavramlar, İlkeler, Teknikler. Ankara: Hacettepe TAS.

Kerxinger, Fred N. 1973. Foundations of Behavioral Research. Holt, Rinehart and Winston.

Krowitz, Mary Jo. 1988. English For Professionals: A Teacher's Perspective. Washington, D.C.: U.S. Information Agency.

Lee, Y.P., et al (eds). 1987. New Directions in Language Testing. Oxford: Pergamon Press.

Levenston, E.A. 1975. "Objective Evaluation of Oral Proficiency of Adult Immigrants to Canada" in Palmer and Spolsky 1975, pp. 66-74.

Liskin-Gasparro, Judith E. 1987. Testing and Teaching for Oral Proficiency. Boston, Massachusetts: Educational Testing Service.

Lombardo, Linda. 1984. "Oral Testing: Getting a Sample of Real Language" in English Teaching Forum (22:1), pp. 2-6.

Low, Graham D., and Y.P. Lee. 1985. "How Shall a Test be Referenced?" in Lee et al 1985, pp. 119-126.

Lowe, J. Edgar. 1975. "Are You Fair in Your Grading?" in Newton 1975, pp. 322-327.

Madsen, Harold S. 1983. Techniques in Testing. New York: Oxford UP.

Matthews, Margaret. 1990. "The Measurement of Productive Skills: Doubts Concerning the Assessment Criteria of Certain Public Examinations" in ELT Journal (44:1). pp. 117-120.

- McBeath, Neil. 1990. "C-Tests--Some Words of Caution" in English Teaching Forum (38:2), pp. 45-46.
- Morrow, Keith. 1984. "Testing Performance in Oral Interaction" in Savignon and Berns 1984, pp. 203-209.
- Murphy, Richard A. 1984. Guidelines to Proficiency: A Training Manual For English Instructors. Washington, D.C.: Information Agency.
- Newton, Anne Covell (ed). 1975. The Art of TESOL: Selected Articles from the English Teaching Forum. Washington, D.C.: English Teaching Forum.
- Newton, Anne Covell (ed). A Forum Anthology: Selected Articles from the English Teaching Forum 1979-1983. Washington, D.C.: Information Agency.
- Nickel, Gerhard. 1975. "Problems of Syllabus, Curriculum and Testing in Connection with Modern Language Programs for Adults in Europe" in Jones and Spolsky 1975, pp. 131-138.
- Oller, Jr., John W. 1972. "Dictation as a Test of ESL Proficiency" in Allen and Campbell 1972, pp. 346-354.
- Oller, John W., and Anne Newton. 1974. Testing English As A Foreign Language. Washington, D.C.: U.S. Information Agency.
- Oller, Jr., John W. 1979. "Explaining the Variance in Tests: The Validation Problem" in Brière and Hinofotis 1979, pp. 61-74.
- Oller, Jr., John W. 1979. Language Tests at School. London: Longman.

Oller, Jr., John W (ed). 1983. Issues in Language Testing Research. Rowley, Massachusetts: Newbury House.

Oller, Jr., John W. 1983. "A consensus for the eighties?" in Oller 1983, pp. 351-356.

Oller, Jr., John W. 1983. "Evidence for a General Language Proficiency Factor: an Expectancy Grammar" in Oller 1983, pp. 3-10.

Palmer, Leslie., and B. Spolsky (eds). 1975. Papers on Language Testing. Washington, D.C.: Georgetown UP.

Peng, Khong Chooi. 1986. Testing Performance Through Interaction Tests in UMSEP" in Peterson 1986, pp. 164-172.

Peterson, Pat Wilcox (ed). 1986. ESP in Practice: Models and Challenges for Teachers. Washington, D.C.: U.S. Information Agency.

Redfield, Michael. 1984. "A More 'Functional' Examination" in English Teaching Forum (22:1), pp. 32-33.

Rivers, Wilga M. 1983. Communicating Naturally in a Second Language: Theory and Practice in Language Teaching. Cambridge UP.

Rivers, Wilga M. 1983. Speaking in Many Tongues: Essays in foreign-language teaching. Cambridge: Cambridge UP.

Savignon, Sandra J., and M.S. Berns (eds). 1984. Initiatives in Communicative Language Teaching: A Book of Reading, UK: Addison-Wesley.

Scott, Mary Lee., and Harold S. Madsen. 1983. "The influence of retesting on test effect" in Oller 1983, pp. 270-279.

Spaventa, Louis J. 1981. "Testing" in Guideposts in TEFL. Washington, D.C.: U.S. Information Agency.

Stansfield, Charles W. (ed). 1986. Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference. Princeton, NJ: Educational Testing Service.

Stevenson, Douglas K. 1985. "Pop Validity and Performance Testing" in Lee et al 1985, pp. 111-118.

Streiff, Virginia. 1983. "The roles of language in educational testing" in Oller 1983, pp. 343-350.

Swain, Merrill. 1984. "Large-Scale Communicative Language Testing: A Case Study" in Savignon and Berns 1984, pp. 185-201.

Underhill, Nic. 1989. Testing Spoken Language: A Handbook of Oral Testing Techniques. Cambridge: Cambridge UP.

Upshur, John A. 1972. "Cross-Cultural Testing: What to Test" in Allen and Campbell 1972, pp. 355-366.

Upshur, John A. 1975. "Objective Evaluation of oral proficiency" in Newton 1975, pp. 328-333.

Upshur, John A. 1979. "Functional Proficiency Theory and a Research Role for Language Tests" in Brière and Hinofotis 1979, pp. 75-99.

Upshur, John A. 1979. "Objective Evaluation of Oral Proficiency in the ESOL Classroom" in Palmer and Spolsky 1975, pp. 52-65.

Valette, Rebecca M. 1977. Modern Language Testing. New York: Harcourt Brace Jovanovich.

Vollmer, Helmut J. 1983. "Competing hypotheses about second language ability: a plea for caution" in Oller 1983, pp. 29-79.

Webber, Richard D., and Desmond M. Allison. 1984. "Library Induction and Performative Testing" in English Teaching Forum (22:2), p. 36.

Wilds, Claudia P. 1975. "The Oral Interview Test" in Jones and Spolsky 1975, pp. 29-44.

Wilkinson, Alix. 1985. "An Original Interview Activity" in English Teaching Forum (23:2). pp. 38-39.

Williams, K.L. 1990. "Three New Tests For Overseas Students Entering Postgraduate and Vocational Training Courses" in ELT Journal (44:1). pp. 55-64.

Zappolo, Celeste. 1981. "A Graded Listening-Comprehension Program" in English Teaching Forum (19:4). pp. 31-36.

GLOSSARY

In order to help the reader understand the present writer's understanding, definitions of some words/terms, and phrases frequently used in testing, are presented. Definitions presented here may be universal as well as specific to the contexts in which they appear in this thesis.

The terms, and phrases are mainly taken from Underhill (1989), Testing Spoken Language, and from the article, "Theoretical and Technical Considerations in Oral Proficiency Testing" by Clark (1975), mostly as they are to be found there or slightly adapted, to fit the purposes of this study.

ADDITIVE MARKING: The assessor has a prepared a list of features to listen for during the test. She awards a mark for each of these features that the learner produces correctly, and adds these marks to give the score. (Also known as "incremental mark system".)

ANALYZE: To analyze a test is to find out how well it is functioning as a measurement tool.

ASSESSOR: This is a person who makes an evaluative judgement, of the learner's oral proficiency level, on the basis of the utterances elicited. The assessor is provided with a rating scale on which she will more precisely indicate what the learner's relative level of proficiency may be. It must be noted that the decision to be made, to some extent, is subjective; and will be based on judgement, personal reaction, intuition, and experience (of the assessor). Participation of more than one assessor in the actual evaluation process usually means a more reliable judgement. (The terms 'rater' and 'assessor' have been used interchangeably in this thesis.)

AUTHENTIC: An authentic task is one which resembles very closely something which one actually does in everyday life.

BIAS: (1) This is either the subjectivity shown in scores given by different raters on free-response tests, or (2) a prejudice for or against values, concepts, or principles held by others.

CLOZE: This is a test procedure which elicits the completion of blanks deleted from a text.

COMMUNICATIVE: When a learner says something that is relevant and true (for himself at least), to someone else who is interested and has not heard it before (from that speaker, at least), then that act of speech is communicative.

CORRELATION COEFFICIENT: This (in testing) describes the relationship between two series of scores. In other words, it gives us some ideas about the extent to which the raters of the tests agree (or disagree). The higher the correlation, the more sure is the tester that a learner's score reflects his or her true level of performance on the subject matter of the test.

CORRELATION MATRIX: This is a table of inter-correlations among the subtests of a test battery.

COVARIANCE: This is the relationship, interdependence or "going togetherness" of two variables.

CRITERION: This is an external principle or standard against which a test is judged or validated (such as course grades, teacher's judgement, or another test).

CUT-OFF SCORE: This is the established score level which represents the minimal language proficiency range deemed essential for admission to or success in a specified program. It is the score point which will differentiate the "masters" from the "non-masters". The masters are those who attain the cut-off scores or grades above them.

EFFICIENCY: This refers to obtaining the best assessments one can, within the limits of time and resources available for the construction and administration of the assessments.

EVALUATION: This is the process of deciding upon, obtaining, and providing, information useful for judging educational procedures and decisions. Evaluation thus involves determining what to measure, constructing or selecting measures (e.g., tests, interviews, observations, questionnaires); administering, scoring, and analyzing them; and making judgements about progress being made by students toward the objectives of language learning; the quality of instruction; the materials; the curriculum; and other relevant teaching-learning factors.

FREE RESPONSE: This is the sort of test situation in which respondents have the liberty to say or write what they choose, usually within certain parameters.

GLOBAL PROFICIENCY: This refers to the individual's ability to carry out various language-use tasks appropriately and effectively in realistic communication settings. (Clark, 1979: 37)

GLOBAL PROFICIENCY TEST: This (in the area of speaking) would presumably require the learner to communicate orally in a variety of speech situations typically encountered in real-life language use most notably face-to-face

conversation with one or more native speakers of the target language. Possible real-life contexts (in the testing of oral proficiency) may include such activities as reading aloud (as in giving a prepared speech), dictating into a tape recorder, talking on the telephone, and conversing face-to-face with one or more interlocutors. Whatever the sort of activity or measurement tool used may be, in all cases, the emphasis is on determining the individual's ability to operate effectively in those presumably real-life language use situations. (Clark, 1975: 10)

HALO EFFECT: This refers to higher or lower grade resulting from previous pleasant or unpleasant experiences (with the writer).

IMPRESSION MARKING: In this category, the assessor awards a mark based on the learner's overall performance, without picking out any special features or using a counting system for errors.

INTERLOCUTOR: This is a person who talks with a learner with the specific purpose of encouraging the learner to display, to the assessor, his oral proficiency in the best way possible. An interlocutor is not necessarily an assessor.

INTERVIEWER: is a person who talks to a learner in an oral test and controls to a greater or lesser extent the direction of the conversation. While exercising this control, she may nonetheless yield the initiative to the learner to redirect the discussion to another area/subject. An interviewer also takes the role of the assessor, and is thus involved in the actual assessment.

LEARNER: This is a person who takes a test in English.

MARKER: This refers, in this study, to someone who is not present at the test itself, but later awards marks to the learner on the basis of an audio or video tape recording. This may be either a routine part of the marking system (as a technique used for second marking as in evaluation of taped tests) or an occasional exercise for the purposes of moderation. (Moderating a test is, specifically, to compose the way assessors award marks, and to take steps to reduce any discrepancies.) The main difference between assessors and markers is that markers never meet the individuals to whom they award marks, while assessors do meet them.

MEASURE: (1) (Noun) Any test results or data which can be stated in number. (2) (Verb) To find the degree or level of knowledge, skill, proficiency, or competence which an individual possesses.

OBJECTIVE: An objective test refers to a test situation in which the assessor only has to decide whether the answer the learner has given to each question is correct or the best. While doing so, there is no or almost no personal judgement on the part of the assessor.

ORAL TEST: This is a repeatable procedure in which a learner speaks, and is assessed on the basis of what she says. An oral test can be given alone, or in an integrated examination aimed at testing other skills as well.

PLACEMENT: This is the process of determining, through tests or other means, the most desirable course or program (where there is more than one choice) for a particular student.

PRACTICALITY: This refers to whether or not administering the test is expensive; whether several administrators

and/or scorers are needed (the more personnel who must be involved in giving and scoring a test, the more costly it becomes); how long the administering and scoring of it will take, the ease with which it can be administered, and "ease of interpretation."

RANDOM SAMPLE: This is an unbiased, representative segment of a larger group.

RATER: Although usually used as a synonym for marker, in marker, in this study it will refer to member of the first jury (who did the actual evaluation of the learners' oral proficiency), and to the assessors the second jury (who, for the second time, awarded marks to learners on the basis of the utterances in the recorded version of the same test for the purposes of moderation and inter-rater reliability).

RATING SCALE: This is a series of short descriptions, on a continuum of different levels of language ability. Its purpose is to describe briefly what a typical learner at each level can do, so that it is easier for the assessor to decide what level or score to assign each learner in a test.

RELIABLE TEST: This is an instrument that produces essentially the same results consistently on different occasions when the conditions of the administration remain the same. Reliability can be measured in different ways: (1) by administering the same test to the same individuals, (2) by administering a different but parallel version of the test to the same individuals, and (3) by administering the test once and measuring the results on a "split-halves" basis, where odd-numbered answer results are measured against the even-numbered results for the same students. Two distinct types of reliability are:

intra-rater reliability, which refers to the extent to which a given scorer is able to consistently assign the same scores to individual tests, that he or she evaluates two or more times in succession; and inter-rater reliability, which refers to the extent to which two or more different raters are able to assign the same scores to a given test performance.

RELIABILITY: This is the consistency of scores obtained by students. (Spaventa, 1981:57)

RESPONSE: This is any oral, written, or physical reaction to a stimulus.

SCORABILITY: This is the ease or difficulty of grading tests.

SPEAKING (Oral Production): This is the ability to use pronunciation, stress, intonation, grammatical structure, vocabulary, and cultural insights, in the normal, authentic functions of communication.

SUBJECTIVE: A subjective test is one in which there may be more than one correct answer to each question; and the evaluation of such a test may, to some extent, at least, depend on the biased judgement of a rater.

SUBTRACTIVE MARKING: In this system, the assessor subtracts one mark from a total, for each mistake the learner makes, down to a minimum of zero.

STIMULUS: This is an agent, action, or condition, which elicits or causes response, usually by providing a subject (a picture, a text, or a particular topic) to talk about.

TEST ADMINISTRATION: This is the process by which test stimuli are presented to the learner (Clark, 1975: 10-28).

1. Mechanical administration refers to procedures in which test booklets, tape records, videotapes, or other inanimate devices are for all practical purposes entirely responsible for test administration: A live examiner's involvement in the testing procedure would be limited to giving general directions and handing out test materials.
2. Human administration, on the other hand, requires the presence of a live examiner who is actively and continuously involved in the testing process: reading test questions aloud, conversing with the learner in an interview situation, and so on.

TEST COMPROMISE: This is the acquisition of knowledge by the students of the test content prior to the administration of the test.

TEST SCORING: This is the process by which the learner's response to the test stimuli are converted to quantifiable or numerically codable data.

1. Mechanical scoring is a scoring process in which student responses are converted automatically, i.e. without any thought or judgement on the part of a human rater, to the appropriate score.
2. Human scoring is the other type of scoring in which one or more persons (assessors) must actually listen to the responses of the learner and exercise a certain degree of thought or judgement in arriving at a rating of the learner's performance (Clark, 1975: p. 12).

Test scoring, both mechanical and human, are further divided into two:

Simultaneous scoring is the one that is carried out on the spot, either during or immediately following the test itself, and there is no need to tape record or in any other way to preserve the learner's responses.

2. Delayed scoring is the process in which the responses of the learner are recorded for evaluation at a later time (Clark, 1975: pp. 12-13).

UNDER-ACHIEVER: This is a student who has a higher potential for learning, but who is not attaining grades commensurate with his/her capacity.

VALID: As referred to by Spaventa (1981), the valid test measures what it sets out to measure, to the extent that it successfully accomplishes the operational purpose for which it is being used. (Clark, 1981: 37)

WASHBACK: This is the effect a test has on teaching practices.

WEIGHTING: This is a procedure by which marks are awarded out of the same total for the different marking categories, and these marks are then multiplied by different factors to give them more, or less, influence in the total.

WEIGHTING OF ITEMS: This is the assigning of point scores to items and procedures on a quiz or test; items may all be weighted equally (i.e., fixed-interval weighting) or may be weighted differently according to their importance or difficulty relative to other items on the test.