AN INVESTIGATION OF THE VALIDITY AND RELIABILITY OF THE SPEAKING EXAM AT A TURKISH UNIVERSITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY
GONCA SAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF ARTS
IN
THE DEPARTMENT OF ENGLISH LANGUAGE EDUCATION

SEPTEMBER 2008

Approval of the Graduate School of Social Sciences

____

Prof. Dr. Sencer AYATA
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Arts.

____

Prof. Dr. Wolf KÖNİG
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Arts.

____

Prof. Dr. Hüsnü ENGİNARLAR
Supervisor

Examining Committee Members

Prof. Dr. Hüsnü ENGİNARLAR _________________________

Assist. Prof. Dr. Nurdan GÜRBÜZ _________________________

Dr. Işıl GÜNSELİ KAÇAR _________________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Gonca SAK

Signature :

ABSTRACT

AN INVESTIGATION OF THE VALIDITY AND RELIABILITY OF THE SPEAKING EXAM AT A TURKISH UNIVERSITY

Sak, Gonca

M.A., Program in English Language Teaching

Supervisor: Prof. Dr. Hüsnü ENGİNARLAR

September 2008, 164 pages

This thesis aims to investigate the validity and reliability of the speaking exam at a Turkish University. For this study, data were obtained through questionnaires, interviews, the students' speaking exam results, TOEFL exam results and departmental speaking exam scores. The results of the questionnaire were used to explore the face validity of the speaking exam. The interviews conducted to examine the content validity of the exam were analyzed in detail and common points from each interview were highlighted. To determine the predictive validity of the exam, Pearson Product Moment Correlation Coefficient and Simple Linear Regression Analysis were conducted. Furthermore, to investigate the construct validity of the exam correlation coefficients between speaking test scores and TOEFL subtest scores were calculated. To estimate the intra and inter-rater reliability level of the exam, correlation coefficients were calculated as well.

The analysis of the results of the questionnaire indicated that the exam has satisfactory face validity. Moreover, the results of the interviews showed that the exam possesses the quality of content validity to a moderately high degree. It was found out that the speaking exam given in preparatory year education does not seem to predict the performances of the students in the departmental speaking exam. Moreover, the statistical analyses done to investigate the construct validity of the exam indicated that there are very low correlations between the speaking exam scores and the other subtests.

It was discovered that the inter-rater reliability of the exam was not as satisfactory as it was expected as the inter-rater reliability of one pair was found relatively low. However, the speaking exam seemed to have satisfactory intra-rater reliability.

Key words: Validity, reliability, testing, testing speaking

ÖZ

TÜRKiYE'DEKİ BİR ÜNİVERSİTEDEKİ KONUŞMA SINAVININ GEÇERLİLİK VE GÜVENİRLİLİK ÇALIŞMASI

Sak, Gonca

Yüksek Lisans, İngiliz Dili Eğitimi

Tez Danışmanı: Prof. Dr. Hüsnü ENGİNARLAR

Eylül 2008, 164 sayfa

Bu çalışmanın amacı bir Türk üniversitesindeki konuşma sınavının geçerlilik ve güvenirliğinin araştırılmasıdır. Bu çalışma için veriler anketlerden, görüşmelerden, öğrencilerin konuşma sınav sonuçlarından, TOEFL sınavı sonuçları ve bölüm konuşma sınav sonuçlarından elde edilmiştir. Anketlerin sonuçları konuşma sınavının yüzeysel geçerliliğini belirlemek amacıyla kullanılmıştır. Ayrıca, sınavın içerik geçerliliğini incelemek için, yapılan görüşmeler detaylı bir şekilde analiz edilmiş ve her görüşmeden elde edilen ortak noktaların üzerinde durulmuştur. Yordama geçerliliğini belirlemek için kullanılan veriyi analiz etmek için ise, Pearson Korelasyon ve basit doğrusal korelasyon analizi yapılmıştır. Sınavın yapı geçerliliğini araştırmak içinse speaking sınav sonuçları ve TOEFL sınavının her alt bileşeninin sonuçları arasındaki korelasyon katsayıları hesaplanmıştır. Sınavın değerlendirici iç tutarlılığı ve değerlendiriciler arası tutarlılık düzeyini tahmin etmek için de, yine korelasyon katsayıları hesaplanmıştır.

Anket sonuçlarının analizi, sınavın yeterli yüzeysel geçerliliğe sahip olduğunu göstermiştir. Ayrica, sınavın içerik geçerliliğini araştırmak için yapılan görüşmeler sonucunda, sınavın yüksek ölçüde içerik geçerliliği vasfına sahip olduğu saptanmıştır. Yordama geçerliliğini belirlemek için yapılan istatistiksel araştırmalar, hazırlık yılı eğitiminde yapılan konuşma sınavının öğrencilerin bölüm konuşma sınavındaki performanslarını belirlemediğini göstermiştir.
Ayrıca, yapı geçerliliğini incelemek için yapılan analizler konuşma sınavı ve diğer alt testler arasında düşük korelasyonlar olduğunu göstermiştir.

Sınavın değerlendiriciler arasındaki tutarlılık düzeyi beklendiği kadar yeterli çıkmamıştır çünkü bir çiftin değerlendiriciler arası tutarlılığı diğerlerine gore düşük bulunmuştur.Diğer yandan, konuşma sınavının yeterli değerlendirici iç tutarlık vasfına sahip olduğu görülmütştür.

Anahtar kelimeler: Geçerlilik, güvenirlilik, sınama, sözel sınav

To my family for their endless support and love…

ACKNOWLEDGEMENTS

I would like to express my special thanks to my thesis advisor, Prof. Dr. Hüsnü Enginarlar for his invaluable guidance, constant encouragement and support at every stage of this thesis study. Thanks also go to the other jury members Assist. Prof. Dr. Nurdan GÜRBÜZ and Dr. Işıl GÜNSELİ KAÇAR for their helpful suggestions and comments.

I offer sincere thanks to the Chairperson of Foreign Languages Department of TOBB University of Economics and Technology, Assist. Prof. Dr. Semih Şahinel, who trusted and provided me the opportunity to conduct this study there.

Thanks are extended to all my ex-colleagues for their cooperation and participation in this study.

I owe special thanks to all my friends who motivated and supported me throughout this study.

Finally, I am sincerely grateful to my family, who have encouraged and believed in me throughout in my education.

TABLE OF CONTENTS

PLAGIARISM……………………………………………………………………………..….iii

ABSTRACT………………………………………………………………………………...iv

ÖZ………………………………………………………………………………………..vi

ACKNOWLEDGEMENTS…………………………………………………………..…ix

TABLE OF CONTENTS…………………………………………………………....x

LIST OF TABLES……………………………………………………………………..xiv

LIST OF FIGURES... xvi

CHAPTER

I. INTRODUCTION…………………………………………………………...1

1.0 Presentation…………………………………………………........1

1.1 Background to the Study……………………………………...1

1.2 Context of the Study………………………………….…...3

1.3 Statement of the Problem…………………………………...5

1.4 Purpose and Scope of the Study……………….........6

1.5 Research Questions……………………………………......8

1.6 Significance of the Study…………………….…………...8

1.7 Definition of Concepts......................................9

II. REVIEW OF LITERATURE...11

2.0 Presentation...11

2.1 Testing Speaking..11

2.2 The Problems of Testing Speaking......................13

2.3 Methods of Testing Speaking............................16

2.4 Concepts Related to Validity and Reliability..........18

2.4.1 Validity...17

2.4.1.1 Types of Validity and Their Uses.18

2.5 Reliability..21
2.5.1 Methods of Determining Reliability.........21
2.5.1.1 External Consistency Procedures. 22
2.5.2.2 Internal Consistency Procedures..22
2.5.2 Reliability of Rater Judgments..............22
2.5.2.1 Inter-rater Reliability................23
2.5.2.2 Intra-rater Reliability................23
2.6 Research Studies on the Validity and Reliability of Exams..23
2.7 Research Studies on the Validity and Reliability of Speaking Tests...26
2.8 Conclusion..33

III. RESEARCH METHODOLOGY..34
3.0 Presentation..34
3.1 The Design of the Study.......................................35
3.2 Participants...35
3.2.1. Students..35
3.2.2. Raters..36
3.2.3. Informants..37
3.3 Data Collection Instruments..................................37
3.3.1. Questionnaire...37
3.3.2. Semi-Structured Interviews..........................38
3.3.3. Video Recordings of C Level Students.............38
3.4.3. Students’ Scores..39
3.4 Data Collection Procedures...................................39
3.4.1. Face Validity..39
3.4.2. Content Validity...41
3.4.3. Predictive Validity......................................41

3.4.4. Construct Validity......................................42
3.4.5. Reliability...42
3.5 Data Analysis..43

IV.DATA ANALYSIS & INTERPRETATION OF THE RESULTS...45
4.0 Presentation..45
4.1 Analysis of Data..45
4.2 Analysis of the Responses to Face Validity Questionnaire..46
4.2.1 Responses to the Multiple Choice and Multiple Response Questions.............................46
4.2.2 Responses to Open Ended Items...............56
4.2.2.1 Results of Open-ended items.........56
4.2.3 Interpretation of the Results of the Face Validity Questionnaire......................................61
4.3 Analysis of the Responses to Content Validity Interview...61
4.4 Predictive Validity Analysis...................................67
4.5 Construct Validity Analysis...................................74
4.6. Reliability Analysis..77
4.6.1 Inter-rater Reliability..................................78
4.6.2 Intra-rater Reliability..................................79

V. CONCLUSION...84
5.0 Presentation...84
5.1 Summary of the Study...84
5.2 Results...85
5.3 Assessment of the Study......................................97
5.4 Recommendations..99

5.4 Implications for Further Research........................100

REFERENCES..102

APPENDICES...108

A. Preliminary Qualifying Exam.........................108
B. Speaking Exam Schedule.............................109
C. The Speaking Exam....................................110
D. A Sample Speaking Exam.............................156
E. The Speaking Exam Rubric.............................158
F. The Face Validity Questionnaire......................159
G. Content Validity Interview Questions..............164

LIST OF TABLES

TABLES

Table 1 The Total Percentages of the Students Answering FVQ Question 3..51

Table 2 The Frequencies of the Responses Given to FVQ Question 3..52

Table 3 The Total Percentages of the Students Answering Question 4..53

Table 4 The Frequencies of the Responses Given to FVQ Question 4..54

Table 5 The Results of Open Ended Items: Item 7.....................56

Table 6 The Results of Open Ended Items: Item 8.....................57

Table 7 The Results of Open Ended Items: Item 9…….................58

Table 8 The Results of Open Ended Items: Item10...................60

Table 9 The Correlation Coefficients between Preparatory Class Speaking Exam Grade and Departmental Speaking Exam Grade.69

Table 10 The Bivariate Regression Analysis of Preparatory Class Speaking Exam and Departmental Speaking Exam Grades..........70

Table 11 The Bivariate Regression Analysis of Preparatory Class Speaking Exam and Departmental Speaking Exam Grades of Students taking the September Proficiency...............................71

Table 12 The Bivariate Regression Analysis of Preparatory Class Speaking Exam and Departmental Speaking Exam Grades of Students taking the December Proficiency................................73

Table 13 The Correlation Coefficients between the Speaking Exam Scores and the Scores of each Subtest and the Total.................75

Table 14 Bivariate Correlation Coefficients between All the Test Components..76

Table 15 The Correlation Coefficients of Each Pair's Ratings..78

Table 16 The Correlation Coefficients of 1st pairs' 1st and 2nd Ratings of the Same Students...79

Table 17 The Correlation Coefficients of 2nd pairs' 1st and 2nd Ratings of the Same Students..80

Table 18. The Correlation Coefficients of 3rd pairs' 1st and 2nd Ratings of the Same Students...81

LIST OF FIGURES

FIGURES

Figure 1 Subjects' Perceptions of the Speaking Exam's Face Validity (FVQ item 11)..47

Figure 2 Subjects opinions on the most accurate way to assess someone's English speaking ability (FVQ question 1)..................48

Figure 3 Subjects' opinions on the extent the speaking exam they took reflected the characteristics of the spoken language in real life situations (FVQ question 2)..50

Figure 4 Subjects' opinions on the difficulty of understanding the test instructions during the test (FVQ question 5)......................54

Figure 5 Subjects' opinion on the attitude of the teachers' attitude during the exam (FVQ question 6)..55

Figure 6 The Scatter Plot of Two Grades..................................68

CHAPTER I

INTRODUCTION

1.0 Presentation

This chapter contains four sections. The first is background to the study. In this section, some information is given about performance assessment in general. Then, the context of the study is presented. Next, the purpose and the scope of the study are highlighted. Finally, the significance of the study is pointed out.

1.1 Background to the Study

With the rise of the communicative approach, the role of speaking ability has become more prominent in language teaching. As a result, performance testing, especially testing the speaking ability has become one of the important issues in language testing. Due to the nature of speaking ability, there are many limitations in this area. The basic problem in testing oral ability is the need to set tasks that form a representative sample of the population of oral tasks eliciting the behavior that truly represents the candidates' ability. In other words, there are many factors that affect our impression of how well someone can speak a language. Since the nature of the speaking skill itself is not usually well-defined, there is disagreement on which different aspects of the speaking skill should be measured. Grammar, vocabulary, and pronunciation are often measured. Moreover, fluency and appropriateness are also usually considered. Because the elements of speaking are plentiful, evaluation of it is not easy either. There may be some inconsistencies in the evaluation process as speaking requires a candidate to use language in some way due to its interactive nature (Luoma, 2004). Moreover, because it involves human raters, the scoring of oral ability is generally highly subjective. Brown (1996) highlights the problem as follows "… the subjective nature of the scoring procedures can lead to evaluator inconsistencies or shifts having an affect on students' scores and affect the scorer reliability adversely" (p.191). Therefore, "the marking system is a vital part of an oral test" (Underhill, 1987, p. 88).

There are also practical restrictions on testing oral performance. These include the administrative costs, difficulties of testing a large number of students either individually or in small groups, training the examiners and the total amount of time and the number of examiners needed for administering the tests. Despite all these constraints, today many institutions are testing students' oral performance through tasks such as interviews, role plays, or oral presentations that are expected to yield evidence about their competence in speaking. Due to this, it should be the responsibility of the institutions to take into consideration the extent to which a test can be shown to produce scores which are representative of a candidate's ability, the constructs wished to be measured and the instruments to be developed in order to provide the necessary information. Given this, "well documented and research-verified explanation" is increasingly required if the validity and reliability of test score interpretation and use are to be supported both logically and with empirical evidence. This should be the primary point that testing should be concerned with as the correctness of interpretations of abilities from test scores can be justified if evidence on underlying abilities or constructs attempted to measure is provided (Weir, 2005).

The notion that language testing is not just about designing instruments for data collection points out the need "to offer a blueprint of the types of evidence to be provided if the correctness of the interpretations of abilities from test scores are to be justified" (Weir, 2005, p. 2) Therefore, a test should be validated by collecting evidence to support the fact that test is doing the job that it is supposed to be doing. This necessarily involves providing data relating to different validities together with the various reliabilities. Yet, as Weir (2005) states, these are not all-or nothing studies and even if only a small section of the validity canvas is filled, it is still an improvement on a test with no validity attached to it (p.220).

1.2 Context of the Study

TOBB University of Economics and Technology, the Department of Foreign Languages was established in 2003 and served nearly 1450 students between 2004 and 2008. The number of English language instructors currently employed at the institution is 51.

TOBB University of Economics and Technology is a Turkish medium university; however, each student enrolling is required to have a certain proficiency level of English to be eligible for the freshman year. Students who are not proficient in English are required to study at the Department of Foreign Languages for one year. There are three main programmes, namely levels, of the preparatory programme at TOBB ETU Department of Foreign

Languages. The students are placed in appropriate levels according to their scores in two different stages (APPENDIX A). The objectives of the levels are determined according to the Common European Framework. Thus, C programme stands for the intermediate level students and the students studying in this programme attend one semester and have the chance to take the second TOEFL-ITP in December provided that they receive a GPA of 65/100 and do not exceed 10 % of the total attendance. Similarly, B programme stands for the pre-intermediate level initially and A program students are at the elementary level.

At the beginning of each academic year, the students take the Preliminary Qualifying Exam, which includes structure, reading and listening sections. The students who cannot pass this exam become A level students. The students who can pass this exam take the writing and speaking exams. If they cannot pass these two components, they will become B level students. The students who pass the writing and speaking sections have the right to take the proficiency exam, which is the TOEFL-ITP. Those who pass the TOEFL-ITP are eligible to start the freshman year. The other students who fail the TOEFL-ITP become C level students and have the right to take the proficiency exam which is given in December (APPENDIX B).

At the Department of Foreign Languages in A and B levels five hours a week is devoted for speaking, but is integrated with listening. In C levels no classroom time is spent on it. However, each C level class has one hour speaking session after the class. Speaking is assessed three times a year. Before the students take the TOEFL-ITP, they are required to take the speaking exam as well as the writing. The speaking exam, which is in September, is taken by both the newly registered students and the ones who fail in the previous academic year. The December speaking exam is taken by only C level students. Lastly, A and B level students are required to take the July speaking exam. Speaking and writing components are important as the results of these tests determine whether the students can attend the TOEFL-ITP and pass the preparatory class.

All the speaking exams given at the preparatory programme include the same tasks. In addition, the assessment scale used for all the tasks is also the same. The exam consists of three parts. In the first part of the exam, the candidates are required to answer general questions about everyday life. The test taker is asked to describe a picture in the second part of the exam. In this part, the candidate is also required to answer interpretative questions related to the picture. In addition, personal questions related to the picture’s main topic are posed. In the last part of the speaking exam the candidate is asked to speak on his own after picking up one topic card from a box in order to express his/her personal opinion on the topic (APPENDIX C). The test takers’ performances in all these tasks are evaluated using the TOEFL-IBT speaking assessment criteria, which is a holistic scale (APPENDIX D). However, the institution has modified the scale to meet their needs.

1.3 Statement of the Problem

Tests are important as the decisions made about students’ performance and knowledge level are influenced by the interpretation of the scores obtained through them. The speaking exam given at TOBB University of Economics and Technology, Department of Foreign Languages also plays an important role in the decisions made about the proficiency level of the students as the students who cannot get an average of 60 out of speaking and writing components are not given the chance to take the TOEFL-ITP exam. Moreover, the students passing the preparatory class also have speaking exams as freshmen and they need to be successful in these in order to pass their courses. The speaking exams implemented in departmental English courses 101, 102 and 201 are similar to the speaking component of TOEFL IBT exam as the students need to get at least a 94 from TOEFL IBT in order to graduate from their university. This clearly indicates that English language education at this university does not end with education in preparatory class.

All these mean that special steps should be provided to ensure the reliability and the validity of the speaking assessment. In other words, in order to claim that this assessment is well founded, some evidence should be generated from test scores.

Weir states (2005) that language testing is not just about creating the instruments for data generation (p.1). This presents the need to show the relationships between the testing instruments and constructs that it attempts to measure as only in this way can more confidence in the interpretation of the results be gained.

1.4 Purpose and Scope of the Study

The purpose of this study is to investigate the reliability and the validity of the speaking assessment implemented at TOBB University of Economics and Technology Department of Foreign Languages. In other words, the aim is to find out if the speaking exam given is doing the job that it is supposed to be doing.

Therefore, data related to content, predictive, construct and face validity, together with reliability indices was collected and analyzed.

The following steps were followed while conducting the study:

i. preparing and administering a questionnaire to examine face validity

ii. administering interviews with the informants on the content of the exam

iii. obtaining both the speaking exam scores of the students in preparatory class and departmental English courses

iv. analyzing the correlations between the scores of two speaking exams

v. obtaining the December TOEFL-ITP scores of C level students and the scores of December speaking exam

vi. correlating each component of the December TOEFL-ITP scores with the speaking exam scores

vii. obtaining the scores of the 6 raters who were pairs in the December speaking exam

viii. having the raters grade the performance of the students once more on videotape

ix. analyzing the correlations between the pairs and within the raters themselves.

If the results of this study show that, the speaking exam given is not adequately valid and reliable, the researcher will make some recommendations for the speaking exam and propose some solutions for TOBB University of Economics and Technology Department of Foreign Languages. If the results indicate that the speaking exam given is valid and reliable, the institution may prefer to continue with more confidence.

1.5 Research Questions

This study sets out to answer the following research questions regarding speaking assessment at TOBB University of Economics and Technology.

1. How valid is the test?

To answer the first question, the following sub-questions need to be investigated:

1.a How satisfactory is the test with respect to face validity?

1.b How satisfactory is the test with respect to content validity?

1.c How satisfactory is the test with respect to predictive validity?

1.d How satisfactory is the test with respect to construct validity?

2. How reliable is the test?

To answer the second question, the following sub-questions need to be investigated:

2.a How satisfactory is inter-rater reliability?

2.b How satisfactory is intra-rater reliability?

1.6 Significance of the Study

This study on the validity and the reliability investigation of the speaking exam implemented at TOBB University of Economics and Technology Department of Foreign Languages is significant for four reasons.

First of all, it is obvious that many educators accept the fact that it is important to test students' competence in speaking through performance based tests in a direct way. However, these attempts can also bring the issues of validity and reliability of the exams implemented. Although many studies have been conducted in different parts of the world on issues concerning the validity and the reliability of the assessments, the number of studies in the field of English Language Testing conducted in Turkey is not many. Therefore, there is a need for similar studies in Turkey in order to obtain more information about testing and validating speaking performance in Turkish context.

Secondly, speaking assessment is implemented three times in one academic year at TOBB University of Economics and Technology Department of Foreign Languages. Since the students are required to take the speaking exam and the results of the exam play an important role in making a decision about students' performances, the speaking assessment must to be evaluated.

Another significance is that this study will shed light on how well this assessment is evaluating the oral performances of the students studying at this department. Therefore, whether there is a need to use a more valid and a more reliable test will be determined only after some evidence is generated about the instrumental value of it. This will also help the examining bodies have more confidence in their interpretation of the scores available to them.

Lastly, this research study will also be valuable for other people in other institutions who would like to validate their tests in order to justify the correctness of their interpretations. They may take this research study as a model and investigate the quality of their own assessment tools.

1.7 Definition of Concepts

Oral Test: is a repeatable procedure in which a learner speaks, and is assessed on the basis of what he/she says. It can be used alone or combined with tests of other skills (Underhill, 1987, p.7).

Testee / candidate: other terms for a test taker.

Interviewer: is a person who talks to a learner in an oral test and controls to a greater extent the direction and the topic of the conversation (Underhill, 1987, p.7).

Interlocutor: is a person who talks with a learner in an oral test, and whose specific aim is to help and encourage the learner to display, to the assessor, his oral fluency in the best way possible (Underhill, 1987, p.7).

Assessor: is a person who listens to a learner in an oral test and makes an evaluative judgment on what he/she hears (also examiner and tester) (Underhill, 1987, p.7).

Marker/ Rater/ Scorer: is the judge or observer who observes a rating scale in the measurement of oral proficiency (Davies et al., 1999, p. 44).

Objective: is the type of scoring where no judgment is required on the part of the scorer (Hughes, 1990 p.22).

Subjective: is the type of scoring where judgment is required on the part of the scorer (Hughes, 1990 p.22).

Validity: deals with whether a test measures what it is supposed to (Underhill, 1987, p.9).

Reliability: is the consistency of evaluation of results (Grounlound & Linn, 1990, p.48).

Validation: is the process of test evaluation to ensure the defensibility and the fairness of test interpretations based on test performance (McNamara, 2000, p. 48).

CHAPTER II

REVIEW OF LITERATURE

2.0 Presentation

In this chapter, testing speaking, the difficulties of testing speaking, concepts related to validity and reliability, speaking test methods and studies on the validity and reliability of the exams will be reviewed.

The first part of the literature review is on testing speaking. First, testing speaking will be discussed. Then, the problems of testing speaking, some concepts related to validity and reliability will be outlined. Moreover, formats of speaking tests will be identified.

In the last part, some studies aimed at investigating the validity and reliability of exams will be reviewed and the results of these studies will be presented as well.

2.1 Testing Speaking

Fulcher (2003) states that the theory and the practice of testing second language speaking is the youngest field of language testing (p.1). That is because it was not until the Second World War that testing speaking became a focus of attention (Fulcher, 1997). Before that testing second language was avoided as the language skills emphasized in language classrooms were the skills of comprehension but not production (Ferguson, 1998). Since the focus in the language classroom started to move from classical approaches in instruction and testing to a more communicative approach, the need to measure language learners' productive skills has arisen. Due to these changes, language teaching has also emphasized the improvement of speaking skills (Hall, 1993). However, assessing speaking is challenging as many factors influence one's impression of how well someone can speak a language (Luoma, 2004). It is also demanding because test scores are expected to be accurate and appropriate for its purpose.

In learning a second or a foreign language, most of the learners find speaking the most difficult skill to master because it requires oral communication that consists of both listening and speaking (Nunan,2002). It is clear that the oral skills are one of the most important to be emphasized. However, many schools or institutions do not even try to measure oral performance. In addition, although it takes its place in their curriculum, not enough attention is paid to it as oral tests are qualitatively different from other tests due to the difficulty of treating oral tests in the same way as other more conventional tests (Underhill, 1987, p.3). Similarly, Lado (1961) states that testing speaking is "the least developed and least practiced in the language testing field" (p.239). Moreover, Chaudlary (1997) highlights the insufficiency of studies on testing speaking. These all indicate that testing speaking is considered the most challenging of all language exams in its phases: preparing, administration and scoring (Madsen, 1983, p.147).

Kitao and Kitao (1996) point out that "in spite of the difficulties inherent in speaking, a speaking test can be a source of beneficial backwash effect since it will encourage the teaching of speaking in class" (p.2). Ur (1996) also supports including oral proficiency tests in language exams:

> In principle, a language test should include all aspects of language skills-including speaking. Speaking is not just "any skill"- it is arguably the most important, and therefore, should take priority in any language test. If you have an oral proficiency test at the end of a course, then this will have a "backwash effect": teachers and students will spend more time on developing skills during the course itself. Conversely, if you do not have such a test they will tend to neglect them. Students who speak well but write badly will be discriminated against if all or most of the test is based on writing (p. 134).

However, Kitao and Kitao (1996) also mention the problems and the difficulties of speaking as sometimes it is necessary to test a large number of students, which makes it essential to develop a system of assessment that can be applied as objectively as possible. Moreover, Grounlound (1998) states some practical limitations of testing speaking which make most language tests not include speaking tests such as the amount of time necessary and the inconsistencies in the judging process of learners' performances.

2.2 The Problems of Testing Speaking

Hughes (1990) explains that too often language tests have a harmful effect on teaching and learning, and fail to measure accurately whatever they are intended to measure (p.1). However, information about people's language skills and ability is sometimes essential. Therefore, when they are tested, the conclusions drawn out of scores should be justified by eliminating the problems which stem from their reliability and validity. The reliability and validity of speaking assessment should also be ensured by using special procedures due to its interactive nature (Luoma, 2004, p. 170). As a result, the literature mainly focuses on the reliability and validity of the exams. Moreover, problems related to speaking test administration, practical constraints, the criteria used to evaluate oral communication and the different nature of speaking from other skills have also been discussed in the literature.

As Underhill states "an oral test is an encounter between two human beings; it is designed by humans, administered by humans, taken by humans and marked by humans" (p.105). This clearly indicates the difficulty of assessing speaking ability with exactness. Therefore, because of its different nature, speaking tests show more questions of validity and reliability than written tests. This calls for the need to ask different kinds of questions in order to evaluate if the test works properly, which is called validation.

Tests of speaking ability show questions of validity and reliability. Validity means whether a test works properly or not. In other words, a test is said to be valid if it measures accurately what is intended to measure (Hughes, 1990, p.76). Having highly valid speaking tests is difficult as "it involves the simultaneous use of a wide variety of different abilities that often develop at different rates" (Harris, 1969, p. 81). Therefore, designing speaking tests is a great problem for test writers as how the content is constructed should be carefully planned by all the people involved in this process. What information should be given by testing instruments and procedures and the purposes of using tests need to be specified (Norris, 2000, p.18).

As stated previously, reliability, which means the stability of scores, is one of the problems of testing speaking. Ur (1996) also highlights that the most significant problem of testing speaking is reliability since there may be variations in examiners' judgment in assessing different examinees. Therefore, such problems as those resulting from inconsistencies between raters, scores, different implementations of the same test and limited guidelines or criteria need to be carefully considered by applying special procedures like evaluating rater reliability, designing effective rating scales and training raters in order to standardize the procedures applied during assessment.

In addition to these, the administration of speaking tests can be challenging due to practical constraints on testing oral communication. These include a necessary number of examiners to test a large number of students, administrative costs, total amount of time needed to implement the speaking exams, equipment and facilities needed for testing and preparation and resources necessary for training the raters (Hughes, 1989; Cohen, 1980; Weir, 1990).

Furthermore, assessing oral ability is problematic due to its being evaluated by human raters and the number of the raters as well. It is claimed by Alderson, Clapham & Wall (1995) that scoring of oral ability is highly subjective and this is one of its characteristics. Heaton (1990a) also expresses the importance of the rater and the difficulty of making objective judgments:

> …..success in communication often depends as much on the listener as on the speaker: a particular listener may have a better ability to decode the foreign speaker's message or may share a common nexus of ideas with him or her, thereby making communication simpler. Two native speakers will not always, therefore, experience the same degree of difficulty in understanding the foreign speaker (Heaton, 1990a; 88).

In relation to this, Brown (1996) also states that "…. the subjective nature of the scoring procedures can lead to evaluator inconsistencies or shifts having an effect on students' scores and affect scorer reliability adversely" (p. 191). In addition, the number of the scorers can also affect the reliability of the scores. Underhill (1987) states "the more assessors you have for any single test…..the more reliable the score will be" (p.89). Furthermore, the roles of interlocutors and raters can cause problems in the assessment of oral performance. The rater acting as an interlocutor at the same time, can be problematic as it becomes harder for an interlocutor to assign scores to test takers while interacting with them as well (Weir, 1995; p.41). These issues also need to be considered while designing and conducting speaking tests.

Establishing the criteria necessary to evaluate oral performance is also one of the drawbacks in assessing speaking due to the different nature of speaking. Deciding on the constructs that should be measured like grammar, vocabulary, pronunciation, fluency and accuracy to evaluate oral communication is still questioned. Kitao and Kitao (1996) mention that "a speaker can produce all the right sounds but not make any sense, or have great difficulty with phonology and grammar and yet be able to get the message across" (p.1). There can be questions on which factors to measure while testing speaking and even the values assigned to each element cause disagreement as well.

2.3 Methods of Testing Speaking

The development of the ability to communicate successfully in the target language is the goal of teaching spoken language and this should involve both comprehension and production (Hughes, 1990). Therefore, it is apparent that testing spoken language is a hard task to accomplish. As the aim is to elicit behavior that represents test takers' ability, setting the right tasks that give valid and reliable information about their performances is significant. There are various methods in assessing oral performance which should be chosen according to the objective of a particular test programme.

The type of the interaction intended may determine the tasks chosen for assessing the speaking ability. Due to this, if the test taker is alone and does not communicate with the testers excluding instructions, test tasks like the following will do: oral presentation/ report, verbal essays, sentence transformation, reading aloud, describing pictures/ maps/ diagrams or re-telling a story. However, sometimes the examinee can be alone but the examiner can be there to communicate with the candidate as well. For this situation, oral interviews and conversational exchanges are suitable as the interaction task requires student-examiner information gap. In addition to this, interaction tasks can include student-student information gap, which requires testing two examinees according to their communication with each other. Role play activities, describing pictures/ maps/ diagrams, paired interviews and giving instructions can be given as examples for this type of interaction. (Hughes, 1990; Weir, 1990; Underhill 1987; Foot 1999).

For all the types mentioned, test takers' performances can be recorded for scoring. Yet, these audio and visual aids may increase the stress level of some candidates during the exam. However, as Perren (1968) states, "spoken language is fugitive. It cannot be re-scanned and reassessed in context like writing unless the performance is recorded as it occurs" (p.108).

It is clearly seen that it would take considerable amount of time and effort to decide on the tasks that would elicit the speaking ability of candidates depending on the needs of institutions or organizations since the suitability of the elicitation techniques and the content is of great of importance if oral performance is desired to be tested in a valid and a reliable way.

2.4 Concepts Related to Validity and Reliability

2.4.1 Validity

Henning (1987, cited in Alderson, et al., 1995, p. 170) defines validity as "appropriateness of a given test or any kind of its component parts as a measure of what is purported to measure". In other words, it means whether a test works properly or not. There are mainly four common types of validity. These are face validity, content validity, criterion validity and construct validity.

2.4.1.1 Types of Validity and Their Uses

To begin with, face validity is concerned with "if the test appears to test what the name of the test implies" (Dick, & Hagerty, 1971, p. 95). Does it seem like a reasonable way to gain the information the researchers are attempting to obtain? Does it seem well constructed? Does it seem as though it will work reliably? Therefore, face validity is determined impressionistically; for example, by asking students whether the exam was appropriate to their expectations and giving questionnaires to administrators or other users.

As the name suggests, content validity is concerned with whether or not the content of the test is sufficiently representative and comprehensive for the test to be a valid measure of what is supposed to measure (Henning, 1987, p. 94). To address this issue, testers or people interested in test validation may need to focus particularly on the organization of the different types of items that they have included on the test and the specifications for each of those item types (Brown, 2005, p.221). Although this validation process may take many forms, the goal is to indicate that the test is a representative sample of the content it claims to measure.

The concept of criterion validity involves "demonstrating validity by showing that the scores on the test being validated correlate highly with some other, well-respected measure of the same construct" (Brown, 2005, p.233). As Weir (1990) states "this is a predominantly quantitative and a posteriori concept" and is divided into two types: concurrent and predictive validity (p. 27). Concurrent validity compares a new instrument with those more established, that supposedly measure the same things. It is established when the test and the criterion are administered at about the same time (Hughes, 1990, p.27). When concurrent validity is investigated, one needs to administer a reputable test of the same ability to the same test takers concurrently or within a few days of the administration of the test to be validated. Then, the scores of the two different tests are correlated using some formula for the correlation coefficient and the resultant correlation is reported as a concurrent validation.

Predictive validity, which is also estimated in this study, differs from concurrent validity in that instead of collecting the external measures at the same time as the administration of the external test, the external measures will only be gathered some time after the test has been given (Alderson, et al., 1995, p. 180). Predictive validation is generally done for proficiency tests. One simple way of validation of this type is to give students a test, and then later on at some point in the future give them another test of the ability the initial test is intended to measure (Alderson, et al., 1995).

The next type of validity is the construct validity and this is the most difficult validity type to explain as it is regarded as a superordinate form of validity to which external and internal validity contribute (Alderson, et al., 1995), p.183). Similarly, Anastasi (as cited in Weir, 1990) expresses that "the content, criterion related and construct validation do not correspond to distinct or logically coordinate categories, on the contrary, construct validity is a comprehensive concept which includes other types" (p.153).

Ebel and Frisbie (1991) explain it as follows:

> The term *construct* refers to a psychological construct, a theoretical conceptualization about an aspect of human behaviour that cannot be measured or observed directly. Examples of the construct are intelligence, achievement, motivation, anxiety, attitude, dominance and reading comprehension. *Construct validation* is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect to mean them. (p.108)

To provide evidence for construct validity, it is essential to indicate that the test correlates highly with indices of behaviour that it is expected to correlate with and also that it does not correlate significantly with variables that it is not expected to correlate with (Weir, 1990, p.23).

2.5. Reliability

The concept of reliability is defined as "the consistency of measurement" (Bachman and Palmer, 1996, p. 19). In other words, a test is reliable to the extent that whatever it measures, it measures it consistently. A measure is considered reliable if a person's score on the same test given twice is similar. It is significant to remember that reliability is not measured; it is estimated.

In addition to these, a test cannot be valid unless it is reliable. If a test does not measure something consistently, it cannot always be measuring precisely. Yet, it is also possible for a test to be reliable but not valid. For example, a test can give the same results all the time even though it is not measuring what it is claimed to. Hence, even though reliability is a must for validity, it alone is not adequate (Alderson, et al., 1995).

The reliability of a test is quantified in the form of a reliability coefficient. The reliability coefficient allows one to compare the reliability of different tests. The ideal reliability coefficient is 1.00, which means that the test would give the same results for a particular set of test takers regardless of the time of administration. This indicates that the reliability coefficient is used to estimate the reliability of a test. However, there are many ways through which reliability coefficients are arrived at.

2.5.1 Methods of Determining Reliability

There are several ways to find out the reliability of an instrument. The diverse procedures can be classified into two groups (Kumar, 1996, p.141):

1. external consistency procedures,
2. internal consistency procedures,

2.5.1.1. External Consistency Procedures

External consistency procedures compare cumulative test results with each other as a means of verifying the reliability of the measure. Test-retest and equivalent-forms are the basic external consistency strategies devised to estimate the stability of a test over time.

2.5.1.2 Internal Consistency Procedures

Internal consistency is the extent to which tests or procedures assess the same characteristic, skill or quality. These procedures aim to establish the items measuring the same phenomenon have similar results. Testers generally prefer internal consistency strategies to estimate the internal consistency reliability in order to avoid the effort and difficulty involved in the external consistency strategies (Brown, 2005, p.176).

2.5.2. Reliability of Rater Judgments

Whenever humans are used as a part of the measurement system, the reliability and the consistency of the results should be explored. When testing students' productive skills (speaking and writing), raters are essential. In such situations whether two raters are being consistent in their judgments should be determined by relying on inter-rater and intra-rater reliability (Brown, 2005, p. 185).

2.5.2.1 Inter-rater reliability

This type of reliability is estimated by examining the scores of two raters and calculating the correlation coefficient between the two sets of scores. The correlation between these ratings would give you an estimate of the reliability or consistency between the raters.

2.5.2.2 Intra-rater reliability

Intra-rater reliability is estimated by gathering two sets of scores produced by the same rater for the same group of students. Then, the correlation coefficient between those two sets of scores is calculated. The reliability coefficients provide estimates of the consistency of judgments over time.

2.6 Research Studies on the Validity and Reliability of Exams

Many studies have been done to evaluate the reliability and validity of the exams. The subject matter of these studies has mainly been on the validity or the reliability of tests other than speaking due to the difficulties mentioned in testing speaking. However, investigating those studies may enable the researcher to gain insight into the procedures applied during the validation studies as well.

To begin with, some researchers have examined the face validity and content validity of tests. To illustrate, the content validity of the end-of-course assessment administered at Hacettepe University, Department of Basic English in the 1997-1998 academic year was investigated by Ösken (1999). To assess the validity of the exam a questionnaire was administered to the instructors.

Their perceptions on whether the end-of-course assessment reflected the contents of the course books were examined. Moreover, to determine the content validity the researcher also compared the number of test items with the frequencies of the course objectives. This way the researcher also wanted to determine the consistency between the end of course assessment test and the course objectives. The questionnaire given indicated that the end-of-course assessment reflected the course content from the point of view of the instructors. Nevertheless, the content validity analysis showed that the items in the test were not chosen by considering the frequencies of course objectives.

Nakamura (2006) investigated the face validity and content validity of a pilot English placement test as well as its reliability and practicality. To examine the face validity of the exam, informal questionnaires and discussions were conducted with 809 first year university students. Content validity was examined in a non-statistical way, by discussing the test items with the instructors. The discussions were mainly on the test construct and the testing method. The results of the study indicated that the test had enough validity. Yet, the researcher highlighted the importance of conducting studies on predictive validity, concurrent validity and more systematic studies on face validity and practicality to improve the test.

Another researcher, Serpil (2000) examined the content validity of tests as well. The study was conducted at Anadolu University School of Foreign Languages and the researcher focused on the content validity of midterm achievement tests administered at the institution. Similar to Ösken's study, questionnaires were distributed to the instructors to investigate their perceptions of the test content and teaching objectives. In other words, their ideas on how well the test items represented the intermediate course material content were explored. Moreover, the instructors were interviewed to discover the teaching objectives. After these procedures were completed, the content of the test and the course material were compared. Then, the content of the tests and teaching objectives were compared. As a result of these, it was found that the instructors' perceptions of midterm tests' representativeness of the course content was moderate to high. Yet, this result conflicted with the degree of the tests' representativeness of the course material which was low. The comparison between the content of the tests and the teaching objectives also resulted in low correlation. It was speculated by the researcher that the study may have resulted in this way due to the insufficiency of definitely determined testing criteria and course objectives.

In addition to these studies, predictive validity of the exams was investigated by some researchers. For example, Dooey (1999) examined the predictive validity of the IELTS (International English Language Testing System) test as an indicator of future academic success. The subjects included both foreign and native students entering first year of a graduate course on the basis of their IELTS scores. In the analyses IELTS test scores and semester weighted averages (SWAs) were used. As a result of this study, it was found that 15 out of 23 native speakers who did not have any difficulty with English became unsuccessful academically and foreign students who couldn't meet the admission criteria in terms of their English level were still successful academically. Therefore, Dooey claimed that future academic success was not guaranteed by high IELTS scores.

The reliability of the exams was studied as the precision of the interpretations needed justifying as well. Some researchers examined the scorer reliability of the tests. Manola and Wolfe (2000) examined the reliability of the essay writing section of the TOEFL. They aimed to investigate to what degree raters' judgments were affected by computer based and hand writing essay mediums for the Test of English as a Foreign Language (TOEFL). 152,951 TOEFL examinees participating in regular TOEFL administrations were involved in this study and their papers were scored by two independent groups of trained judges. It was found out that agreeing on word processed essays was easier than the handwritten ones according to the raters. The researchers concluded that, it was not just to make conclusions about the examinees performances by using the scores gathered from them as the inferences drawn from hand written essays were suggested to have decreased validity.

Cardoso (1998) focused on the reliability of English tests administered in Brazil as part of the university exam. Two reading tests were under analysis and internal consistencies of the two test scores were statistically analyzed. The result of the study indicated that both tests were reliable with the reliability coefficients of 0.912 and 0.83.

2.7 Research Studies on the Validity and Reliability of Speaking Tests

Several studies in language testing have already been conducted in an attempt to analyze the different aspects of speaking tests. Shohamy (1994, cited in Iwashita, Brown, McNamara and O'Hagan, 2007, p.7) states that insights from such analysis provide invaluable contribution to defining the construct of speaking in oral tests. However, only few of these studies focused on the issues of validity and reliability due to the nature of speaking ability and complexity of measuring spoken utterances. Examining how other researchers have investigated speaking tests can shed light on the process followed in this research study.

Some researchers have looked at the validity of speaking tests. To begin with, Nakamura (1997) investigated the construct validity of an English Speaking Test. One of the purposes of this study was to examine if the proposed nine traits (pronunciation, grammar, discourse, fluency, content, vocabulary, comprehensibility, interactional competence and sociolinguistic competence) were relevant and separable parts of speaking ability. Moreover, the research was also designed to discover the extent to which the proposed construct of speaking is reflected in other standardized tests such as the Test of Spoken English (TSE) or the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) (Nakamura, 1997, p. 14). Twenty nine college students took the set of tests (a writing test, an interview test and a tape mediated speaking test) and seven English teachers, who are native speakers of English scored the test using the 1-4 point scale rating sheet. Tape mediated tests and interview tests were given since another purpose of this study was to examine whether there was a relation between methods and language ability. Factor analysis was adopted to examine the categorization of traits and the relationship between the factor and method. As a result of this study, Nakamura concluded that an interview test and a tape test are measuring the whole communication ability from different modes. He also added that this might be true for TSE and ACTFL because of the similar components of these tests. In addition to these, the study showed that the nine traits were functioning as factor construct elements in both direct oral communication ability (Interview test) and semi direct oral communication ability (tape mediated test). However, they all maintain their own characteristics which cannot be measured by others.

O'Sullivan, Weir and Saville (2002) also examined the validity of speaking test tasks within the University of Cambridge Local Examinations Syndicate (UCLES) by addressing the match between the intended and the actual test taker with respect to a blueprint of language functions representing the construct of spoken language ability. In order to achieve this, an observation checklist was designed for both a priori and posteriori analysis of speaking task output. With the help of this checklist, the output elicited by the task was examined without resorting to limited analysis of transcripts. Thus, this study provided additional source of validation evidence. The checklist was prepared by referring to relevant literature, namely, by considering the informational and interactional functions of a speaking test. Then, it was adapted to more closely mirror the desired outcomes of spoken language test tasks in the UCLES Main Suite by evaluating draft checklists in order to arrive at an operational one. The results of this study indicated that operational version of these checklists was feasible although further modification was required. O'Sullivan, Weir and Saville also concluded that the checklists on their own are not satisfactory enough to offer evidence to show the construct validity of a spoken test but could be supplements to other procedures.

Another study by Cumming, Grant, Mulcahy-Ernt and Powers (2004) also focused on speaking test tasks as well as writing. As to speaking, the content validity, educational appropriateness and the perceived authenticity of prototype tasks for a new TOEFL were evaluated. Seven highly experienced instructors of English as a Second Language (ESL) at three universities were asked to rate their students' abilities and to check their students' performances in order to determine if this new version of the TOEFL test corresponded to the domain of academic English necessary for studies at English-medium universities or colleges in North America. Moreover, they aimed to investigate if the test tasks fulfilled the purposes that they were designed for. As the authors suggest the rationale for this study followed from recent conceptualizations of the centrality of construct validity in test development since it requires various kinds of evidence from different sources about the interpretation of test scores and conclusions drawn out of the test results. The instructors taking part in this study completed questionnaires profiling their professional qualifications, rated their students' proficiency in the field test of prototype tasks for the new TOEFL. It was observed that performances of most of the instructors' students on the prototype test tasks were equivalent or better than their usual performance in class and they had positive views. However, they realized some problems and suggested ways as to their content and presentation.

Iwashita, Brown, McNamara and O'Hagan (2007) investigated the nature of speaking proficiency in English as a second language in the context of a larger project to develop a rating scale for a new international test of English for Academic Purposes, TOEFL IBT. As mentioned in the article, the study can also be thought to address issues of test score validation since it provides evidence in interpreting the conclusions about learners' abilities made on the basis of scores given by examiners using rating scales. Spoken test performances elicited through five different tasks and five different proficiency levels were examined using a range of measures of grammatical accuracy, complexity, vocabulary, pronunciation and fluency. The results of the study indicated that a set of features had the strongest impact on the overall assigned score. These were *Vocabulary* and *Fluency*. The study also showed that even if one feature of language is not as good as other features, the level of the overall proficiency of that test taker is not merely determined by their performance on that particular aspect of language. As to the five different levels, it was observed that level 5 and level 4 learners had clearly better performances, but the performance of level 1 learners was not always the worst, that is, for some features the performances of the speakers were not as expected. The results presented are thought to have strong implications for scale development as in general the contribution of different features of performance to overall assigned score is an issue for the interpretability of scores.

The issues of validity and reliability were also the focus of Salabary's (2000) study as minor structural changes ACTFL (American Council on the Teaching of Foreign Languages)- OPI (Oral Proficiency Interview) were explained addressing the concerns related to the reliability and the validity of the instrument. Three problems of OPI were mentioned, i.e., lack of features of conversational interaction, limited range of interactional contexts and lack of specification of content areas to be addressed. Therefore, Salabary (2000) suggested broadening the scope of interactional formats represented in oral performance tests by using simulations such as role plays as they are more authentic. In addition to this, it was highlighted that the assignment of differential weights to each category of the assessment criteria of the OPI is the most significant liability of it. To remedy this situation, designing a more specific set of criteria is recommended. As a result, the principled selection of tasks (i.e., conversational interaction samples) to be included in an oral proficiency interview, the selection and identification of what criterion or norm will be pursued, some specification of the developmental process of L2 learning, the explicit identification and description of the components of communicative language ability, and the explicit assignment of weights to each category of overall competence should provide the points of departure for the modifications of any future revision of the ACTFL-TTM.

In addition to these, Sawaki's (2007) study on the construct validation of analytic rating scales in a speaking assessment addressed score dependability, convergent validity of analytic rating scales and the relationship of the analytic scores to the overall score. The issue of validation has been mostly researched in the context of L2 performance assessment based on analytic rating scales and in this research, the responses of 214 participants to two role play speaking tasks in the Speaking section of the Language Ability Assessment System (LAAS) Spanish test were analyzed using the Confirmatory factor analysis (CFA) and multivariate generalizability theory (G theory). The raters of the speaking tasks were 15 graduate students and faculty members at the Department of TESL/ Applied Linguistics and the Department of Spanish and Portuguese at UCLA. The results of this study showed that there are extremely high correlations among the LAAS analytic rating scales. For example, the correlations among Vocabulary, Cohesion and Grammar were very high, which meant that a test taker scoring high on one of these constructs tended to score high on the other two as well. Therefore, it was concluded that these empirical interrelationships show that the scales are related to one another (convergent validity) and each scale provides information about a unique aspect of the test taker's language ability (discriminant validity).

The performances of test takers during oral exams were also analyzed in He's and Dai's study (2006). A corpus based study was conducted by analyzing the conversations that took place in College English Test–Spoken English Test (CET-SET) group discussion to examine its validity. During the analysis, the degree of interaction among test takers in the conversations was observed. The degree of interaction was analyzed by means of a checklist of eight interactional language functions (IFL) included in CET-SET syllabus. Quantitative analysis showed that the frequency of the occurrence of IFLs was very low. This inadequate elicitation of IFLs was thought to pose a problem for measuring their speaking ability in terms of the ability to engage in communicative interaction. The study suggests that conversational features do not appear in speaking tests just because speaking partners are introduced with equal power. Therefore, the research team would like to determine if grouping has any influence on candidates' performance. Since the grouping was done by a computer, this issue could not be investigated. In addition to this, the topics in the exam were examined. Due to this, the research team would like to take a closer look at the data to see if there are any task specific trends, hoping that the findings will give the test designers some idea of the topics that are engaging and able to elicit more ILFs included in the test syllabus.

2.8 Conclusion

In this chapter the researcher has reviewed issues related to testing speaking. Moreover, research studies on issues of validity and reliability of assessments have been highlighted. It has been observed by the researcher that research studies have mainly been conducted to evaluate the validity and reliability of tests of receptive skills due to the limitations mentioned. Moreover, it has been realized that different aspects of speaking ability were analyzed more than examining the reliability and the validity of oral proficiency exams. The study described aims to conduct such an analysis on an oral proficiency exam.

CHAPTER III

RESEARCH METHODOLOGY

3.0 Presentation

In this chapter, first the design of the study is explained. Then, the participants of the study, the students and the raters are presented. After that, data collection procedures and data collection instruments are explained. Finally, information on the data analysis and interpretation is provided.

3.1 The Design of the Study

This study was designed to examine the validity and the reliability of the speaking exam implemented in at TOBB University of Economics and Technology Department of Foreign Languages. In order to be able to investigate the validity and the reliability of the speaking exam aforementioned, different kinds of instruments were used to collect data. Therefore, this is both a quantitative and qualitative research study.

In the quantitative part of the study, a face validity questionnaire and the scores of C level students were used in order to answer the first, the third and the fourth parts of the first research question. The scores were used in order to examine the second research question as well. In the qualitative part of this study, interviews were held in order to examine the second part of the first research question, which is on content validity.

This study addresses the following research questions:

1. How valid is the test?

To answer the first question, the following sub-questions need to be investigated.

1.a How satisfactory is the test with respect to face validity?

1.b How satisfactory is the test with respect to content validity?

1.c How satisfactory is the test with respect to predictive validity?

1.d How satisfactory is the test with respect to construct validity?

2. How reliable is the test?

To answer the second question, the following sub-questions need to be investigated.

2.a How satisfactory is inter-rater reliability?

2.b How satisfactory is intra-rater reliability?

3.2. Participants

3.2.1. Students

The students participating in this present study (*N*=70) were members of C level preparatory class students who took the December TOEFL-ITP. This group involved students who registered in 2007-2008 academic year. The group was chosen mainly as C level students have the chance to take the December TOEFL-ITP and enter the freshmen year as irregular students if they pass it. Since the researcher aimed to find out the predictive validity of the speaking exam given in the preparatory year, the speaking exams that the students had in their departmental English courses were needed in order to investigate it.

3.2.2. Raters

In order to answer the second research question six raters participated in the study. They were English instructors employed at TOBB University of Economics and Technology, Department of Foreign Languages. The participants were selected for the study on the basis of their willingness to participate. The researcher explained the process of this research study to the instructors who rated the December speaking exam and six of them volunteered to participate in the study.

All of the participants are female and non-native speakers of English. The participants' ages ranged from 25 to 40 years. Their experience in teaching English ranged from three to fifteen years. Among the six instructors, three of them were teaching speaking during the 2007-2008 fall and spring semesters. One of these instructors is also the speaking coordinator of the department. Moreover, they all had experience in assessing the speaking ability as this exam is given three times in each academic year.

The raters who participated in this research were also pairs in the assessment of the December speaking exam. The volunteering instructors were chosen among pairs in order to calculate the inter-rater reliability of the exam as each student who took the exam was rated by two separate raters, but received only a single score.

To estimate the inter-rater reliability of the exam, Pearson correlation coefficients of the 1st ratings of each pair were computed. In other words, the scores assigned by each rater was correlated with their partners to determine how consistent their ratings are.

The intra-rater reliability of the exam was computed by conducting Pearson correlation using the 1st and 2nd ratings of each pair for the same students. In order to gather data for this part of the study, these six pairs who volunteered to take part in this study graded the performances of the students once more. This time the ratings were done by watching and listening to the audio and video recordings made during the speaking exam.

3.2.3. Informants

In the qualitative part of this exploratory research in order to answer the second part of the first research question, the content validity of the speaking exam was investigated. Since this depends on the logical reasoning of the informants and the researcher, three people took part in this part of the study. One of them was the chairperson and the other two were the academic and administrative coordinators of the Department of Foreign Languages. Two of the informants were male. The informants' ages ranged from 30 to 45. Their years of experience ranged from 10 to 25.

3.3. Data Collection Instruments

The instruments employed in this study were a questionnaire, interviews, video recordings and scores of students.

3.3.1 Questionnaire

As pointed out by Dörnyei (2003) "questionnaires are uniquely capable of gathering large amount of information quickly in a form that is readily processable" (p.1). In this study the researcher aimed to collect quantitative data to answer the first part of the first research question on the face validity of the exam implemented.

A face validity questionnaire, used in another study was found in the literature and adapted for this study (Moritoshi, 2002) (APPENDIX E). As the construct of the questionnaire was students' ideas and attitudes about the speaking exam and its characteristics, the questionnaire aimed to find data on whether the test given is appropriate to their expectations (see 3.4.1).

3.3.2. Semi-structured Interviews

To answer the research question on content validity, the researcher developed ten open ended questions and conducted three semi-structured interviews with the informants. The answers given to the questions asked by the researcher were noted down in order to be analyzed later on. The interviews held in order to collect data on content validity are explained in a more detailed way in the 3.4.4 part of the chapter.

3.3.3. Video Recordings of C Level Students

Another instrument used in this study is video recordings of C level students taking the speaking exam administered in December 2007-2008 fall term.

After receiving permission from TOBB University of Economics and Technology Department of Foreign Languages administration, the researcher made use of the video recordings of the students who took the December speaking exam. The recorded performances of the students taking the speaking exam were rated once more by the instructors volunteering in order to examine the intra-rater reliability of the exam.

3.3.4. Students' Scores

For the quantitative part of the study, 2007-2008 academic year speaking assessment scores of C level students' scores were examined. The December TOEFL-ITP scores and speaking exam grades of the students were used. Moreover, the departmental speaking exam scores of the students passing the December TOEFL-ITP were used to investigate the predictive validity. Since not all the students passing the December TOEFL-ITP were able to take departmental English courses 101 and 102 for different reasons, the speaking scores of the students passing the September TOEFL-ITP were used in order to calculate the predictive validity as well. Briefly, speaking scores of 42 students passing the September TOEFL-ITP exam and speaking scores of 18 students passing the December TOEFL-ITP were made use of as the scores of the departmental speaking exam of these students (totally 60) were obtained. All of these students registered in 2007.

3.4. Data Collection Procedures

In this section, how various aspects of the validity and the reliability of the speaking exam were analyzed is outlined.

3.4.1. Face Validity

The first set of data was collected through an 11-item face validity questionnaire (FVQ) (APPENDIX E) given *N*=70 of the C level students who took the December TOEFL-ITP. The questionnaire was adapted and administered to ascertain the subjects' views on how speaking should be tested generally and their reactions to certain aspects of the speaking exam given in particular. Although only item 11 pertains particularly to the subjects' understanding of the test's face validity, the other items were included in order to provide additional quantitative and qualitative information which might justify the responses of the subjects for item 11.

The FQV administered was written both in English and Turkish to maximize the comprehension and depth of subjects' responses. After the questionnaire was adapted, it was shown to the supervisor of this thesis, to an English instructor and the chairperson of TOBB University of Economics and Technology, Department of Foreign Languages. They gave suggestions on the wording, format and the length of the statements in the questionnaire. To illustrate, item 5 was "Was it difficult to remember the test instructions during the test?" and was changed into "To what extent was it difficult to understand the test instructions during the test?" since the options to be chosen were worded like "very difficult, difficult, neither easy nor difficult, easy and very easy" in order not to answer the question with yes/no statements again and again. Furthermore, these three experts suggested clarifying what "looking like real life situation" means. Therefore, item 2 was changed into "To what extent did the speaking exam you took reflect the characteristics of the spoken language in real life situations?"

After the necessary changes were made on the construction of some of the statements related to their clarity, it was piloted on a group of students (*N*= 15) who took the speaking exam as well. The subjects in the piloting group were required to mark the ambiguous and unclear statements. Using the piloting data, the questionnaire items were revised, some statements were reworded or changed in order to eliminate the uncertainties. After the revisions were made, according to the information gathered from the piloting group, the questionnaire was administered to the sample group.

3.4.2. Content Validity

In order to examine the content validity of the exam, the necessary data was collected through an interview held with a group of expert judges, namely the chairperson of the department and two coordinators. They were asked questions on the content of the exam and the syllabus in order to determine if the exam includes an adequate number of items that tap the construct (APPENDIX F). The interviews were analyzed by the researcher to examine the content validity of the instrument by focusing on the commonly given answers.

3.4.3. Predictive Validity

In order to examine the criterion-related validity of the speaking exam implemented, its predictive validity was investigated by calculating the Pearson Product Moment correlation coefficient and conducting regression analysis between the scores of the speaking exam and the scores of the speaking exam given in departmental English courses. Since both are supposed to measure the speaking ability of the students, the researcher wanted to determine if the speaking assessment examined predicts scores on some criterion measure. First, the researcher gathered the scores of the C Level students taking the December speaking exam. Then, the names of the students passing the speaking exam were taken from the administration. The students starting the freshmen year as irregulars and getting the departmental English courses at the same time were identified since not all the students took the departmental English courses due to being exempted from them or clashes with other courses in their weekly programmes. Having identified those, the researcher entered the data for each student in SPSS programme and calculated the correlation coefficient between the two scores. Although the results could have been interpreted after calculating the correlation coefficient, the researcher wanted to be more sure of the results obtained. Therefore, a simple linear regression analysis was also conducted to verify the obtained results.

3.4.4. Construct Validity

Data for construct validity was obtained from the scores of the 2007 registered C Level students who took the December TOEFL-ITP. The overall construct validity of the exam is discussed by correlating the scores obtained for each component of the TOEFL-ITP with the speaking scores. This is one way of assessing the construct validity as the correlations between the different components of the test are expected to be fairly low. Alderson, et al. (1995) states that “ the reason for having different components is that they all measure something different and therefore contribute to the overall picture of language ability by the test” (p. 184).

The relationship between the speaking scores and total TOEFL scores was also analyzed by conducting Pearson Moment correlation.

3.4.5. Reliability

Two different statistical procedures are commonly used to produce estimates of reliability. These are correlation and Kuder-

Richardson internal consistency formulae. Each of the three common procedures - test/re-test, parallel form and split half – gives information about the reliability of a test. However, as Underhill suggests these classical measures of test reliability have little relevance for oral tests because they are designed for rigid, pre-planned tests consisting of a fixed number of individual questions (1987,p.106). More useful information could be gathered by comparing each marker's scores with her/his own scores and with the scores of other markers. Based on this, the inter-rater and intra-rater reliability of the speaking exam was estimated. In order to calculate the inter-rater reliability of the speaking exam given, the grades of two raters were correlated. Furthermore, to calculate the intra-rater reliability, three pairs were asked to grade the same students' performances once more after the exam was administered. The video recordings of the students that they graded were given to the raters and they were requested to grade them by watching and listening to the recordings using the same criteria. After that, the second grades that they gave were collected to be analyzed.

3.5 Data Analysis

The data analysis was performed in five steps. Firstly, frequencies for each of the items of the face validity questionnaire were calculated. Then, the answers given to the open ended questions in the questionnaire were analyzed by the researcher in order to find out how satisfactory the face validity of the exam is.

Secondly, the interviews held to examine the content validity of the exam were analyzed by focusing on the commonly given answers.

Next, the predictive validity of the exam was examined by correlating the preparatory class speaking exam scores of the C level students' with their scores in the speaking exam given in departmental English courses. It was computed by means of Pearson Product Moment Correlation Coefficient.

For the fourth step, in order to examine the construct validity, the scores obtained for each component of the December TOEFL-ITP were correlated with the speaking exam scores.

Then, to investigate the reliability of the speaking exam, the grades given by each rater were correlated within themselves and with their partners as well.

CHAPTER IV

DATA ANALYSES AND INTERPRETATION OF RESULTS

4. 0. Presentation

This chapter presents data analysis and interpretation of results. First, the face validity questionnaire results are presented and discussed. Then, the second set of data collected in order to investigate the content validity of the exam through interviews is presented and examined. Next, predictive validity analysis is presented. After this, construct validity analysis of the exam is displayed and finally, the inter-rater and intra-rater reliability analysis of the exam is presented. Furthermore, the results of all the research questions are interpreted and discussed.

4.1 Analysis of the Data

This study aims to investigate the validity and the reliability of the speaking exam given in preparatory classes at TOBB University of Economics and Technology Department of Foreign Languages.

Therefore, in this study, different sets of data were collected and used to investigate different types of validity and reliability.

To examine the face validity of the exam implemented, a questionnaire was used to find out students' ideas and attitudes about the exam. The content validity of the exam was investigated by conducting interviews with the informants. To investigate the construct validity of the exam, the December TOEFL-ITP scores and the speaking and writing exam scores of the 2007 registered C level students were used. Likewise, the scores of the speaking exam under research and the scores of speaking exam given in Departmental English 101 and 102 courses were used to examine the predictive validity of the exam. Additionally, the speaking exam scores given by the raters were used to estimate the inter-rater and intra-rater reliability of the exam.

4.2 Analysis of the Responses to the Face Validity Questionnaire

The face validity questionnaire was presented to the students after they took the speaking exam. The questionnaire contained 11 questions. Seven of the questions were multiple choice type and 4 of them were open ended. Among the multiple choice questions, there were two questions to which the students could give more than one answer. Thus, the results obtained for each type were analyzed and presented independently.

Frequencies and percentage analyses for the multiple choice item questions and multiple response questions were analyzed using the SPSS programme.

4.2.1 Responses to the Multiple Choice and Multiple Response Questions

Since only item 11 pertains particularly to subjects' understanding of the test's face validity, the frequency and percentages of students' responses for that item was given for the ease of analysis and interpretation in Figure 1. The other items which were included in order to provide additional quantitative and qualitative information which might justify the responses of the subjects for item 11 were presented later on.

Of the 70 subjects who took the test, all of them answered item 11 directly relating to their perception of the speaking exam's face validity. The results for that item are presented in Figure 1.

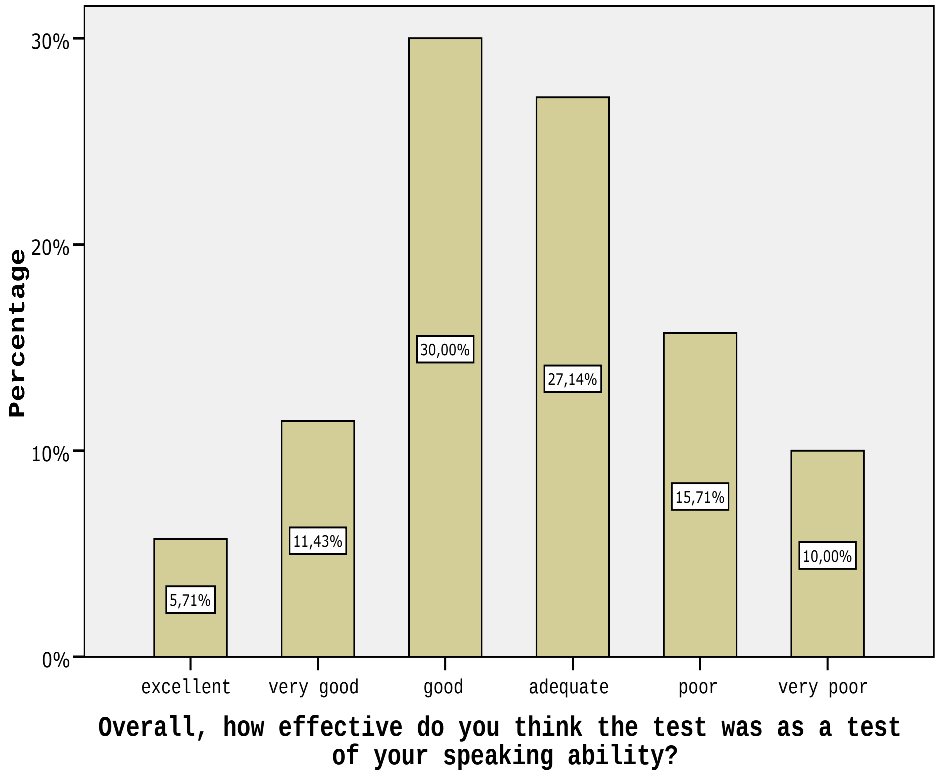


Figure 1. Subjects' Perceptions of the Speaking Exam's Face Validity (FVQ item 11)

These data cannot be averaged to find a mean value due to being at an ordinal level. When the figure is examined, it is seen that nearly 47% of the sample think that the speaking exam has a good-excellent, i.e, satisfactory face validity, while 27% think that its face validity is adequate. Moreover, nearly 16% of the sample regards its face validity as being poor and 10% percent of the subjects find it very poor as being a test of their speaking ability. To assist in the interpretation of the overall effectiveness of the exam, further analysis was performed on the other FVQ items, the results of which are indicated in the following figures and tables.

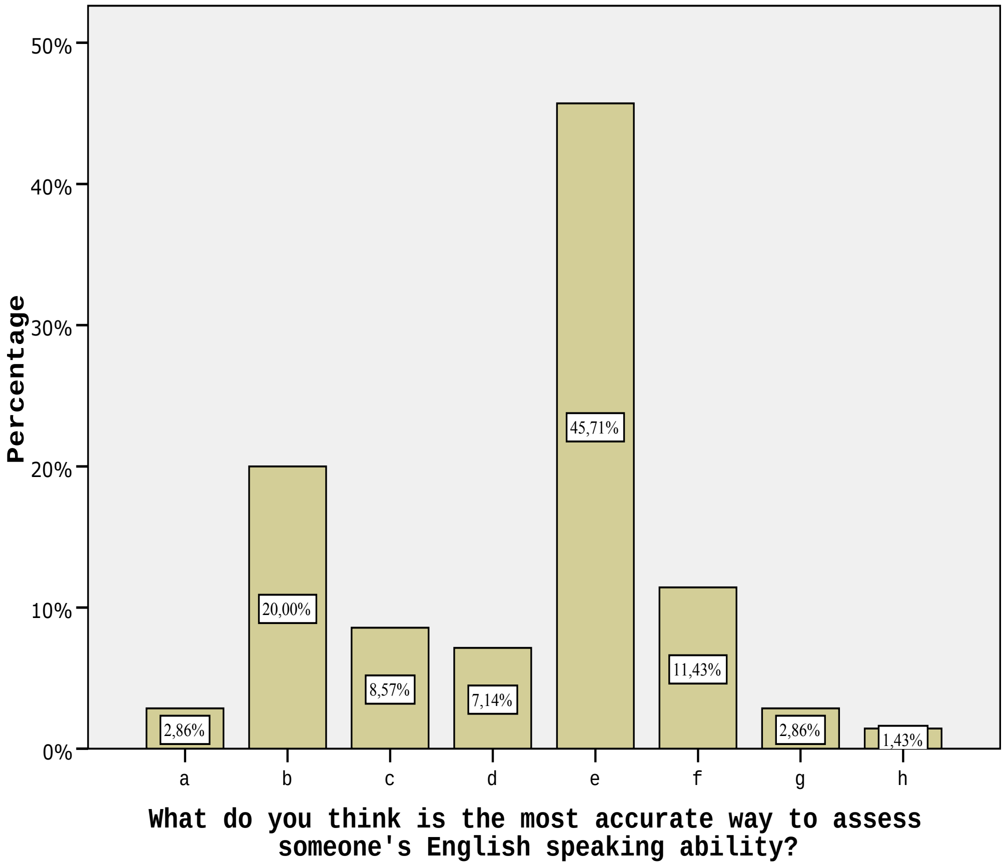


Figure 2. Subjects' opinions on the most accurate way to assess someone's English speaking ability (FVQ question 1)

The first FVQ question includes eight items regarding the most accurate way to assess someone's English speaking ability:

(a) write a script of a dialogue or talk, (b) read a dialogue or talk, and then answer comprehension questions about it, (c) listen to a dialogue or talk, and then answer comprehension questions, (d) a written test of vocabulary and grammar useful during speaking, (e) speak with a native speaker on a given topic in English, (f) speak with a non-native speaker in English on a given topic in English, (g) another way and (h) I am not sure.

As indicated in Figure 2 above, nearly 46 % of the subjects involved in the study reported that "speaking with a native speaker on a given topic in English" is the most accurate way to assess someone's speaking ability. According to the responses of the subjects it is seen that "reading a dialogue or a talk and then answering comprehension questions about it" is another mostly preferred and accurate way of measuring someone's speaking ability with the percentage of 20. Additionally, nearly 3% of the students suggested other ways to assess someone's speaking ability like having a small chat, practicing throughout the whole semester with a native or non-native speaker and assessment of the class teacher during the classroom activities.

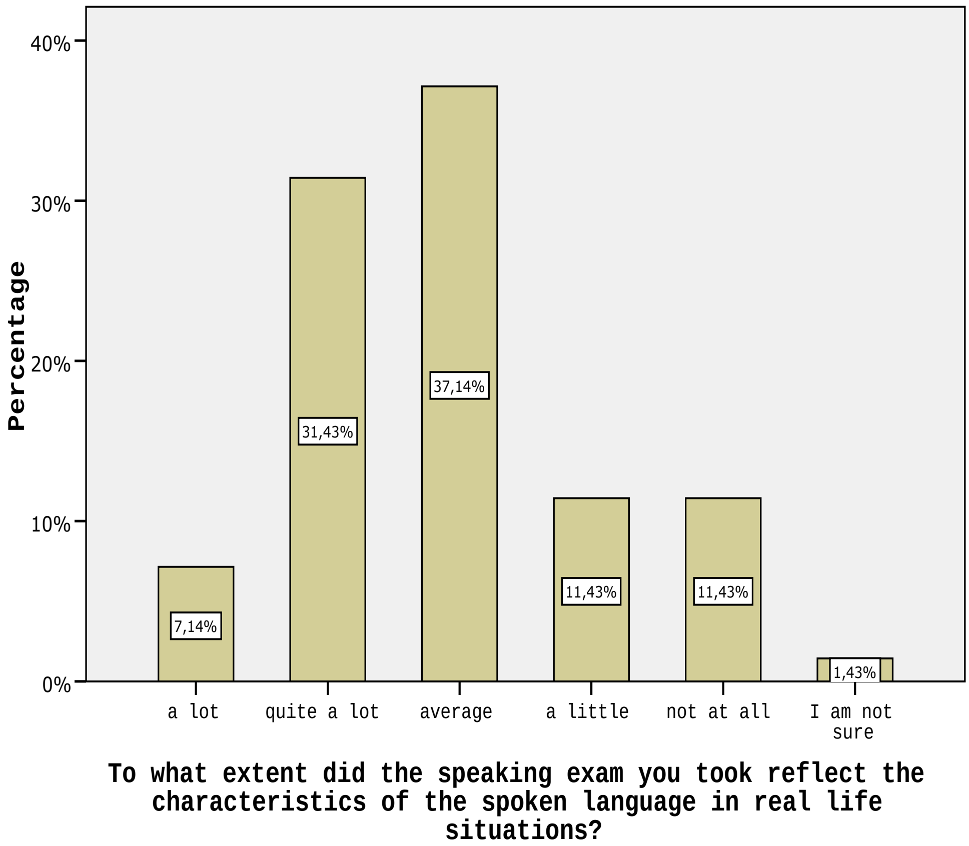


Figure 3. Subjects' opinions on the extent the speaking exam they took reflected the characteristics of the spoken language in real life situations (FVQ question 2)

As indicated in Figure 3 above, 37% of the subjects think that the degree to which the speaking exam reflects the characteristics of the spoken language in real life situations is average. However, 31% of the subjects reported this as quite a lot. This may indicate that more than half nearly of (68%) the subjects find the speaking exam they took satisfactory in reflecting the characteristics of the spoken language in real life situations. This is quite satisfactory.

Table 1 below indicates the percentages of the students answering the third question of the questionnaire. Due to the design of the questionnaire only the students choosing either **a**, **b** or **c** items in the second question were required to answer the third one. Therefore, not all of the students answered this question. As presented in Table 2, 38.6% of the students who answered the third question reported that one of the reasons why the exam reflected the characteristics of the spoken language in real life situations is that they were able to express their ideas and emotions (note that the total percentage of cases exceeds 100% since this is a multiple response question and students gave more one answer if they desired). Similarly, 35% of the students said that because they were able to speak, the exam they took reflected characteristics of the spoken language in real life. Moreover, 33% of the students reported that the reason for the exam's reflecting characteristics of the spoken language is its being mostly spontaneous and not writing a script for what they would say.

Table 1.

The Total Percentages of the Students Answering FVQ Question 3

Case Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
	57	81.4%	13	18.6%	70	100.0%

a. Dichotomy group tabulated at value 1.

Table 2.

The Frequencies of the Responses Given to FVQ Question 3

Frequencies

		Responses		Percent of Cases
		N	Percent	
Why do you think the speaking exam you took reflect the characteristics of the spoken language in real life situations?[a]	It had the parts of a normal dialogue.	15	15.6%	26.3%
	I was able to speak enough.	20	20.8%	35.1%
	I was able to ask questions freely.	8	8.3%	14.0%
	I was able to express my ideas and emotions.	22	22.9%	38.6%
	The teacher didn't tell me whether my opinion or answer was right or wrong.	9	9.4%	15.8%
	It was mostly spontaneous & I didn't write a script for what I would say	19	19.8%	33.3%
	I am not sure.	3	3.1%	5.3%
Total		96	100.0%	168.4%

a. Dichotomy group tabulated at value 1.

Likewise, for the fourth question asking why the exam they took didn’t reflect the characteristics of the spoken language in real life situations the students could give more than one response if they wanted. Furthermore, not all the students were required to answer this question due to the design of the questionnaire. When all these are considered, it is seen that 38.5% of the students who answered the fourth question reported that one of the reasons why the exam didn’t reflect the characteristics of the spoken language in real life situations is that they could not speak enough (note that the total percentage of cases exceeds 100% since it is a multiple response question) (see Table 3 & 4). Moreover, 33.3% of the students showed not being able to express their ideas and emotions as the reason for the exam's not reflecting the characteristics of the spoken language in real life situations (see Table 3 & 4).

Table 3.

The Total Percentages of the Students Answering FVQ Question 4

Case Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
	13	18.6%	57	81.4%	70	100.0%

a. Dichotomy group tabulated at value 1.

Table 4.

The Frequencies of the Responses Given to FVQ Question 4

Frequencies

		Responses		Percent of Cases
		N	Percent	
Why do you think the speaking exam you took didn't reflect the characteristics of the spoken language in real life situations?	It didn't have the parts of a normal dialogue.	2	13.3%	15.4%
	I could not speak enough.	5	33.3%	38.5%
	I could not express my ideas and emotions.	4	26.7%	30.8%
	It was spontaneous and I could write a script for what I would say.	2	13.3%	15.4%
	Other reason(s).	1	6.7%	7.7%
	I am not sure.	1	6.7%	7.7%
Total		15	100.0%	115.4%

a. Dichotomy group tabulated at value 1.

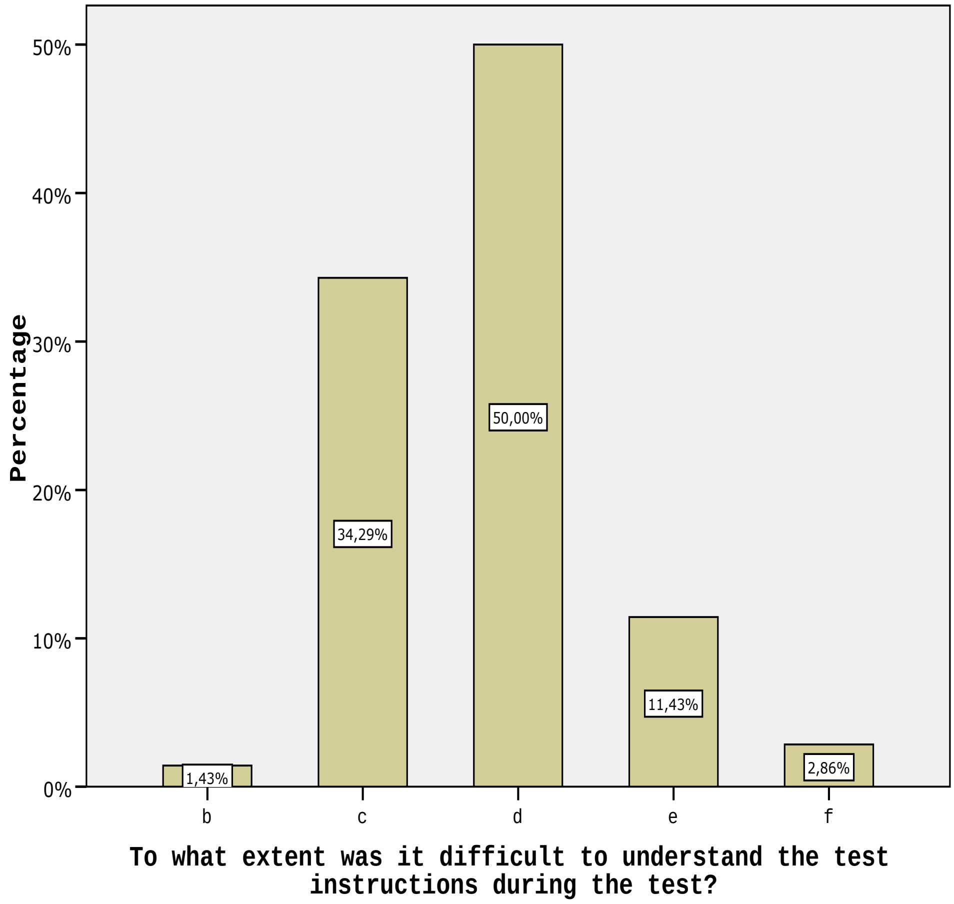


Figure 4. Subjects’ opinions on the difficulty of understanding the test instructions during the test (FVQ question 5)

The fifth question on the questionnaire includes six response categories regarding the question on the difficulty of understanding the test instructions during the test: (a) it was very difficult, (b) it was difficult, (c) it was neither easy nor difficult, (d) it was easy, (e) it was very easy and (f) I am not sure. As indicated in the Figure 4 above, although there are six response categories on the questionnaire, since none of the subjects chose item **a**, it cannot be seen in the figure. When the percentages are analyzed, it is seen that half of the students who took the questionnaire found understanding the test instructions "easy" during the test. This question may have emerged like this due to the speaking exam practices the students participate in throughout the whole semester. During those practices, the students get familiar with the tasks and the instructions of the exam.

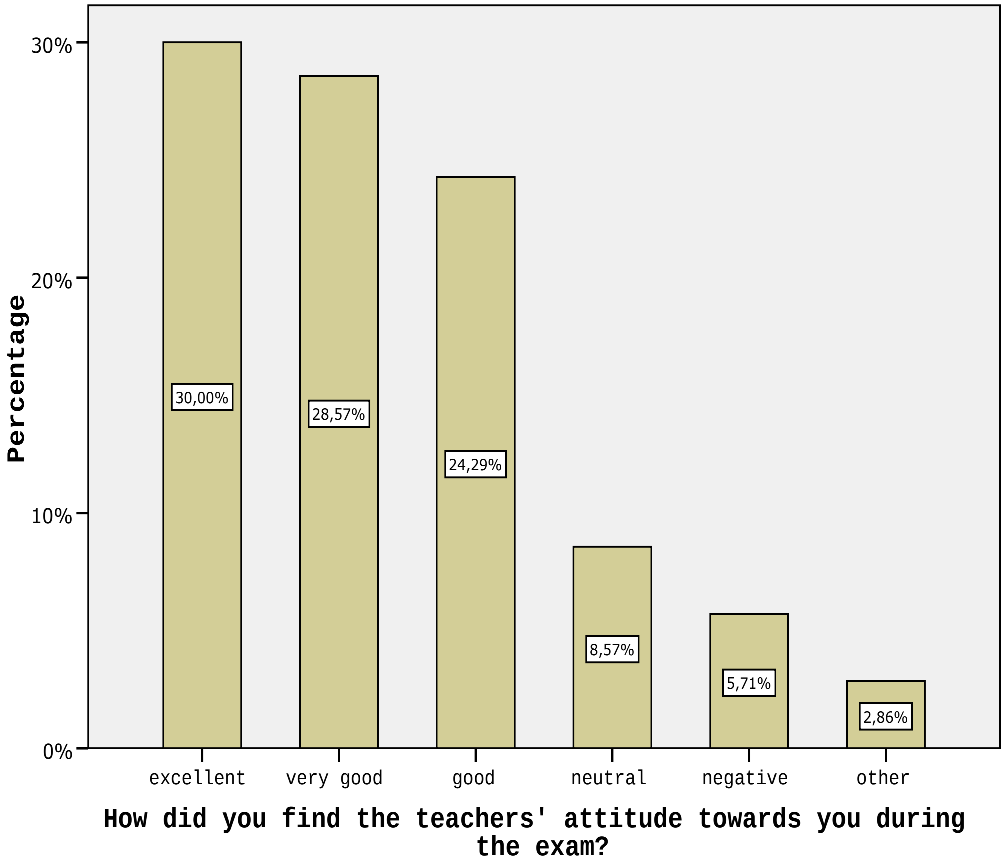


Figure 5. Subjects' opinion on the attitude of the teachers during the exam (FVQ question 6)

In the figure above, the opinions of the students about the attitude of the teachers during the exam are presented. The students seem to be content with the attitude of the teachers as nearly 83% of the subjects reported their attitude as being good to excellent. A small group of the students-2.80%- chose the "other" response category and noted that the attitude of the teachers' was below average and some teachers made them feel more stressed.

4.2.2 Responses to Open Ended Items

The face validity questionnaire included 4 open-ended items which were to be completed by the students. The responses to these items were analyzed via cross-case analysis, listing the common answers given by the students to show general tendencies.

4.2.2.1 Results of Open-ended items

The first open-ended item, which is the 7th question in the questionnaire, aimed to find out the students' comments about the speaking exam's procedure if they had any complaints to report.

Table 5.

The Results of Open Ended Items: Item 7

If you have any comments about the speaking test procedures, please write them below.	
Answer	**Frequency**
It was good.	20
There should be more interaction between the teachers and the students during the exam.	10
In the last part, the students should be given the chance to choose more than one topic.	3
It should be like a daily chat on daily topics.	2

As seen in Table 5 above, students had different comments on the speaking test procedures. Only 35 students answered this question and the majority of the students answering this question reported that the speaking test procedures were good. 10 of the students said that there should be more interaction between the examiners and the test takers. Related to the topics talked over, 3 students mentioned that it would be better to have more choices in the third part of the exam, where the students are required to choose a topic and state their ideas about it. Furthermore, some students wanted the exam procedures to be like a daily chat on daily topics as they stated this would decrease the stress level they have by giving them the chance to express their ideas more easily.

Table. 6

The Results of Open Ended Items: Item 8

What was the aspect of the speaking exam you liked most?	
Answer	**Frequency**
Teachers' positive attitude	30
Choosing a topic and talking about it	7
Choosing a picture and talking about it	7
Being with teachers and speaking with them	5
Its resemblance to daily speech (especially with the aid of warm up questions in the 1st part)	4
Getting the exam alone without any other students	3
Teachers' not correcting our mistakes	3
Having only two examiners	1

As Table 6 shows, for the 8th question which is about the aspect of the speaking exam that the students liked most, half of the students answering the question stated that it was the positive attitude of the teachers towards the students during the exam. Some of the students also noted that due to their positive attitude, they were able to rid themselves of the stress they had. There were two more frequently given answers to this question, which are about the tasks included in the exam. One of them was choosing a topic and talking about it and the other one was choosing a picture and talking about it. There were some other responses as well. For example, being with teachers and speaking with them, its resemblance to daily language especially with the aid of warm up questions, getting the exam alone without any other students, teachers' not correcting their mistakes and having two examiners only were among the other aspects of the speaking exam that the students liked most.

Table. 7

The Results of Open Ended Items: Item 9

What was the aspect of the speaking exam you disliked most?	
Answer	**Frequency**
The use of microphones and webcams during the implementation of the exam.	16
The inadequacy of the time allotted for each student	9
Speaking on your own in the third part of the exam, not having a dialogue	5
Choosing a picture and talking about it	4
Its causing stress	3
Not being given enough time for the last part of the exam	2

Including some unknown vocabulary items	2
Waiting for your turn (for the exam time)	2
Cliché and easy questions	1
Its determining our eligibility to take the TOEFL exam	1
Not looking like real life	1
Being tested alone	1
Its being very formal	1
Not being administered at previously announced time	1

Table 7 above shows the students' responses about the aspect of the speaking exam that they disliked most. 49 of the students answered this question. The majority of the students stated that the use of microphones and webcams during the implementation of the exam was the aspect of the speaking exam that they disliked most. Some of the students giving this response to this question also noted that the use of microphones and webcams caused stress during the exam. In addition, the inadequacy of the time allotted for each student was another common answer for this question. Moreover, the third part of the exam where the students were required to express their own ideas was also stated as the most disliked aspect of the exam by five students since it wasn't a dialogue. Furthermore, four students answering this question chose the second part of the exam where they are required to choose a picture and talk about it, as the most disliked aspect of the exam. As indicated in Table 7 above, there are also some other responses although they are not frequent.

Table. 8

The Results of Open Ended Items: Item 10

How could the test be improved? Please write your comments in Turkish if you have any.	
Answer	**Frequency**
Including more interaction and a dialogue	10
Including various topics	8
Extending the time limit of the exam	3
Giving a written exam	1
Not using any technological devices during the implementation	1

The tenth question aimed at finding students' ideas on how the test could be improved. Two common answers were given (see Table 8). Firstly, 10 students out of 23 stated that the test could be improved by including more interaction in it. In other words, many students were of the opinion that there should be a dialogue between the test takers and examiners as in this way listening could be integrated into the exam as well. Some of them said that it would be better if the examiners are native speakers. Secondly, as indicated in Table 8 above, the students reported that different topics should be included if the exam is to be improved so that the test takers can have more opportunities as well. Related to this, some students noted that they should be given the chance to change the topics they choose in the third part of the exam as the test taker may not have any ideas about the topic he chooses even in his native language. In addition to this, there were some other suggestions like extending the limit of the exam, giving a written exam and not using technological devices during the exam.

4.2.3 Interpretation of the Results of the Face Validity Questionnaire

The results of the questionnaire seemed to suggest that not all the students are content with the speaking exam given due to various reasons. However, when the answers are considered in general, the speaking exam seemed to possess face validity quality to a satisfactory degree in the eyes of the students and that the exam can be bettered by paying attention to the reasonable points stated by the subjects involved in the study.

4.3 Analysis of the Responses to Content Validity Interview

As mentioned before, one way of collecting data in order to analyze the content validity of an exam is getting the views of the experts like the instructors, teachers or the administrators of an institution since it is recommended to rely on a panel of experts who are familiar with the constructs that the exam measures. Therefore, interviews were conducted with the chairperson and two coordinators working at TOBB University of Economics and Technology Department of Foreign Languages in order to investigate the content validity of the exam. For the first question which was about for whom and what the test was designed for, it was stated by the first interviewee that the test was designed for the students completing one year intensive of preparatory class education. More specifically, it was designed for students completing 360 hours of education in C levels, 990 hours of education in B levels and 1260 hours of education in A levels.

Moreover, it was added that the test was designed for newly registered students as well since it aims to test both the proficiency level of the students and their readiness before English language education. In other words, the first interviewee also pointed out that the newly registered students are also given the speaking exam no matter what their English learning background is. Yet, the students are required to take the PQE (APPENDIX A) before they take the speaking exam, which is the subject matter of the research. The second, the third and the fourth interviewees also highlighted that the test was designed for preparatory class students.

For the second part of the first question it was stated by the first interviewee that the test was designed first of all in a way to reflect the principle of test in the way you teach. As it was mentioned while designing the learning situations, four basic language skills are focused during the instruction. Moreover, it was emphasized that grammar teaching is integrated into reading and listening and the text books used also are chosen according to this principle. The first interviewee stated that "Since the importance of speaking is emphasized in classroom instruction, it cannot be ignored while testing language proficiency as well…therefore, it can be said that language abilities of the learners can be better tested through speaking and writing since they are productive skills…in order to see how well language learners reflect their abilities, this component is included in the proficiency exam…." It was also stated that English language education does not end with the preparatory class at TOBB. It continues in the 1st and 2nd years of university education since the ultimate point that the students want to be brought to is to score at least 94 points from the TOEFL IBT.

The other interviewees stated that this test was mainly designed to test oral proficiency as all other skills were also tested in the proficiency exam implemented at the institution.

For the 2nd question on the appropriateness of the test to the students at the institution it was mentioned by all the interviewees that the institution wants to develop the speaking skills of the students; therefore, enough attention is given to speaking skills during the academic year. As a result, direct testing of speaking is implemented in order to see if the students can express themselves using the target language in an effective way.

For the third question all the interviewees reported that there are not any test specifications for the speaking exam at hand. In other words, they came to a conclusion that technically no specifications were developed but the exam used is constructed and modified by taking the international Common European Framework descriptors. Therefore, although there is no table of specifications that can be used to evaluate the content, it was pointed out that the content is described by those indicators dividing the learners into levels. This indicates that the institution has a rationale behind for the test that they have been using. Moreover, as indicated by all of the interviewees according to the levels determined by the Common European Framework, this speaking exam is designed for B1 level learners "who can deal with most situations likely to arise whilst traveling in an area where the language is spoken, can produce simple connected texts on topics which are familiar or of personal interest, can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.." Based on this, it was claimed that the test content is appropriate to this indicator as the institution wants the learners to reach this level of proficiency at the very least when they complete preparatory class education. However, as the second interviewee states "….B1 is our target group, yet it also differentiates the students who may reach B2 and C1 levels and their TOEFL exam scores also indicate this…", it seems to be taken by the students who may do better in terms of the functions described by B1 level indicator.

Since the content is determined by the indicator mentioned above, the interviewees stated that the items or tasks in the test match what the test as a whole is supposed to assess for the fifth question. All the interviewees agreed that the speaking exam was meant to be generally spontaneous. Therefore; it is suitable to the descriptor mentioned above. All the interviewees seemed to agree that the students are expected to produce simple connected texts on topics which are familiar or of personal interest, describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans..." Due to this the last part of the exam where the candidates are asked to state their ideas on a chosen topic was given as an example to support the fact that tasks match what the test as a whole is supposed to assess. Moreover, the first part of the exam where the students answer general warm up questions seems to match with the function requiring the students to produce simple sentences about their personal interests, experiences, dreams, hopes and ambitions.

For the 6th question whether the test produces a good sample of the contents of the syllabus of the preparatory class, all the interviewees stated that it didn't need to reflect it as this was supposed to assess the proficiency level of the students. Yet, they all pointed out that when the activities included in the syllabus were considered, they all build the base necessary for a successful performance in the exam. It was told by the interviewees that all preparatory class levels prepare presentations which give them the chance to interact both with their teacher and peers as the presentations include discussion sessions at the end. Moreover, they stated that conversation clubs are designed for all levels by the administration in order to foster the speaking skills of the students. In these conversation clubs, several speaking activities are done in order to give the students a chance to use the target language. Furthermore, the content of the conversation clubs is said to be modified through the end of the second semester by including exam oriented activities in order to meet the needs of the learners and to reduce their anxiety level as well.

For the seventh item that is questioning how well the tasks/ items of the test reflect the characteristics of speaking ability, similar answers were given by the interviewees. They all mentioned that since it cannot be wholly authentic, they do their best in order to test it as directly as possible. Therefore; the items included in the exam reflect the characteristics of speaking ability fairly well since it was thought to include several speech acts like greeting, describing something, expressing ideas, exemplifying and agreeing or disagreeing in different parts of the exam. Moreover, the second interviewee also stated that some kind of interactiveness can also be seen in the first part of the exam where the students answer some warm up questions asked by the examiners.

Furthermore, all the interviewees agreed that no research was conducted to determine the test content. However, they stated that the speaking exams appropriate to different proficiency levels can be found in the literature. As a result, the institution preferred to modify the speaking part of the FCE and IELTS exam in order to meet their objectives. Similarly, it was stated that no research was conducted to evaluate test content. Yet, as stated by the first interviewee, feedback taken from the students and the instructors are considered by the administration at times.

In order to answer the tenth question, the interviewees analyzed each part of the exam to see if the tasks and topical contents are relevant to the target language use. They commonly mentioned that the first and the third parts of the exam resemble real life situations. It was stated that the first part includes basic questions about the test taker's home town, family, work or study, leisure and future plans. And the last part requires the exam taker to express their opinions by supporting them with specific examples and evidence. It was pointed out that these can be included in the situations that the test taker is likely to encounter. For the second part it was said that the way it was implemented didn't resemble a real life situation. However, one may also need to describe a place, a person or a thing in real life as well. Therefore, considering the themes of the pictures they concluded that part did not also present a big discrepancy between the real life and exam situation. Additionally, the second interviewee stated that "speaking construct is a broad domain and you cannot include all the speech acts or likely situational uses in it. Therefore, you need to make a logical sampling of this broad domain by deciding on the indicators showing that one can effectively express himself using the target language. However, sometimes these choices are influenced by practical constraints…" This may indicate that the institution is aware of the difficulties of testing speaking and they do their best to implement it as well as possible.

Taking the answers given during the interview into consideration it can be concluded that the interviewees are of the opinion that the speaking exam implemented meets the objectives of their programme as they aim to assess all language skills in their proficiency exam. Moreover, since this is "a proficiency exam measuring people's ability in a language, regardless of any training they may have had in that language", they all think it does not have to be directly based on the content and the objectives of language courses given in their department. However, regardless of this, it is seen that they still try to prepare their students for the speaking exam with the help of the exam practices done throughout the semester in lessons and conversation club activities. Due to this, it is seen that the implementation of the speaking exam has a beneficial backwash effect on teaching and learning as they encourage oral ability throughout the semester by testing oral ability in their proficiency exam.

4.4 Predictive Validity Analysis

In the quantitative part of the research, the predictive validity of the speaking exam given in the preparatory year was examined. In order to investigate it, the departmental speaking exam scores of the students passing the December TOEFL-ITP were used. Since not all the students passing the December TOEFL-ITP were able to take departmental English courses 101 and 102 for different reasons, the scores of the students passing the September TOEFL-ITP were also used in order to calculate the predictive validity as well.

First of all, the preparatory speaking exam scores of the students passing either the September or December TOEFL-ITP exam were collected. Next, to be able to conduct this part of the research, the preparatory speaking exam scores of the students who were able to get 101 and 102 English courses were found out. Then, the departmental speaking exam scores of the same students were collected as well. After this, prior to Pearson Product Moment correlation coefficient analysis, a scatter plot of these two grades, as shown in Figure 6, was obtained to give a rich descriptive picture of the relationship between two variables. Next, correlation coefficients were computed between the two test scores in order to indicate the relationship between them. Moreover, to verify the obtained results from the correlation analysis and to determine whether there is a linear relationship between preparatory class speaking exam grades and departmental speaking exam grades, regression analysis was also used.

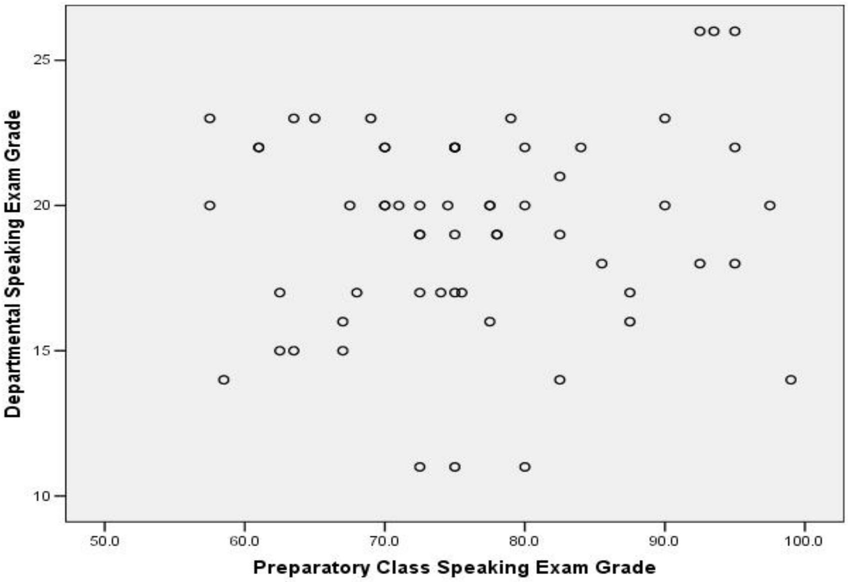


Figure 6. The Scatter Plot of Two Grades

As can be inferred from the scatter plot above, there is no linear pattern in the scatter plot indicating the absence of a linear relationship between preparatory class speaking exam grades and departmental speaking exam grades. Moreover, the correlation coefficient computed supported the result indicated by the scatter plot (see Table 9).

Table 9. *The Correlation Coefficients between Preparatory Class Speaking Exam Grade and Departmental Speaking Exam Grade*

Correlations

		Preparatory Class Speaking Exam Grade	Departmental Speaking Exam Grade
Preparatory Class Speaking Exam Grade	Pearson Correlation	1	.120
	Sig. (2-tailed)		.361
	N	60	60
Departmental Speaking Exam Grade	Pearson Correlation	.120	1
	Sig. (2-tailed)	.361	
	N	60	60

As seen in the table above, the correlation coefficient between preparatory class speaking exam grades and departmental speaking exam grades is 0.120. This shows that there is a weak linear association between these grades. Moreover, this correlation is not statistically significant since p-value is greater than 0.05. However, bivariate linear regression analysis was also conducted to verify the obtained results (see Table 10).

Table 10. *The Bivariate Regression Analysis of Preparatory Class Speaking Exam and Departmental Speaking Exam Grades*

Model Summary

Model	R	R Squared	Adjusted R Square	Std. Error of the Estimate
1	.120[a]	.014	-.003	

a. Predictors: (Constant), Preparatory Class Speaking Exam Grade

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	54.030	10.923		4.946	.000
	Preparatory Class Speaking Exam Gr	.131	.142	.120	.922	.361

a. Dependent Variable: VAR00001

As indicated in Table 10, R Squared is 0.014, which shows that only 1.4 % of the variation in departmental speaking exam grade is explained by preparatory class speaking exam grade. In other words, there is only 1.4% agreement between one set of scores and the other. This is a very low percent indicating that this regression line is useless. Moreover, this fact is justified by the p-value of the regression coefficient. The p-value is greater than 0.05, which means that this regression line is statistically insignificant. This may be due to the fact that the two speaking tests are differently constructed, as preparatory class speaking exam is conducted with human examiners, yet the other one with computers like the speaking section of the TOEFL IBT exam. Yet, in order to get those courses 101 and 102, which include that speaking exam, the students need to complete a preparatory year of education by getting all the necessary exams including the one which is the subject matter of this study.

Regardless of these results, in order to see if there is a difference between the students taking the September TOEFL and the December TOEFL exams, correlation coefficients of these two groups were also computed to examine the predictive validity. Since the data, the results of which are shown in the Table 9 above, included the 2007 registered students who passed the speaking, writing and the TOEFL exam implemented in September, it was thought that there might be some differences between the students taking all those exams in December. Therefore, bivariate regression analysis was also calculated separately for each group in order to see if the speaking exam implemented in preparatory year predicts the performance of the students' scores in the speaking exams given in 101 and 102 departmental English courses.

Table 11. *The Bivariate Regression Analysis of Preparatory Class Speaking Exam and Departmental Speaking Exam Grades of students taking the September Proficiency*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.233[a]	.054	.031	3.637

a. Predictors: (Constant), Preparatory Class Speaking Exam Grade

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	13.195	3.927		3.360	.002
	Preparatory Class Speaking Exam Gr	.076	.050	.233	1.518	.137

a. Dependent Variable: Departmental Speaking Exam Grade

The table above presents the findings of the regression analysis which was computed using the scores of 42 students taking the September speaking exam. Using their scores of the departmental speaking exam, bivariate regression analysis was conducted. As indicated in Table 11, R Square is 0.054 which shows that only 5.4% of the variation in departmental speaking exam grade is explained by preparatory class speaking exam grade. In other words, there is only 5.4% agreement between one set of scores and the other. This is a very low percentage indicating that this regression line is useless. Moreover, this fact is justified by the p-value of the regression coefficient. The p-value, which is 0.137, is greater than 0.05. This means that this regression line is statistically insignificant.

Table 12. *The Bivariate Regression Analysis of Preparatory Class Speaking Exam and Departmental Speaking Exam Grades of students taking the December Proficiency Exam*

Model Summary

Model	R	R Squared	Adjusted R Square	Std. Error of the Estimate
1	.258 [a]	.066	.008	3.041

a. Predictors: (Constant), Preparatory Class Speaking Exam Grade

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	25.976	6.164		4.214	.001
	Preparatory Class Speaking Exam G	-.089	.083	-.258	-1.067	.302

a. Dependent Variable: Departmental Speaking Exam Grade

Table 12 above indicates the findings of the regression analysis which was computed using the scores of the 2007 registered students taking the December speaking exam and taking 101, 102 departmental English courses in the second term as well. Using these two sets of scores, bivariate regression analysis was conducted. As indicated in the table, R Square is 0.066 which shows that only 6.6 % of the variation in departmental speaking exam grade is explained by preparatory class speaking exam grade, which means there is only 6.6% agreement between one set of scores and the other. This is also a very low percentage, indicating that this regression line is useless like in the other two regression analyses above. Moreover, this fact is justified by the p-value of the regression coefficient. The p-value, which is 0.302, is greater than 0.05. This means that this regression line is statistically insignificant.

As seen in all the calculations to examine predictive validity, there is no significant relationship between two sets of scores. As the findings of the regression analyses suggest, preparatory class speaking exam grades do not explain the departmental speaking exam grades statistically. That is, the level of agreement between one set of scores and the other is very low.

4.5 Construct Validity Analysis

As mentioned earlier, one way of assessing the construct validity of a test is "to correlate the different test components with each other" (Alderson, et al., 1995, p.184). The correlations between different test components are expected to be fairly low as they all contribute to the overall picture of the language ability by measuring something different. However, if the components of a test correlate very highly with each other, the two tests may be questioned if they are testing the same skills or the same thing. On the other hand, the correlations between the whole test and each subtest might be expected to be around +.7 or more as the overall score is a more general measure of language ability than each subtest (Alderson, et al., 1995, p. 184). Therefore, in order to analyze the construct validity of the exam, Pearson Product Moment Correlation Coefficients between the scores of the students in the speaking exam and the scores of the students in each subtest of the TOEFL exam were computed.

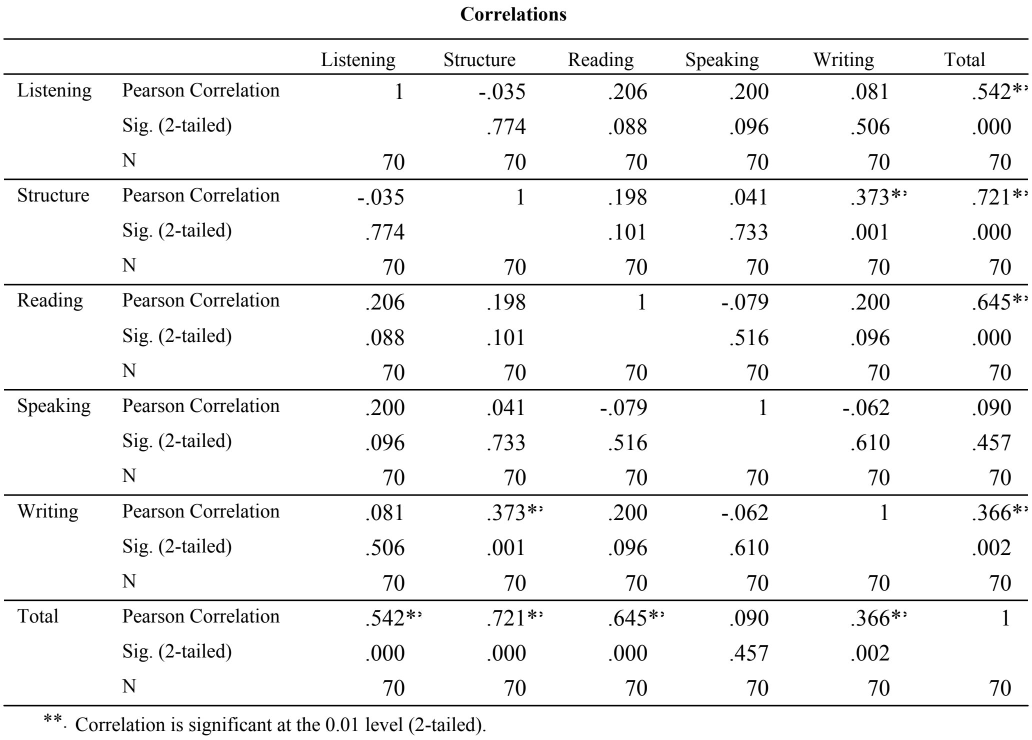
Table 13. *The Correlation Coefficients between the Speaking Exam scores and the Scores of each Subtest and the Total*

		Listening	Structure	Reading	Speaking	Writing	Total
Speaking	Pearson Correlation	.200	.041	-.079	1	-.062	.090
	Sig. (2-tailed)	.096	.733	.516		.610	.457
	N	70	70	70	70	70	70

In the above table, the correlations between speaking test scores and the scores of other test components are shown. It is seen that all of the correlation coefficients are very low indicating that there is no strong linear association between the speaking test score and the scores of the other test components. The highest correlation in this table is between the speaking and the listening scores (0.2). However, it is also fairly low. Furthermore "Sig. (2-tailed)" row shows that none of these six correlations (including the one with the total) are statistically significant since all of the p-values are greater than 0.05. These low correlations between the speaking scores and the other subtests indicate that they are testing different constructs. Moreover, the correlation between the speaking and the total scores is not statistically significant as the p-value is greater than 0.05.

As mentioned, bivariate correlation coefficients between all of the components were also computed to better analyze the results. The findings of this calculation are indicated in Table 14.

Table 14. *Bivariate Correlation Coefficients between all the test components*



Correlations

		Listening	Structure	Reading	Speaking	Writing	Total
Listening	Pearson Correlation	1	-.035	.206	.200	.081	.542**
	Sig. (2-tailed)		.774	.088	.096	.506	.000
	N	70	70	70	70	70	70
Structure	Pearson Correlation	-.035	1	.198	.041	.373**	.721**
	Sig. (2-tailed)	.774		.101	.733	.001	.000
	N	70	70	70	70	70	70
Reading	Pearson Correlation	.206	.198	1	-.079	.200	.645**
	Sig. (2-tailed)	.088	.101		.516	.096	.000
	N	70	70	70	70	70	70
Speaking	Pearson Correlation	.200	.041	-.079	1	-.062	.090
	Sig. (2-tailed)	.096	.733	.516		.610	.457
	N	70	70	70	70	70	70
Writing	Pearson Correlation	.081	.373**	.200	-.062	1	.366**
	Sig. (2-tailed)	.506	.001	.096	.610		.002
	N	70	70	70	70	70	70
Total	Pearson Correlation	.542**	.721**	.645**	.090	.366**	1
	Sig. (2-tailed)	.000	.000	.000	.457	.002	
	N	70	70	70	70	70	70

**. Correlation is significant at the 0.01 level (2-tailed).

As indicated in the above correlation matrix which includes bivariate correlations between the all test components, the highest correlations are between the structure test scores and the total scores (0.721) and between the reading test scores and the total scores (0.645). Out of these 15 correlations, five of them are statistically significant, namely, structure and writing, listening and total, structure and total, reading and total, writing and total. If the individual component scores are embedded in the total scores for the test, the correlations are expected to be high (around +.7 or more) as the correlations will be partly between the test component and itself. The individual components, reading, listening and structure are embedded in the total scores of the TOEFL-ITP indicated in Table 14. However, only the correlation between the structure scores and the total test scores is above +.7, indicating that this component has a strong effect on the final total score. The correlations between the other two test scores (reading and listening) and the total scores are also statistically significant but not above +.7 as suggested in the literature although they are embedded in the total scores for the test. The two individual components, which are not included in the total test scores are writing and speaking. Similarly, high correlations are expected between these two sets and the whole test since overall score is a more general measure of language ability than each subtest (Alderson, et al., 1995, p. 184). However, only the writing component appears to have a statistically significant correlation, but it is not as high as expected as well (0.366). The correlation between the speaking scores and the total test scores is 0.090. It is an interesting fact that all of the bivariate correlations between the total test score and the test components are statistically significant except the one between the total and the speaking test scores. The fact that the writing and speaking correlations are on the low side (0.366 and 0.090) may be due to the fact that these subtests proved to be unreliable and correlations between unreliable tests lead to low correlation coefficients as the results are partly due to chance (Alderson, et al., 1995, p. 185).

4.6. Reliability Analysis

The reliability analysis of the speaking exam investigated was done in two steps. Since this is a test of production where raters' judgments affect the decision to be made about the performances of the students, intra-rater and inter-rater reliability levels were examined by calculating the correlation coefficients of the scores given by the raters. First, the inter-rater reliability analysis part of the study was presented. The results are displayed in tables and the results discussed. Next, the results of intra-rater reliability analysis were explained and displayed in tables as well.

4.6.1 Inter-rater Reliability

Table 15. *The Correlation Coefficients of each pair's ratings*

		First rater first rating	Second rater first rating
First rater, first rating	Pearson Correlation	1	.910(**)
	Sig. (2-tailed)		.000
	N	19	19
Second rater, first rating	Pearson Correlation	.910(**)	1
	Sig. (2-tailed)	.000	
	N	19	19
		Third rater first rating	Fourth rater first rating
Third rater ,first rating	Pearson Correlation	1	.914(**)
	Sig. (2-tailed)		.000
	N	18	18
Fourth rater, first rating	Pearson Correlation	.914(**)	1
	Sig. (2-tailed)	.000	
	N	18	18
		Fifth rater first rating	Sixth rater first rating
Fifth rater, first rating	Pearson Correlation	1	.487(**)
	Sig. (2-tailed)		.035
	N	19	19
Sixth rater, first rating	Pearson Correlation	.487(**)	1
	Sig. (2-tailed)	.035	
	N	19	19

In the above table, the correlation coefficients estimating the inter-rater reliabilities of three pairs of raters are given. It was estimated by looking at the scores produced by two raters in each pair. The scores were lined up in columns and the results were obtained by calculating a correlation coefficient between two sets of scores on SPSS. As indicated in the table, the correlation coefficients obtained for the first two pairs are 0.910 and 0.914, respectively, indicating quite high inter-rater reliabilities. However, the inter-rater reliability of the third pair of raters is 0.487, which is fairly low. Regardless of this, the correlation coefficients for all pairs are statistically significant with p-values are smaller than 0.05.

4.6.2 Intra-rater Reliability

Table. 16 *The Correlation Coefficients of 1st pairs' 1st and 2nd Ratings of the Same Students*

		First rater first rating	First rater second rating
First rater first rating	Pearson Correlation	1	.776(**)
	Sig. (2-tailed)		.000
	N	19	19
First rater second rating	Pearson Correlation	.776(**)	1
	Sig. (2-tailed)	.000	
	N	19	19

		Second rater first rating	Second rater second rating
Second rater first rating	Pearson Correlation	1	.933(**)
	Sig. (2-tailed)		.000
	N	19	19
Second rater second rating	Pearson Correlation	.933(**)	1
	Sig. (2-tailed)	.000	
	N	19	19

In order to estimate the intra-rater reliability of the first pair, the correlation coefficients are obtained as indicated in Table 16 above. The intra-rater reliabilities of the first two raters' ratings are 0.776 and 0.933 respectively, which are quite high. Moreover, both of the correlation coefficients estimating these reliabilities are statistically significant since their p-values are satisfactory. The same procedures were followed to find out the intra-rater reliability level for the raters of the second pair.

Table. 17 *The Correlation Coefficients of 2nd pairs' 1st and 2nd ratings of the same students*

		Third rater first rating	Third rater second rating
Third rater, first rating	Pearson Correlation	1	.560(**)
	Sig. (2-tailed)		.016
	N	18	18
Third rater ,second rating	Pearson Correlation	.560(**)	1
	Sig. (2-tailed)	.016	
	N	18	18

		Fourth rater first rating	Fourth rater second rating
Fourth rater, first rating	Pearson Correlation	1	.727(**)
	Sig. (2-tailed)		.001
	N	18	18
Fourth rater ,second rating	Pearson Correlation	.727(**)	1
	Sig. (2-tailed)	.001	
	N	18	18

The correlation coefficients obtained for the second pair are as indicated in Table 17 above. The intra-rater reliability levels of the second two raters' ratings are 0.560 and 0.727. The intra-rater reliability of the third rater's ratings is not high. However, the intra-rater reliability of the fourth rater's ratings is quite high. Moreover, both of the correlation coefficients estimating these reliabilities are statistically significant since the p-values are very satisfactory.

Table 18 below indicates the correlation coefficients calculated for the third pair in order to estimate intra-rater reliability.

Table. 18 *The Correlation Coefficients of 3rd pairs' 1st and 2nd Ratings of the Same Students*

		Fifth rater first rating	Fifth rater second rating
Fifth rater, first rating	Pearson Correlation	1	.796(**)
	Sig. (2-tailed)		.000
	N	19	19
Fifth rater, second rating	Pearson Correlation	.796(**)	1
	Sig. (2-tailed)	.000	
	N	19	19

		Sixth rater first rating	Sixth rater second rating
Sixth rater ,first rating	Pearson Correlation	1	.582(**)
	Sig. (2-tailed)		.009
	N	19	19
Sixth rater ,second rating	Pearson Correlation	.582(**)	1
	Sig. (2-tailed)	.009	
	N	19	19

The intra-rater reliability levels of the fifth and sixth raters' ratings are 0.796 and 0.582. The intra-rater reliability of the fifth rater's ratings is quite high whereas the intra-rater reliability of the sixth rater’s ratings reliability is not as high as his partner. Moreover, both of the correlation coefficients estimating these reliabilities are statistically significant since the p-values are satisfactory.

When all the tables representing the intra-rater reliability levels are considered, it can be concluded that all of the intra-rater reliability coefficients are statistically significant as they are smaller than 0.05. However, among the six raters, the second rater’s ratings seem to be the most reliable, whereas the third and sixth rater’s ratings seem to be the least reliable as they are lower than .70, which is the adequate level for oral tests (Brown, 1996, Hughes, 1989, Lado, 1961).

As can be seen from the reliability analysis part of the research, the inter-rater reliability indices of the raters participating in the study are good except for one pair as their correlations are below .70 (see Table 18). Moreover, the intra-rater reliability estimated by correlating the first and the second ratings of each rater shows differences. Except for the third and the sixth rater, they are generally satisfactory. Although the correlations indicating the intra-rater reliability of the third and the sixth raters’ ratings are statistically significant, it can be claimed that they are not so high when the minimum desirable level for oral tests (.70) is considered.

These values obtained may have had some impact on validity studies which were also included in this research as there is an inevitable conflict between reliability and validity in language tests (Underhill, 1982, p. 17). This clearly indicates that there is no point in measuring something reliably unless what is measured is known. However, it is also known that reliability is a pre-requisite for validity. Therefore, it is difficult to have both reliable and valid tests especially when assessing the communicative ability of language learners, which puts forward the need to regard validation studies as ongoing processes checking these two conflicting concepts continuously.

The results obtained and their implications will be discussed in a more detailed way in the following chapter.

CHAPTER V

CONCLUSION

5.0 Presentation

In this chapter, first a summary of the study is given. Second, the results obtained are reviewed and discussed. Next, an assessment of the study is presented. Finally, some implications are given for further research.

5.1 Summary of the Study

This study on the validation of a speaking exam was carried out at TOBB University of Economics and Technology, Department of Foreign Languages. The subjects were 2007 registered level students taking the September and the December TOEFL-ITP exam and three informants from the administration. The researcher was also an instructor at this department between 2005 and 2007.

This study focused on the speaking exam given in preparatory year education to find out if the exam is a valid and a reliable one. For this, a questionnaire was implemented, interviews were conducted and the scores of students were made use of in order to apply different statistical calculations. In other words, the results of this study were obtained through questionnaires, interviews, the students' speaking exam results, TOEFL-ITP exam results and departmental speaking exam scores. The questionnaires were analyzed by frequencies and percentages of responses and the results of the questionnaires were used to determine the face validity of the speaking exam. The results of the questionnaires were displayed in tables and figures. Furthermore, to examine the content validity of the exam, the interviews were analyzed in detail and common points from each interview were emphasized. To analyze the data used to determine the predictive validity of the exam, Pearson Product Moment Correlation Coefficients were calculated and Simple Linear Regression Analysis was conducted. Similarly, to investigate the construct validity of the exam Pearson Product Moment Correlation Coefficients between speaking test scores and each subtest scores were calculated. To estimate the intra and inter-rater reliability level of the exam, correlation coefficients were calculated for these as well. All the statistical results have been presented in tables and diagrams in the preceding chapter.

5.2 Results

This section discusses the findings of the study and draws conclusions about the research questions outlined in Chapters 1-4. Each subsection relates to one of the research questions. Where relevant, references to other reported research in the literature are presented.

This study set out to answer the following research questions regarding speaking assessment at TOBB University of Economics and Technology.

1. How valid is the test?

To answer the first question, the following sub-questions need to be investigated.

1.a How satisfactory is the test with respect to face validity?

1.b How satisfactory is the test with respect to content validity?

1.c How satisfactory is the test with respect to predictive validity?

1.d How satisfactory is the test with respect to construct validity?

2. How reliable is the test?

To answer the second question, the following sub-questions need to be investigated.

2.a How satisfactory is inter-rater reliability?

2.b How satisfactory is intra-rater reliability?

To answer the first set of research questions, the face validity of the exam was first investigated. The face validity questionnaire was give- to the students to find out students' ideas and attitudes about the speaking exam and its characteristics. All the items in the questionnaire were analyzed by calculating the frequencies and the percentages. Based on item 11 pertaining particularly to subjects' understanding of the test's face validity, it was seen that 47% of the subjects think the exam has satisfactory face validity while 57% of them do not. Another, 27% of the subjects perceived the speaking exam's face validity as adequate. When this finding and the results presented in Figure 1 are considered, the face validity of the exam may be fairly described as satisfactory. Therefore, it can be claimed that the exam possesses this quality to an adequate degree.

The reason the subjects think that such a direct measure has only moderate face validity may be understood from the results presented in Figure 2 as nearly 46% of the subjects involved in the study reported that “speaking with a native speaker on a given topic in English” is the most accurate way to assess someone’s speaking ability. The results shown in Figure 3 provide one possible explanation: that 37% of the subjects think that the speaking exam they took reflected the characteristics of the spoken language in real life situations to an average degree. However, 31% of the subjects reported this as quite a lot. This may indicate that more than half of (nearly 68%) the subjects find the speaking exam they took satisfactory in reflecting the characteristics of the spoken language in real life situations. Despite this, the rest had two common reasons given for their views. They were as follows: they could not speak enough and they weren’t able to express their ideas and emotions (see Table 4). The reasons for this are not clear but possible causes may include the following: Firstly, for the open ended item questioning the most disliked part of the exam, the majority of the students reported the use of microphones and web cams. This may possibly have caused stress during the exam and the students may not have expressed their ideas and emotions as well as they desired. As the second most disliked aspect of the exam, the subjects reported the inadequacy of time allotted for each student. Due to this, the students might have thought that they couldn’t speak enough.

The students seemed not to have any problems with the instructions of the exam and the attitude of the teacher’s during the exam (see Figures 4 and 5). The instructions were found mostly easy since the students have the chance to practice enough for the exam during the semester. Moreover, the positive attitude of the teachers, which was found good-excellent by 83% of the students, may have affected the clarity of the instructions in a positive way. In addition to these, most of the students commenting on the exam procedures seem to be content with it as they said that it was good. Related to the aspect of the exam that the students liked most, the majority of the students again reported that it was the positive attitude of the teachers. However, as mentioned earlier, the use of microphones and web cams were found to be the most disliked aspect of the exam.

The students commenting on how the exam can be improved seemed to be of the opinion that more interaction should be included in the exam. Furthermore, they stated that the variety of the topics discussed in the exam should be increased. Underhill (1987) emphasizes the importance of choosing topics by saying:

> Choosing the topic is very important. It should be relevant to the aims of the programme or the needs of the learners and should contain new information or put over a new point of view. It should not be so specialized that only the speaker himself is interested, nor should it be so general that it has no apparent purpose other than as a language exercise (p.47).

This clearly indicates that special attention should be paid while choosing the topics to be discussed in the exam as the performance of the students should not be influenced by the difficulty of the topic itself.

In the light of these, it is hoped that some work can be done to enhance the speaking exam's face validity and the additional information gained from this questionnaire will be of great use for this effort.

For the second part of the first research question, as a result of the interviews held to investigate the content validity of the exam it was found out that all the informants who participated were of the opinion that the content of the speaking exam they implement should not necessarily need to be related to the content or the objectives of their language programme since it is supposed to be an oral proficiency exam. Therefore, as they stated, the exam was designed to measure the students' ability in English regardless of any training they have had. Regardless of this, the institution still seemed to prepare their students for the speaking exam during the preparatory year education with the help of in class and conversation class activities. It is obvious that the students are somehow familiar with the content of the speaking exam although it was stated by the informants that the content of the exam does not need to be a reflection of what had been taught. As it is known, to determine the content validity of a language test, the test's content should be examined to see if it includes a representative sampling of what has been taught in a particular course and if the content is in line with the predetermined course objectives and test specifications (Anastasi, 1988; Bachman, 1991; Brown 1996; Heaton 1990; Hughes, 1989). As a result, since this exam is supposed to assess the oral proficiency level of the students and when the wide spectrum of the aspects determining one's ability to speak English is considered, the institution seems to make an effort to include a representative sample of these as much as possible. However, one of the points that all the informants mentioned was related to the interaction which can be increased during the implementation of the exam. Although it is included in the first part of the exam, where the candidates answer general questions about everyday life, it is clear from the interviews and the results of the FVQ that all the participants believe in the importance of it as it is regarded as one of the aspects which can enhance the authenticity of the exam. Yet, although interaction is limited to a minimum amount, the institution's determination to measure it in a direct way cannot be disregarded as well.

Considering these and the fact that this speaking exam is given as an oral proficiency test, it can be said that it possesses the quality of content validity to a moderately high degree.

To answer the third part of the first research question statistical calculations were carried out as mentioned before. The results of the Pearson Product Moment Correlation Coefficient indicated that there is a weak linear association between preparatory speaking exam and departmental speaking exam grades. The correlation found is not statistically significant since the p-value which is 0.120 is greater than 0.05. The bivariate linear regression analysis conducted to verify the obtained results also indicated that only 1.4 % of the variation in departmental speaking exam grades is explained by preparatory class speaking exam grades (see Table 9). This means that there is only 1.4% agreement between one set of scores and the other. This fact was also justified by the p-value of the regression coefficient. The p-value is greater than 0.05, and it means that this regression line is statistically insignificant. Regardless of this, regression analysis was calculated for two groups separately to see whether there is a difference between the students taking the September TOEFL and the December TOEFL-ITP exams. The results of them also indicated that the agreement between one set of scores and the other is very low. The R Squares obtained for the students passing the September and December speaking exams are 0.054 and 0.066 respectively (see Table 10 and 11)

All the results of these statistical calculations done to determine the predictive validity of the exam show that there is no significant relationship between two sets of scores. Moreover, as the findings of the regression analyses suggest, preparatory class speaking exam grades do not explain the departmental speaking exam grades statistically. That is, the level of agreement between one set of scores and the other is low.

However, these do not mean that this exam is not doing the job that it is supposed to be doing and it should not be used since similar studies in literature have also had similar results. For example, Dooey (1999) examined the correlation between IELTS test scores and semester weighted averages and found out that future academic success was not guaranteed by high IELTS scores, which indicates that high IELTS grades do not predict future academic success. Similarly, Ösken (1999) investigated the predictive validity of midterm achievement tests administered at Hacettepe University, Department of Basic English (DBE) and the study indicated that some of the midterm achievement tests had only a moderate amount of predictive validity. The researcher speculated that this was because of the differences between the form and content of the tests.

In the same way, these results obtained in predictive validity part of the study may have emerged in this way due to the differences between the forms and the contents of the two tests, as each speaking test is differently constructed. Preparatory class speaking exam is done with human examiners, but the other one is computer based like the speaking section of the TOEFL IBT exam. Furthermore, the content of the two exams differ. In the preparatory class speaking exam, as explained before, there are three different sections each of which requires the candidates to express their ideas according to the nature of the tasks included. However, the TOEFL IBT, like the speaking exam of departmental English courses 101 and 102, includes 2 tasks to express an opinion on a familiar topic and 4 tasks to speak based on what is read and listened to. As can be seen, the students are required to integrate what they listen to or read with their speaking. Therefore, the grading is done by taking these into consideration as well. In other words, for those 4 tasks, the students cannot receive full credits if they cannot understand and integrate what they have read or listened to no matter how well or fluently they speak. This presents the difference between the criteria used to assess the performances of the students for each of the speaking exam given. This may also possibly have affected the obtained results of correlations and regression analyses.

As a result of the points mentioned earlier, the speaking exam given in preparatory year education does not seem to have a satisfactory predictive validity when the scores are correlated with the departmental speaking exam scores.

In order to answer the fourth part of the first research question, which was on the construct validity of the exam, Pearson Product Moment Correlation Coefficients between the scores of the students in the speaking exam and the scores of the students in each subtest of the TOEFL exam were computed.

As a result of the calculations, low correlations were found, indicating that there is no strong linear association between the speaking test scores and the scores of other test components (see Table 13). The low correlations that were found out between the speaking exam scores and the other subtests indicate that they are testing different constructs. Moreover, as a result of the bivariate correlation coefficients computed between all of the components of the test, it was seen that all of the bivariate correlations between the total test scores and the test components are statistically significant except for the one between the total and the speaking test scores contrary to what is claimed in the literature (see Table 14). It is known that when the individual component score is included in the total score for the test, the correlation will be inflated artificially. Therefore, it is normal to expect high correlations (around +.7 or more) between the structure, reading and listening components and the total test score as they are included in the total score. However, in this study only the correlation between the structure scores and the total test scores is above +.7, indicating that this component has a strong effect on the final total score. The correlations between the other two test scores (reading and listening) and the total scores are also statistically significant but not above +.7 although they are embedded in the total scores as well. The correlation between the writing scores and the total scores is also statistically significant but not high enough. This is an interesting fact since all of the bivariate correlations between the total test scores and the test components are statistically significant except for the one between the total and the speaking test scores. In addition, the fact that the writing and speaking correlations are on the low side (0.366 and 0.090) may be due to the fact that these subtests proved to be somewhat unreliable and correlations between unreliable tests may lead to low correlation coefficients for validity as the results are partly due to chance.

Considering all these, it can be claimed that the speaking exam given has certain degree construct validity as the correlations between the speaking test scores and the scores of other test components are very low. Yet, the insignificant correlation between the total and the speaking test score is interesting.

For the second set of research questions including two sub sections, the Pearson Product Moment Correlation Coefficients were calculated to estimate the rater reliability of the exam. Similar to the findings of Halleck (1996) investigating the inter-rater reliability of trained raters on Oral Proficiency Interviews (OPI), as a result of the correlations computed, statistically significant results were obtained since all p-values were satisfactory (see Table 15). However, contrary to the high correlations ranging from 0.93 to 0.83 found in Halleck's (1996) study, not all the correlations found in this inter-rater reliability analysis are as high as they preferably should be. The correlation coefficients obtained for the first two pairs are 0.910 and 0.914 respectively, which are quite high inter-rater reliabilities. However, the inter-rater reliability of the third pair is 0.487, which is fairly low although it is also statistically significant.

Similarly, another study reporting lower correlation coefficients is that of Jafarpur (1988). An FSI-type oral interview was used in his study and it was conducted at Shiraz University. The performances of 58 students were scored by 3 raters and inter-rater reliability was reported as between 0.58 and 0.65. The researcher indicated that since the raters were not language teachers who received some training, low correlations may have emerged.

In the light of these, it can be said that the inter-rater reliability of the exam is not as satisfactory as is expected since the correlation of the scores of the third pair is fairly low. However, estimating the inter-rater reliability levels of all the pairs who took part in the scoring could have given more sound results. In addition, the published evidence on inter-rater reliability suggests that high correlation coefficients are generally achieved when multiple trained raters are used to score performances (Fulcher, 2003, p.142). This points to the importance of training the raters and the use of at least 2 raters in any speaking test in order to avoid possible reliability problems.

To answer the second sub-section of the second research question, the correlation coefficients were computed as well. As a result of these correlations calculated to estimate the intra-rater reliability of the exam, it was found out that the speaking exam has satisfactory intra-rater reliability as the correlations of 4 of the raters are higher than .70 (see Tables 16, 17 and 18). Only two of the raters' correlations (the third and the sixth rater) between their first and second ratings are low (0.560 and 0.582) although they are also significant as p-values are satisfactory. Two of the raters' grading may have been lower than the others because the second assessment of the performances of the students were made by watching and listening to the recordings made during the first implementation of the exam. As pointed out by Nancy (1980), ratings made on the spot may be somewhat different than ratings made from recordings as there may be a tendency to be more attracted to the enthusiasm and presence of students when rating on the spot than when rating recordings (p.17). Since the physical conditions were not the same, the results could have been affected by this. However, in a study reported by Shohamy (1983, as cited in Fulcher 2003, p. 141), high intra-rater reliability level (0.99) was found although the performances of 32 students were rerated by using the recorded tapes of interviews, which may possibly indicate that there may also be some other reasons behind the low correlations found for some of the raters.

As mentioned earlier, this research in a way draws our attention to the ongoing conflict between reliability and validity as the relationship between two is difficult to understand (Alderson, et al., 1995, p. 187). It is obvious that reliability is a pre-requisite for validity. Therefore, the problems related to reliability may influence the validity of the exam, which may also be the case in this study. For example, in the predictive validity part of the study, low correlations may have emerged as the true scores are not known. The observed scores used to compute correlations may have been affected by the unreliability of the tests (William, 2000, p.4). Similarly, the low correlation between the speaking scores and the total TOEFL test scores may have emerged due to the scoring. However, these do not mean that the exam investigated should not be used. All these indicate that the tension between these two complex terms should be paid enough attention since it is possible for a test to be reliable but invalid as well. To eliminate the problems and to enhance the validity and reliability of the exam, special procedures can be applied if the results obtained from the test need to be justified. Nevertheless, as Underhill (1982) points out:

> The main problem…may be stated simply: high reliability and high validity are seemingly incompatible…If you believe real language occurs in creative communication between two or more parties with genuine reasons for communicating, then you may accept that the trade-off between reliability and validity is unavoidable (p.17).

Due to this, the primary concern of the test writers or the institutions trying to validate their examinations should be to increase the quality of their tests as much as possible by taking the issues of reliability and validity into consideration.

5.3 Assessment of the Study

This research study focused on the face, content, predictive, construct validity and rater reliability of the speaking exam implemented at TOBB University of Economics and Technology, Department of Foreign Languages. Therefore, the findings of this study cannot be generalized to other institutions or departments executing a similar exam. However, the methods and the procedures used in this study may serve as a model for other similar contexts.

This present study can be improved in several ways. Firstly, the number of participants can be increased. For example, all preparatory class students taking this exam can be involved in this study. Moreover, more raters participating in the study as then it would make it easier to generalize the results. More raters may have brought further insights to the results investigated in the study.

Secondly, the face validity and the content validity of the test could be assessed by asking the opinions of the instructors working at the department as well.

Finally, the limitations of this study should be considered in order to improve this study. Some of the limitations that need to be stated are as follows:

1. The face validity questionnaire was given to 70 subjects who were C level students in 2007-2008 academic year. The TOEFL-ITP and speaking exam scores of the same group of students were used to determine the construct, predictive validity and to estimate the rater reliability of the exam. This group was mainly chosen to be able to examine the predictive validity of the exam as well since the students passing the December TOEFL-ITP exam were able to have 101 and 102 Departmental English courses which include departmental speaking assessment. However, including all the preparatory class students in the investigation of face validity part of the study could pave the way for a better understanding of the results. In order to investigate the perceptions of more students, the FVQ can be conducted after the July TOEFL-ITP exam as well.

2. Intra-rater reliability of the exam was investigated with the help of 6 raters. Not all the raters who took part in the December TOEFL-ITP speaking exam were required to assess the performances of the same students that they graded before. The raters taking part in this study were chosen on a voluntary basis. However, if all the raters participated, the results could be easier to interpret and generalize. This could not have been implemented due to the intense schedule and work load of the instructors working at the department.

3. The content validity of the exam was determined by holding interviews with "experts" as suggested in the literature. The comments were made based on the ideas and perceptions of

informants. Nevertheless, a questionnaire could also have been given to the other instructors employed at the department of validity could also be helpful for the enhancement of the speaking exam.

5.4 Recommendations

As a result of the study conducted, the following recommendations are made as to the speaking exam and its implementation.

1. The scorer reliability analysis of the speaking exam indicated that there are some undesirable differences between the raters in terms of grading. Although the obtained significant statistical results were rather satisfactory, it would be better if the institution held more sessions on standardization so that all instructors, especially newly hired ones, could benefit from them. The differences between the raters may be reduced in this way as all the instructors can have the opportunity to understand the procedures and the scoring of the exam before the implementation.

2. The speaking topics included in the last stage of the speaking exam can be revised as some students reported that they cannot speak about the topic they choose even in their native language. Topics should be reviewed both in quality and quantity. This sheds light on the issue that the topics chosen should sometimes be modified according to the profile of the students since it is their speaking ability which is tested, not their world knowledge.

3. As a result of the predictive validity analysis, it was seen that there is no significant relationship between the preparatory speaking exam and the departmental speaking exam grades

although they claim to test the same constructs. Further analysis should be done about this issue. The contents and the grading procedures of the two speaking exams can be examined in a detailed way to understand the reasons for the obtained results.

5.5 Implications for Further Research

Douglas (2000) reports a similarity between the validation process and a "mosaic":

> Validation is not a once-for-all event but rather a dynamic process in which many different types of evidence are gathered and presented in much the same way as a mosaic is constructed… is a mosaic never to be completed, as more and more evidence is brought to bear in helping us interpret performances on our tests and as changes occur in the process of testing, the abilities to be assessed, the contexts of testing and generalizations test developers want to make (p. 258)

Therefore, further research can be done to investigate other aspects of the speaking exam or different types of validity and reliability as well. This research focused on only some types of validity and reliability of the speaking exam at TOBB University of Economics and Technology, Department of Foreign Languages. It is seen that the institution seems to provide a feasible way of assessing speaking skills while still maintaining requirements of reliability and validity.

This study may also be helpful for teachers and testers who are interested in testing speaking since it investigates the validity and the reliability of the speaking exam implemented in preparatory education at TOBB University of Economics and Technology, Department of Foreign Languages. Moreover, this study can be a model for other validation studies. The teachers, testers and administration at TOBB University of Economics and Technology, Department of Foreign Languages can benefit from the investigation of the speaking exam with respect to teaching, learning and testing.

REFERENCES

Alderson, J. C., Claphamn, C.,& Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, J. D. (1996). *Testing in language programs*. Upper Saadle River, NJ, USA: Prentice Hall Regents.

Brown, J. D. (2005). *Testing in Language Programs*: New York: McGraw-Hill.

Brown, A., Iwashita, N., Mc Namara, T., and O' Hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics, 29* (1), 24-49.

Cardoso, R. M. F. (1998). Authentic foreign language testing in a Brazillian university entrance exam. (ERIC Document Reproduction Service No. ED423675).

Chaudhary, S. (1997). Testing spoken English as a second language. *English Teaching Forum, 35* (2), 22-25.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston: Heinle & Heinle Publishers.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D.E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing,* 21 (2), 107-145. Retrieved February 2, 2008, from http://ltj.sagepub.com/cgi/content/refs/21/2/107.

Dai, Y., He, L. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing,* 23 (3), 370-401. Retrieved February 2, 2008, http://ltj.sagepub.com/cgi/content/abstract/23/3/370

Davies, A., Brown, A., Elder, C., Hill, Kathryn. , Lumley, T., & McNamara, T.(1999). *Dictionary of Language Testing.* Cambridge: Cambridge University Press.

Dooey, P. (1999). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. Retrieved on February, 2002 from the following World Wide Web: http://lsn.curtin.edu.au/tlf/tlf1999/dooey.html

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Dörnyei, Z. (2003). Questionnaires in second language research: *Construction, administration and processing*. Manwah, NJ: Lawrence Erbaum Associates, Inc.

Ebel, R. L., Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Gronlund, N., & Linn, R. L. (1990). *Measurement and evaluation in teaching.* New York: Macmillan Publishing Company.

Grounlound, N. E. (1998). *Assessment of student achievement*. London: Allyn and Bacon.

Halleck, G. B. (1996). Interrater reliability of the OPI : Using academic trainee raters. *Foreign Language Annals, 29* (2), 223-238.

Harris, D. P. (1969). *Testing English as a second language:* New York: Mc Graw-Hill Book Company.

Ferguson, N. (1998). Comprehension and production of the spoken language. *IRAL* (36), 307-322.

Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Longman Education.

Foot, M.C. (1999). Relaxing in pairs. *EFL Journal, 53* (1), 36-41.

Henning,G.(1987). *Guide to language testing.* Cambridge: NewBury House Publishers.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, A. (1990). *Testing for language teachers*. Glasgow: Cambridge University Press.

Jafarpur, A. (1988). Non-native raters determining the oral proficiency of EFL learners. *System*, 16 (1), 61-8.
Kitao, S.K., Kitao, K. (1996). Testing Speaking. (ERIC Document Reproduction Service No. ED 398261).

Kumar, R. (1996). *Research methodology*. London: Sage Publications.

Lado, R. (1961). Language Testing: *The Construction and Use of Foreign Language Tests: A Teacher's Book.* New York: McGrow-Hill Book Company.

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

Madsen, H.S. (1983). *Techniques in testing.* Oxford: Oxford University Press.

Manola, J.R., & Wolfe, E. W. (2000). The impact of composition medium on essay raters in foreing language testing. (ERIC Document Reproduction Service No. ED443836).

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

Moritoshi, T.P. (2002). *Validation of the Test of English Conversation Proficiency*. Master's thesis, University of Birmingham, Birmingham.

Nakamura, Yuji. (1997). Establishing construct validity of an English speaking test. *Journal of Communication, n6, 13-30.*

Norris, J.M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum, 38* (1), 18-22.

Nunan,D.(2002). *Research methods in language learning.* Cambridge: Cambridge University Press.

O'Sullivan, B., Weir. C.J., Saville, N. (2002).Using Observation Checklists to Validate Speaking Tasks. *Language Testing,* 19 (1), 33-56. Retrieved October 30, 2007, from http://ltj.sagepub.com/cgi/content/abstract/19/1/33

Ösken, H. (1999). *An assessment of the validity of the midterm and the end of course assessment tests administered at Hacettepe University Department of Basic English*. Unpublished master's thesis, Bilkent University, Ankara.

Perren, G.E. (1968). Testing spoken language: some unsolved problems. In Davies, Alan (Ed.), *Language Testing Symposium: a psycholinguistic approach*. (pp. 107-132). London: Oxford University Press.

Salabary, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing,* 17 (3), 289-310. Retrieved February 2, 2008, from http://ltj.sagepub.com/cgi/content/abstract/17/3/289

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing,* 24 (3), 355-390. Retrieved February 2, 2008, from http://ltj.sagepub.com/cgi/content/abstract/24/3/355

Serpil, H. (2000). *An assessment of the content validity of the midterm achievement tests administered at Anadolu University Foreign Languages Department.* Unpublished masters thesis, Bilkent University, Ankara.

Underhill, N.(1987). *Testing spoken language.* Cambridge: Cambridge University Press.

Ur,P.(1996). *A Course in language teaching.* Cambridge: Cambridge University Press.

Walter D., Nancy H. (1971). *Topics in measurement: Reliability and Validity*: New York: Mc Graw-Hill.

Weir, C. J. (2005). *Language testing and validation.* New York: Palgrave Macmillan

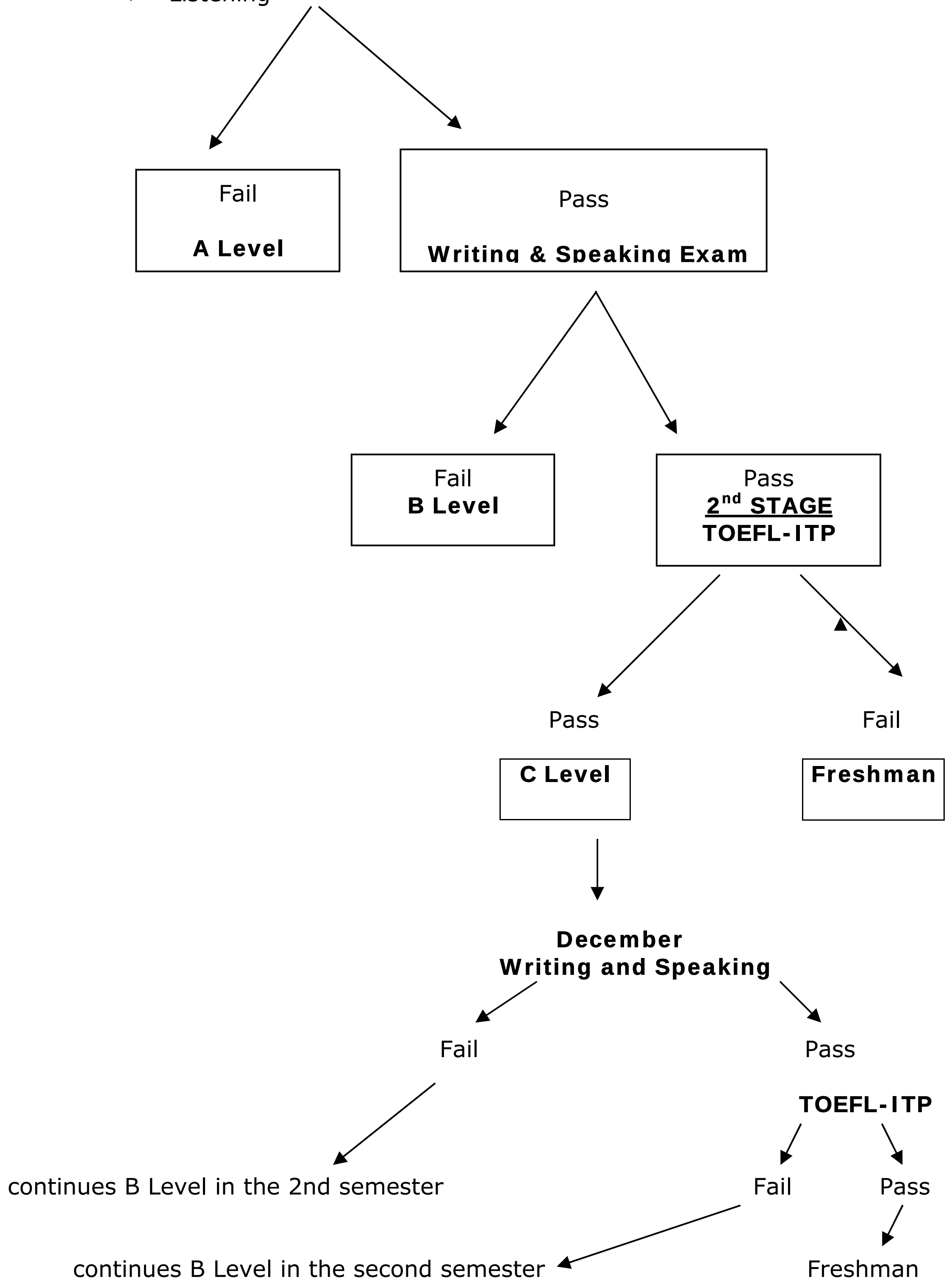
Weir, C.J. (1990). *Communicative language testing*. New York: Prentice Hall.

William, D. (2000). *Reliability, validity and all that jazz*. Retrieved August 15, 2008 from http://www.aaia.org.uk/pdf/2001DYLANPAPER3.PDF

APPENDIX A

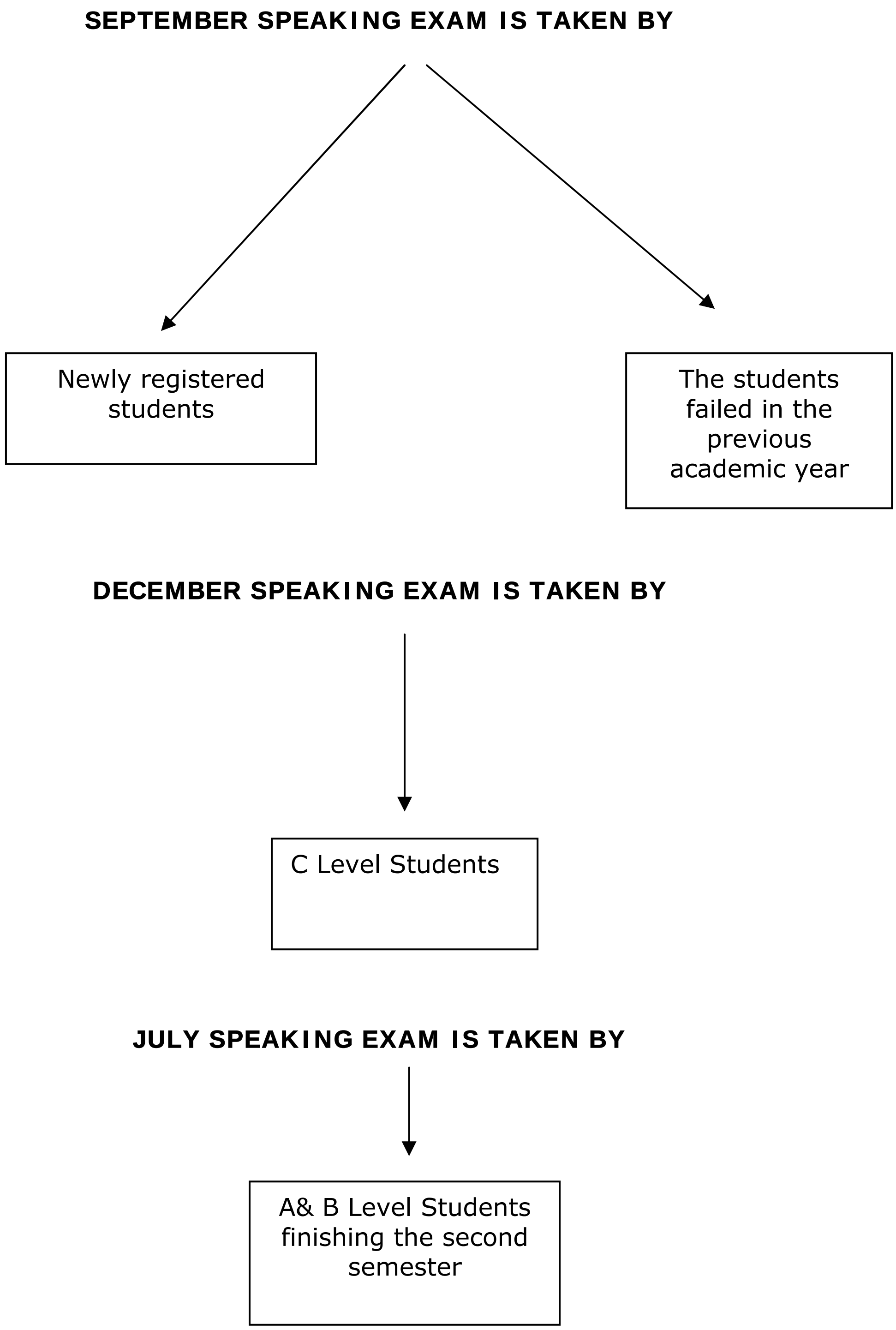
Preliminary Qualifying Exam

- Grammar and Vocabulary
- Reading
- Listening



APPENDIX B

SPEAKING EXAM SCHEDULE



SEPTEMBER SPEAKING EXAM IS TAKEN BY

Newly registered students

The students failed in the previous academic year

DECEMBER SPEAKING EXAM IS TAKEN BY

C Level Students

JULY SPEAKING EXAM IS TAKEN BY

A& B Level Students finishing the second semester

APPENDIX C

THE SPEAKING EXAM

PART I
Warm-up/ Personal Questions (3 minutes)

In the first part of the exam you will ask general questions about everyday life.

- ✓ The interaction is between the instructor and the student. There will be 2 or 3 instructors but only one will speak with a student. The others will take notes.
- ✓ Instructors are not permitted to explain or reword the questions. If the students cannot understand the question(s), the instructors can ONLY repeat the question(s).
- ✓ Students are expected to give answers of a minimum 15 seconds. One word answers are not acceptable.

Sample Questions:
Where are you from?
Who are the most important people in your life?
How close to your school do you live?

PART II
Picture Description and Question Related to the Picture (3minutes)

In the second part of the exam the students will be asked to descirbe a picture.

This section is sub-divided into three sections:
A) Picture Description (1 minute)

The student will choose a picture from a variety of pictures. S/he will be asked to deserib/ talk about the picture.

B) Interpretative Questions (Questions related to the picture-1 minute)

What makes you believe ... ?
Why do you think ... ?

C) Personal Questions related to the Picture's Main Topic. (1minute)

For example, if there' s a picture of people cooking: Do you like cooking?

***** If the student talks about these sections without being asked, there's no need to interrupt him/her with these kind of questions.

PART III

Expressing Opinions (3 minutes)

In the third part of the exam the students will be asked to speak on their own.

- ✓ The student will pick one topic card from the envelope.
- ✓ The students will have one minute to prepare brief notes before they speak.
- ✓ The students are expected to express their personal opinion on the topic.

Tips to Keep in Mind

- ✓ For part 3, although the notess are for the students' own use only they will be collected AT THE END of the test.
- ✓ For part 3, collect the questions AT THE END of the test.
- ✓ STOP the student if he/she goes over time while answering the questions.
- ✓ Do NOT explain the questions or unknown vocabulary to the students. Do not paraphrase the questions. Only repeat the questions.
- ✓ Ease the students with a calm and cheerful face ☺
- ✓ Do NOT exceed the time limit.
- ✓ Mark the students INDIVIDUALLY and calculate the average grade. If the gap between the markers is more than 20 points, the interlocutors are supposed to grade the students again after revising the performance of the student.

STAGE 1- SAMPLE QUESTION

Good morning / afternoon / evening. My name is................ and this is my colleague .... And your name is?

Select a further question for the candidate.

1. What kind of journey did you have to get here today?
2. Do you live with your friends or family?
3. What do you like about the area you live in?
4. What do you do?
5. Do you live in this area?
6. What do you like best about your city/town/village?
7. When did you start leaming English?
8. Do you study any other languages apart from English?
9. How old are you?
10. Which football team do you support?
11. What are your hobbies?
12. Which book did you fast read? What was it about? Did you like it?
13. Which film did you fast see? What was ft about? Did you like it?
14. What kind offilms do you like most?
15. Who is your favorite actor/actress? Why?
16. How many members are there \h your family? Can you describe them?
17. What is your favorite food? Can you cook it?
18. Do you Iike going to parties? When did you last go to a party? How was it?
19. How many hours a day do you watch TV?

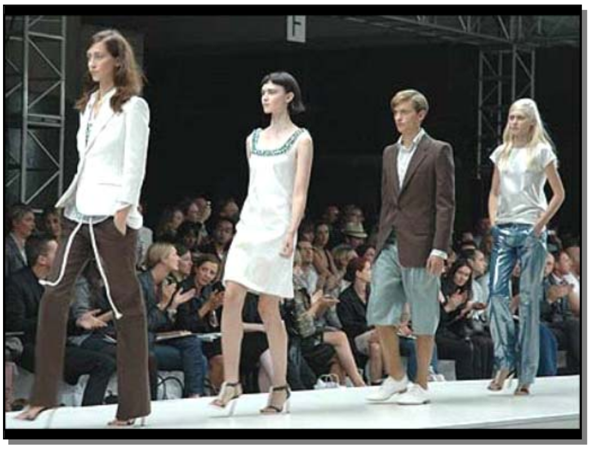
20. What is your favorite program on TV?
21. Do you like shopping? Where do you usually go shopping?
22. Do you like gossiping?
23. Are you a jealous person? When do you act in a jealous way?
24. What do you usually do after school everyday?
25. Do you like summer holiday or winter holiday? Why?
26. Do you like studying?
27. Are you a good student? Why / Why not?
28. Is there any particular person who helped you learn English?
29. Could you please tell us something about the kind of things you read for pleasure?
30. What do you do when you are not working or studying?
31. Could you please tell us about your future plans?
32. What about your early schoolife? What were they like?
33. How ambitious are you?
34. Are you a competitive person?
35. What would you change if you were the Prime Minister of Turkey?
36. How easy or difficult is it nowadays for young people to find a job they really want to do?
37. What would you say has been the most enjoyable period of your life so far?
38. Who are the most important people in your life?
39. How would you describe as a real friend?
40. How do you find out what is happening in the world?
41. What are some of your bad habits?
42. Who is the most attractive in your family?
43. What do you like about the area you live in?

44. What kinds of films do you like most? Why?
45. Who is your favorite actor/actress? Why?
46. What is your favorite food? Can you cook it?
47. Do you like summer holidays or winter holidays? Why?
48. Who are the most important people in your life?
49. What are your hobbies?
50. Could you please tell us about your future plans?

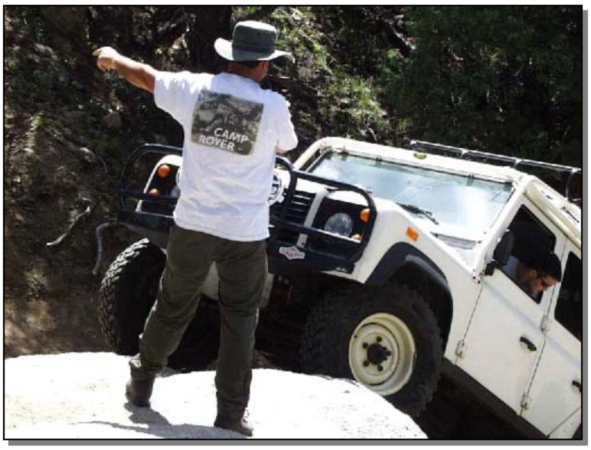
Stage 2

Show all the slides to the candidate and ask them to speak about the one they choose.

Slide 1



Slide 2



Slide 3

Slide 4

Slide 5

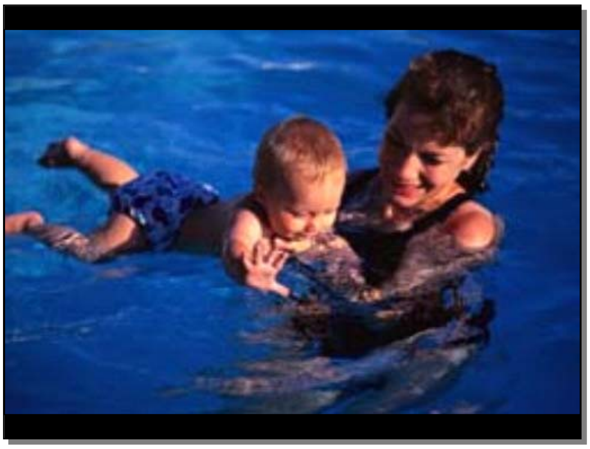
Slide 6

Slide 7

Slide 8

Slide 9

Slide 10



Slide 11

Slide 12

Slide 13

Slide 14

Slide 15

Slide 16

Slide 17

Slide 18

Slide 19

Slide 20

Slide 21

Slide 22

Slide 23

Slide 24

Slide 25

Slide 26

Slide 27

Slide 28

Slide 29

Slide 30

Slide 31

Slide 32

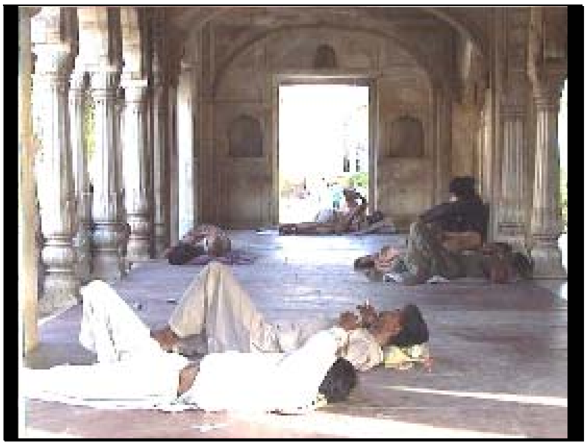
Slide 33

Slide 34

Slide 35

Slide 36

Slide 37



Slide 38

Slide 39

Slide 40

Slide 41

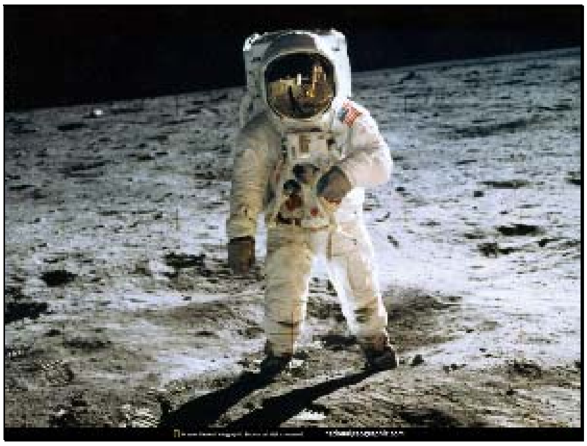
Slide 42

Slide 43

Slide 44

Slide 45

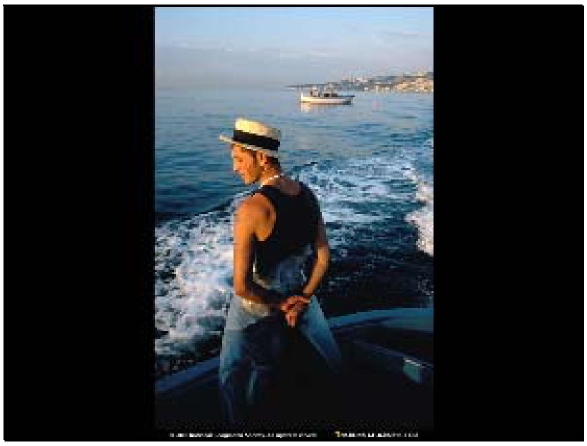
Slide 46



Slide 47

Slide 48

Slide 49



Slide 50

Slide 51

Slide 52

Slide 53

Slide 54

Slide 55

Slide 56

Slide 57

Slide 58

Slide 59

Slide 60

Slide 61

Slide 62

Slide 63

Slide 64

Slide 65

Stage 3

Ask the candidate to choose a question from the envelope. Give them same time to take note. about il The questions are as follows:

SPEAKING TOPICS
1. People attend college or university for many different reasons (for example, new experiences, career preparation, increased knowledge). Why do you think people attend college or university? State your opinion.
2. Do you agree or disagree with the following statement? Parents are the best teachers. State specific reasons.
3. Nowadays, food has become easier to prepare. Has this change improved the way people live? Use specific reasons and examples to support your opinion.
4. It has been said, "Not everything that is learned is contained in books." Compare and contrast knowledge gained from experience with knowledge gained from books. In your opinion, which source is more important? Why?
5. A company has announced that it wishes to build a large factory near your community. Discuss the advantages and disadvantages of this new influence on your community. Do you support or oppose the factory? Explain your position.
6. If you could change one important thing about your hometown, what would you change? Use reasons and specific examples to support your answer.
7. How do movies or television influence people's behavior? Use reasons and specific examples to support your answer.
8. Do you agree or disagree with the following statement? Television has destroyed communication among friends and family. Use specific reasons and examples to support your opinion.

9. Some people prefer to live in a small town. Others prefer to live in a big city. Which place would you prefer to live in? Use specific reasons and details to state your opinion.

10. "When people succeed, it is because of hard work. Luck has nothing to do with success." Do you agree or disagree with the quotation above? Use specific reasons and examples to explain your position.

11. Do you agree or disagree with the following statement? Universities should give the same amount of money to their students' sports activities as they give to their university libraries. Use specific reasons and examples to support your opinion.

12. Many people visit museums when they travel to new places. Why do you think people visit museums? Use specific reasons and examples to support your answer.

13. Some people prefer to eat at food stands or restaurants. Other people prefer to prepare and eat food at home. Which do you prefer? Use specific reasons and examples to support your answer.

14. Some people believe that university students should be required to attend classes. Others believe that going to classes should be optional for students. Which point of view do you agree with? Use specific reasons and details to explain your answer.

15. Neighbors are the people who live near us. In your opinion, what are the qualities of a good neighbor? Use specific details and examples in your answer.

16. It has recently been announced that a new restaurant may be built in your neighborhood. Do you support or oppose this plan? Why? Use specific reasons and details to support your answer.

17. Some people think that they can learn better by themselves than with a teacher. Others think that it is always better to have a teacher. Which do you prefer?

18. What are some important qualities of a good supervisor (boss)? Use specific details and examples to explain why these qualities are important.

19. Should governments spend more money on improving roads and highways, or should governments spend more money on improving public transportation (buses, trains, subways)? Why?

20. It is better for children to grow up in the countryside than in a big city. Do you agree or disagree?

21. In general, people are living longer now. Discuss the causes of this phenomenon.

22. In some countries, teenagers have jobs while they are still students. Do you think this is a good idea? Support your opinion by using specific reasons and details.

23. A person you know is planning to move to your town or city. What do you think this person would like and dislike about living in your town or city? Why?

25. It has recently been announced that a new movie theater may be built in your neighborhood. Do you support or oppose this plan? Why?

26. Do you agree or disagree with the following statement? People should sometimes do things that they do **not** enjoy doing.

27. Do you agree or disagree with the following statement? Television, newspapers, magazines, and other media pay too much attention to the personal lives of famous people such as public figures and celebrities.

28. Some people believe that the Earth is being harmed (damaged) by human activity. Others feel that human activity makes the Earth a better place to live. What is your opinion?.

29. It has recently been announced that a new high school may be built in your community. Do you support or oppose this plan? Why?

30. Some people spend their entire lives in one place. Others move a number of times throughout their lives, looking for a better job, house, community, or even climate. Which do you prefer: staying in one place or moving in search of another place?

31. Is it better to enjoy your money when you earn it or is it better to save your money for some time in the future?

32.. Businesses should hire employees for their entire lives. Do you agree or disagree?

33. Do you agree or disagree with the following statement? Attending a live performance (for example, a play, concert, or sporting event) is more enjoyable than watching the same event on television.

34. Choose **one** of the following transportation vehicles and explain why you think it has changed people's lives - a) automobiles b) bicycles c) airplanes

35. Do you agree or disagree that progress is always good?

36. Learning about the past has no value for those of us living in the present. Do you agree or disagree?

37. Do you agree or disagree with the following statement? With the help of technology, students nowadays can learn more information and learn it more quickly.

38. The expression "Never, never give up" means to keep trying and never stop working for your goals. Do you agree or disagree with this statement?

39. Some people think that human needs for farmland, housing, and industry are more important than saving land for endangered animals. Do you agree or disagree with this point of view? Why or why not?

40. What is a very important skill a person should learn in order to be successful in the world today? Choose **one** skill and use specific reasons and examples to support your choice.

41. Why do you think some people are attracted to dangerous sports or other dangerous activities? Use specific reasons and examples to support your answer.

42. Some people like to travel with a companion. Other people prefer to travel alone. Which do you prefer? Use specific reasons and examples to support your choice.

43. Some people prefer to get up early in the morning and start the day's work. Others prefer to get up later in the day and work until late at night. Which do you prefer? Use specific reasons and examples to support your choice.

44. What are the important qualities of a good son or daughter? Have these qualities changed or remained the same over time in your culture? Use specific reasons and examples to support your answer. Some people prefer to work for a large company. Others prefer to work for a small company. Which would you prefer? Use specific reasons and details to support your choice.

45. People work because they need money to live. What are some **other** reasons that people work? Discuss one or more of these reasons. Use specific examples and details to support your answer.

46. Do you agree or disagree with the following statement? Face-to-face communication is better than other types of communication, such as letters, email, or telephone calls. Use specific reasons and details to support your answer.

47. Some people like to do only what they already do well. Other people prefer to try new things and take risks. Which do you prefer? Use specific reasons and examples to support your choice.

48. Some people believe that success in life comes from taking risks or chances. Others believe that success results from careful planning. In your opinion, what does success come from? Use specific reasons and examples to support your answer.

49. What change would make your hometown more appealing to people your age? Use specific reasons and examples to support your opinion.

50. Do you agree or disagree with the following statement? The most important aspect of a job is the money a person earns. Use specific reasons and examples to support your answer.

51. Do you agree or disagree with the following statement? One should never judge a person by external appearances. Use specific reasons and details to support your answer.

52. Do you agree or disagree with the following statement? A person should never make an important decision alone. Use specific reasons and examples to support your answer.

53. A company is going to give some money either to support the arts or to protect the environment. Which do you think the company should choose? Use specific reasons and examples to support your answer.

54. Some movies are serious, designed to make the audience think. Other movies are designed primarily to amuse and entertain. Which type of movie do you prefer? Use specific reasons and examples to support your answer.

55. Do you agree or disagree with the following statement? Businesses should do anything they can to make a profit. Use specific reasons and examples to support your position.

56. Some people are always in a hurry to go places and get things done. Other people prefer to take their time and live life at a slower pace. Which do you prefer? Use specific reasons and examples to support your answer.

57. Do you agree or disagree with the following statement? Games are as important for adults as they are for children. Use specific reasons and examples to support your answer.

58. Do you agree or disagree with the following statement? Parents or other adult relatives should make important decisions for their older (15 to 18 year-old) teenage children. Use specific reasons and examples to support your opinion.

59. What do you want **most** in a friend - someone who is intelligent, or someone who has a sense of humor, or someone who is reliable? Which **one** of these characteristics is most important to you? Use reasons and specific examples to explain your choice.

60. Do you agree or disagree with the following statement? Most experiences in our lives that seemed difficult at the time become valuable lessons for the future. Use reasons and specific examples to support your answer.

61. Some people prefer to work for themselves or own a business. Others prefer to work for an employer. Would you rather be selfemployed, work for someone else, or own a business? Use specific reasons to explain your choice.

62. Should a city try to preserve its old, historic buildings or destroy them and replace them with modern buildings? Use specific reasons and examples to support your opinion.

63. Do you agree or disagree with the following statement? Classmates are a more important influence than parents on a child's success in school. Use specific reasons and examples to support your answer.

64. If you were an employer, which kind of worker would you prefer to hire: an inexperienced worker at a lower salary or an experienced worker at a higher salary? Use specific reasons and details to support your answer.

65. Many teachers assign homework to students every day. Do you think that daily homework is necessary for students? Use specific reasons and details to support your answer.

67. Some people think that the automobile has improved modern life. Others think that the automobile has caused serious problems. What is your opinion? Use specific reasons and examples to support your answer.

68. Which would you choose: a high-paying job with long hours that would give you little time with family and friends **or** a lower-paying job with shorter hours that would give you more time with family and friends?

69. Do you agree or disagree with the following statement? Grades (marks) encourage students to learn. Use specific reasons and examples to support your opinion.

70. Some people say that computers have made life easier and more convenient. Other people say that computers have made life more complex and stressful. What is your opinion? Use specific reasons and examples to support your answer.

71. Do you agree or disagree with the following statement? The best way to travel is in a group led by a tour guide. Use specific reasons and examples to support your answer.

73. Do you agree or disagree with the following statement? Children should begin learning a foreign language as soon as they start school. Use specific reasons and examples to support your position.

74. Do you agree or disagree with the following statement? Boys and girls should attend separate schools. Use specific reasons and examples to support your answer.

75. Is it more important to be able to work with a group of people on a team or to work independently? Use reasons and specific examples to support your answer.

76. Your city has decided to build a statue or monument to honor a famous person in your country. Who would you choose? Use reasons and specific examples to support your choice.

77. Describe a custom from your country that you would like people from other countries to adopt. Explain your choice, using specific reasons and examples.

78. Do you agree or disagree with the following statement? Technology has made the world a better place to live. Use specific reasons and examples to support your opinion.

79. Do you agree or disagree with the following statement? Advertising can tell you a lot about a country. Use specific reasons and examples to support your answer.

80. Do you agree or disagree with the following statement? Modern technology is creating a single world culture. Use specific reasons and examples to support your opinion.

81. Some people say that the Internet provides people with a lot of valuable information. Others think access to so much information creates problems. Which view do you agree with? Use specific reasons and examples to support your opinion.

82. A foreign visitor has only one day to spend in your country. Where should this visitor go on that day? Why? Use specific reasons and details to support your choice.

83. If you could go back to some time and place in the past, when and where would you go? Why? Use specific reasons and details to support your choice.

84. What discovery in the last 100 years has been most beneficial for people in your country? Use specific reasons and examples to support your choice.

85. Do you agree or disagree with the following statement? Telephones and email have made communication between people less personal. Use specific reasons and examples to support your opinion.

86. If you could travel back in time to meet a famous person from history, what person would you like to meet? Use specific reasons and examples to support your choice.

87. If you could meet a famous entertainer or athlete, who would that be, and why? Use specific reasons and examples to support your choice.

88. If you could ask a famous person **one** question, what would you ask? Why? Use specific reasons and details to support your answer.

89. Some people prefer to live in places that have the same weather or climate all year long. Others like to live in areas where the weather changes several times a year. Which do you prefer? Use specific reasons and examples to support your choice.

90. Many students have to live with roommates while going to school or university. What are some of the important qualities of a good roommate? Use specific reasons and examples to explain why these qualities are important.

91. Do you agree or disagree with the following statement? Dancing plays an important role in a culture. Use specific reasons and examples to support your answer.

92. Some people think governments should spend as much money as possible exploring outer space (for example, traveling to the Moon and to other planets). Other people disagree and think governments should spend this money for our basic needs on Earth. Which of these two opinions do you agree with? Use specific reasons and details to support your answer.

93. People have different ways of escaping the stress and difficulties of modern life. Some read; some exercise; others work in their gardens. What do you think are the best ways of reducing stress? Use specific details and examples in your answer.

94. Do you agree or disagree with the following statement? Teachers should be paid according to how much their students learn. Give specific reasons and examples to support your opinion.

95. If you were asked to send one thing representing your country to an international exhibition, what would you choose? Why? Use specific reasons and details to explain your choice.

96. You have been told that dormitory rooms at your university must be shared by two students. Would you rather have the university assign a student to share a room with you, or would you rather choose your own roommate? Use specific reasons and details to explain your answer.

97. Some people think that governments should spend as much money as possible on developing or buying computer technology. Other people disagree and think that this money should be spent on more basic needs. Which one of these opinions do you agree with? Use specific reasons and details to support your answer.

98. Some people like doing work by hand. Others prefer using machines. Which do you prefer? Use specific reasons and examples to support your answer.

99. Schools should ask students to evaluate their teachers. Do you agree or disagree? Use specific reasons and examples to support your answer.

100. In your opinion, what is the most important characteristic (for example, honesty, intelligence, a sense of humor) that a person can have to be successful in life? Use specific reasons and examples from your experience to discuss the topic.

101. It is generally agreed that society benefits from the work of its members. Compare the contributions of artists to society with the contributions of scientists to society. Which type of contribution do you think is valued more by your society? Give specific reasons to support your answer.

102. Students at universities often have a choice of places to live. They may choose to live in university dormitories, or they may choose to live in apartments in the community. Compare the advantages of living in university housing with the advantages of living in an apartment in the community. Where would you prefer to live? Give reasons for your preference.

103. You need to travel from your home to a place 40 miles (64 kilometers) away. Compare the different kinds of transportation you could use. Tell which method of travel you would choose. Give specific reasons for your choice.

104. Some people believe that a college or university education should be available to all students. Others believe that higher education should be available only to good

students. Discuss these views. Which view do you agree with? Explain why.

105. Some people believe that the best way of learning about life is by listening to the advice of family and friends. Other people believe that the best way of learning about life is through personal experience. Compare the advantages of these two different ways of learning about life. Which do you think is preferable? Use specific examples to support your preference.

106. When people move to another country, some of them decide to follow the customs of the new country. Others prefer to keep their own customs. Compare these two choices. Which one do you prefer? Support your answer with specific details.

107. Some people prefer to spend most of their time alone. Others like to be with friends most of the time. Do you prefer to spend your time alone or with friends? Use specific reasons to support your answer.

108. Some people prefer to spend time with one or two close friends. Others choose to spend time with a large number of friends. Compare the advantages of each choice. Which of these two ways of spending time do you prefer? Use specific reasons to support your answer.

109. Some people think that children should begin their formal education at a very early age and should spend most of their time on school studies. Others believe that young children should spend most of their time playing. Compare these two views. Which view do you agree with? Why?

110. The government has announced that it plans to build a new university. Some people think that **your** community would be a good place to locate the university. Compare the advantages and disadvantages of establishing a new university in your community. Use specific details in your discussion.

111. Some people think that the family is the most important influence on young adults. Other people think that friends are the most important influence on young adults. Which view do you agree with? Use examples to support your position.

112. Some people prefer to plan activities for their free time very carefully. Others choose not to make any plans at all for their free time. Compare the benefits of planning free-time activities with the benefits of not making plans. Which do you prefer - planning or not planning for your leisure time? Use specific reasons and examples to explain your choice.

113. People learn in different ways. Some people learn by doing things; other people learn by reading about things; others learn by listening to people talk about things. Which of these methods of learning is best for you? Use specific examples to support your choice.

114. Some people choose friends who are different from themselves. Others choose friends who are similar to themselves. Compare the advantages of having friends who are different from you with the advantages of having friends who are similar to you. Which kind of friend do you prefer for yourself? Why?

115. Some people enjoy change, and they look forward to new experiences. Others like their lives to stay the same, and they do not change their usual habits. Compare these two approaches to life. Which approach do you prefer? Explain why.

116. Do you agree or disagree with the following statement? People behave differently when they wear different clothes. Do you agree that different clothes influence the way people behave? Use specific examples to support your answer.

117. Decisions can be made quickly, or they can be made after careful thought. Do you agree or disagree with the following statement? The decisions that people make quickly are always wrong. Use reasons and specific examples to support your opinion.

118. Some people trust their first impressions about a person's character because they believe these judgments are generally correct. Other people do not judge a person's character quickly because they believe first impressions are often wrong. Compare these two attitudes. Which attitude do you agree with? Support your choice with specific examples.

119. Do you agree or disagree with the following statement? People are never satisfied with what they have; they always want something more or something different. Use specific reasons to support your answer.

120. Do you agree or disagree with the following statement? People should read **only** those books that are about real events, real people, and established facts. Use specific reasons and details to support your opinion.

121. Do you agree or disagree with the following statement? It is more important for students to study history and literature than it is for them to study science and mathematics. Use specific reasons and examples to support your opinion.

122. Do you agree or disagree with the following statement? All students should be required to study art and music in secondary school. Use specific reasons to support your answer.

123. Do you agree or disagree with the following statement? There is nothing that young people can teach older people. Use specific reasons and examples to support your position.

125. Some people say that physical exercise should be a required part of every school day. Other people believe that students should spend the whole school day on academic studies. Which opinion do you agree with?

126. A university plans to develop a new research center in your country. Some people want a center for business research. Other people want a center for research in agriculture (farming). Which of these two kinds of research centers do you recommend for your country? Use specific reasons in your recommendation.

127. Some young children spend a great amount of their time practicing sports. Discuss the advantages and disadvantages of this. Use specific reasons and examples to support your answer.

128. Do you agree or disagree with the following statement? **Only** people who earn a lot of money are successful. Use specific reasons and examples to support your answer.

129. If you could invent something **new**, what product would you develop? Use specific details to explain why this invention is needed.

130. Do you agree or disagree with the following statement? A person's childhood years (the time from birth to twelve years of age) are the most important years of a person's life. Use specific reasons and examples to support your answer.

131. Do you agree or disagree with the following statement? Children should be required to help with household tasks as soon as they are able to do so. Use specific reasons and examples to support your answer.

132. Some high schools require all students to wear school uniforms. Other high schools permit students to decide what to wear to school. Which of these two school policies do you think is better? Use specific reasons and examples to support your opinion.

133. Do you agree or disagree with the following statement? Playing a game is fun only when you win. Use specific reasons and examples to support your answer.

134. Do you agree or disagree with the following statement? High schools should allow students to study the courses that students want to study. Use specific reasons and examples.

135. Do you agree or disagree with the following statement? It is better to be a member of a group than to be the leader of a group. Use specific reasons and examples to support your answer.

136. What do you consider to be the most important room in a house? Why is this room more important to you than any other room? Use specific reasons and examples to support your opinion.

138. If you could make one important change in a school that you attended, what change would you make? Use reasons and specific examples to support your answer.

139. A gift (such as a camera, a soccer ball, or an animal) can contribute to a child's development. What gift would you give to help a child develop? Why? Use reasons and specific examples to support your choice.

140. Some people believe that students should be given one long vacation each year. Others believe that students should have several short vacations throughout the year. Which viewpoint do you agree with? Use specific reasons and examples to support your choice.

141. Would you prefer to live in a traditional house or in a modern apartment building? Use specific reasons and details to support your choice.

142. Some people say that advertising encourages us to buy things we really do not need. Others say that advertisements tell us about new products that may improve our lives. Which viewpoint do you agree with? Use specific reasons and examples to support your answer.

143. Some people prefer to spend their free time outdoors. Other people prefer to spend their leisure time indoors. Would you prefer to be outside or would you prefer to be inside for your leisure activities? Use specific reasons and examples to explain your choice.

145. Do you agree or disagree with the following statement? Playing games teaches us about life. Use specific reasons and examples to support your answer.

146. Imagine that you have received some land to use as you wish. How would you use this land? Use specific details to explain your answer.

147. Do you agree or disagree with the following statement? Watching television is bad for children. Use specific details and examples to support your answer.

148. What is the most important animal in your country? Why is the animal important? Use reasons and specific details to explain your answer.

149. Many parts of the world are losing important natural resources, such as forests, animals, or clean water. Choose **one** resource that is disappearing and explain why it needs to be saved. Use specific reasons and examples to support your opinion.

150. Do you agree or disagree with the following statement? A zoo has no useful purpose. Use specific reasons and examples to explain your answer.

151. In some countries, people are no longer allowed to smoke in many public places and office buildings. Do you think this is a good rule or a bad rule? Use specific reasons and details to support your position.

152. Plants can provide food, shelter, clothing, or medicine. What is one kind of plant that is important to you or the people in your country? Use specific reasons and details to explain your choice.

153. You have the opportunity to visit a foreign country for two weeks. Which country would you like to visit? Use specific reasons and details to explain your choice.

154. In the future, students may have the choice of studying at home by using technology such as computers or television or of studying at traditional schools. Which would you prefer? Use reasons and specific details to explain your choice.

156. The twentieth century saw great change. In your opinion, what is **one** change that should be remembered about the twentieth century? Use specific reasons and details to explain your choice.

157. When people need to complain about a product or poor service, some prefer to complain in writing and others prefer to complain in person. Which way do you prefer? Use specific reasons and examples to support your answer.

158. People remember special gifts or presents that they have received. Why? Use specific reasons and examples to support your answer.

159. Some famous athletes and entertainers earn millions of dollars every year. Do you think these people deserve such high salaries? Use specific reasons and examples to support your opinion.

160. Is the ability to read and write more important today than in the past? Why or why not? Use specific reasons and examples to support your answer.

161. People do many different things to stay healthy. What do you do for good health? Use specific reasons and examples to support your answer.

162. You have decided to give several hours of your time each month to improve the community where you live. What is one thing you will do to improve your community? Why? Use specific reasons and details to explain your choice.

163. People recognize a difference between children and adults. What events (experiences or ceremonies) make a person an adult? Use specific reasons and examples to explain your answer.

164. Your school has enough money to purchase either computers for students or books for the library. Which should your school choose to buy - computers or books? Use specific reasons and examples to support your recommendation.

165. Many students choose to attend schools or universities outside their home countries. Why do some students study abroad? Use specific reasons and details to explain your answer.

166. People listen to music for different reasons and at different times. Why is music important to many people? Use specific reasons and examples to support your choice.

167. Groups or organizations are an important part of some people's lives. Why are groups or organizations important to people? Use specific reasons and examples to explain your answer.

168. Imagine that you are preparing for a trip. You plan to be away from your home for a year. In addition to clothing and personal care items, you can take **one** additional thing. What would you take and why? Use specific reasons and details to support your choice.

169. When students move to a new school, they sometimes face problems. How can schools help these students with their problems? Use specific reasons and examples to explain your answer.

170. It is sometimes said that borrowing money from a friend can harm or damage the friendship. Do you agree? Why or why not? Use reasons and specific examples to explain your answer.

171. Every generation of people is different in important ways. How is your generation different from your parents' generation? Use specific reasons and examples to explain your answer.

172. Some students like classes where teachers lecture (do all of the talking) in class. Other students prefer classes where the students do some of the talking. Which type of class do you prefer? Give specific reasons and details to support your choice.

173. Holidays honor people or events. If you could create a new holiday, what person or event would it honor and how would you want people to celebrate it? Use specific reasons and details to support your answer.

174. A friend of yours has received some money and plans to use all of it either o to go on vacation o to buy a car Your friend has asked you for advice. Compare your friend's two choices and explain which one you think your friend should choose. Use specific reasons and details to support your choice.

175. The 21st century has begun. What changes do you think this new century will bring? Use examples and details in your answer.

176. What are some of the qualities of a good parent? Use specific details and examples to explain your answer.

177. Movies are popular all over the world. Explain why movies are so popular. Use reasons and specific examples to support your answer.

178. In your country, is there more need for land to be left in its natural condition or is there more need for land to be developed for housing and industry? Use specific reasons and examples to support your answer.

179. Many people have a close relationship with their pets. These people treat their birds, cats, or other animals as members of their family. In your opinion, are such relationships good? Why or why not? Use specific reasons and examples to support your answer.

180. Films can tell us a lot about the country where they were made. What have you learned about a country from watching its movies? Use specific examples and details to support your response.

181. Some students prefer to study alone. Others prefer to study with a group of students. Which do you prefer? Use specific reasons and examples to support your answer.

182. You have enough money to purchase either a house or a business. Which would you choose to buy? Give specific reasons to explain your choice.

APPENDIX D

A SAMPLE SPEAKING EXAM

Examiner: So hello how are you Hatice?
Test taker: Fine, thanks and you?
Examiner: Good, thank you very much. Do you prefer Hatice or Hilal?
Test taker: Hilal
Examiner: OK, I will call you Hilal.
Examiner: Hilal, can you please tell a little bit about yourself?
Testkaker: My name is Hatice Hilal and I was born in 1989. Now I am 18 years old. I came here from Afyon and my department is Industrial Engineering.
Examiner: Can you talk about your family please?
Test taker: Thanks to God I have a big family. I have two sisters and a brother. My father is a doctor, my mother is a housewife. My brother's name is Ahmet and he is studying electrical electronics enginnerin at Boğaziçi University. I have two sisters, one of my sisters is 6 years older than me. The other one is tow years elder than me. My elder sister graduated from Istanbul Kultur University, Law Department. She is in Ankara and we are living together. My younger sister is at high school. That's all.
Examiner: How is your relationship with your brothers and sisters?
Test taker: We are getting well.
Examiner: No fights?
Test taker: My younger sister sometimes.
Examiner: What is the basic topic?
Test taker: For example, I have a boyfriend and my father doesn't know it. My younger sister always talks about it.
Examiner: She says I will tell my father, if you...
Examiner: OK, let's go on to the next stage, picture talking. I will scroll down, you please choose the picture you want to describe.
Test taker: 49... There is a man who is sleeping. In front him there is a table. On this there is ...there is a secret... Maybe he has taken some drugs. Maybe he has problems with his wife and maybe he has made a lot mistakes in his life and nowadays there lots of people like that because especially children and adult , mos of the children are taking drug and somebody forces them to do something like that and also cigarettes... Most of the people smoke and ....
Examiner: Do you smoke?
Test taker: No and I hate drugs.
Examiner: Do you know anyone who takes drugs?

Test taker: No, I don’t know.
Examiner: Do you know the results of taking drugs?
Test taker: I think they have a lot of problems in their lives and they take I think.
Examiner: Ok, next stage. You are going to choose a topic. Please tell us the number and you have got one minute and you can take notes.
Test taker: 88
Examiner: OK
Examiner: Are you ready?

Test taker: I ask a famous person, maybe how did you become a famous person? Otherwise what did you do to become a famous person? I don’t know actually I don’t care about famous person.
Examiner: A famous person can also be an important politician or a doctor, or a Nobel prize winner. Try to think in those terms...Not celebrity
Test taker: Actually I am not interested in politic, actually I don’t listen them on TV but only one person I am interested. Can I say his name? Is it a problem?
Examiner: Sure
Test taker: Nihat Genç. I think he has really good opinions.
Examiner: If you had the chance, what would you ask?
Test taker: Actually I cannot ask anything but I listen him. Maybe I can ask about economy and I don’t know.
Examiner: Ok, thank you Hilal.

Name: H.H.N
Grade: 78

APPENDIX D

SCORES	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4 (85-100)	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well paced flow (fluid expression) Speech is clear. It may include minor lapses or minor difficulties with pronunciation or intonation patterns which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate) Some minor (or systematic) errors are noticeable but do not obscure meaning	Response is sustained and sufficient to the task. It is generally well-developed and coherent; relationships between ideas are clear (or clear progression of ideas)
3 (61-84)	The response addresses the tasks appropriately, but may falls short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation or pacing are noticeable and may requires listener effort at time (though overall intelligibility is not significantly affected)	The response demonstrates automatic and effective use of grammar and vocabulary and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or in accurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it doesn't seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at time not be immediately clear.
2 (31-60)	The response addresses the task but the development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation awkward intonation or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition)	The response is connected to the task, though the number of ideas presented or development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support) At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1 (1-30)	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort: delivery is choppy, fragmented, or telegraphic frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit (or prevent) expression of ideas and connections among ideas. Some low-level responses may really heavily on practiced or formulaic expressions.	Limited relevant content is expressed**.** The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of prompts.
0 - Speaker makes no attempt to respond OR response is unrelated to the topic. OR repeatedly uses expressions such as "yes", "no", "I t know" etc.				

APPENDIX E

TOBB ETU SPEAKING EXAM
THE FACE VALIDITY QUESTIONNAIRE

Please check (√) your answers in the boxes. If necessary, please write your comments in Turkish in the spaces provided.
(Lütfen cevaplarınızı kutulara işaretleyiniz. Gerekirse, yorumlarınızı ayrılan boşluklara Türkçe olarak yazınız.)

1. **What do you think is the most accurate way to assess someone's English speaking ability?**
   Birisinin İngilizce konuşma becerisini ölçmenin en doğru yolu sizce nedir?
   Please choose *only one* of the answers below.
   Lütfen aşağıdaki cevaplardan sadece birisini seçiniz.

a. Write a script of a dialogue or talk. □
(Bir diyaloğun veya konuşmanın metnini yazmak.)

b. Read a dialogue or talk,and then answer comprehension questions about it. □
(Bir diyaloğu veya konuşmayı okumak, ve sonra onunla ilgili kavrama soruları cevaplamak.)

c. Listen to a dialogue or talk, and then answer comprehension questions. □
(Bir diyaloğu veya konuşmayı dinlemek, ve sonra onunla ilgili kavrama soruları cevaplamak.)

d. A written test of vocabulary and grammar useful during speaking. □
(Konuşmada yararlı olacak dilbigisi kurallarını ve kelimeyi içeren yazılı bir test.)

e. Speak with a native speaker on a given topic in English. □
(Ana dili İngilizce olan biriyle verilen bir konu üzerinde İngilizce konuşmak.)

f. Speak with a non-native speaker in English on a given topic in English. □
(Ana dili İngilizce olmayan biriyle verilen bir konu üzerinde İngilizce konuşmak.)

g. Another way (please write): ____
(Başka bir yöntem (lütfen yazınız):

h. I am not sure. □
(Emin değilim.)

2. To what extent did the speaking exam you took reflect the characteristics of the spoken language in real life situations?
Girdiğniz konuşma sınavı gerçek hayattaki konuşmaya ne kadar benziyordu?
Please choose *only one* of the answers below.
Lütfen aşağıdaki cevaplardan sadece birisini seçiniz.

a. A lot. (Çok)	□	Go to question 3. (Üçüncü soruya gidiniz.)
b. Quite a lot. (Oldukça.)	□	Go to question 3. (Üçüncü soruya gidiniz.)
c. Average. (Orta.)	□	Go to question 3. (Üçüncü soruya gidiniz.)
d. A little. (Az.)	□	Miss question 3. Go to question 4. (3.soruyu atlayıp, dördüncü soruya geçiniz.)
e. Not at all. (Hiç.)	□	Miss question 3. Go to question 4. (3.soruyu atlayıp, dördüncü soruya geçiniz.)
f. I am not sure. (Emin değilim.)	□	Miss questions 3 and 4 question 5. (3. ve 4. soruyu atlayıp 5.soruya geçiniz.)

3. **Why do you think the speaking exam you took reflect the characteristics of the spoken language in real life situations?**
Sizce girdiğiniz konuşma sınavı neden gerçek hayattaki konuşmaya benziyordu?

Please choose any of the answers below. You can choose more than one answer.
Lütfen aşağıdaki cevaplardan her hangi birisini seçiniz. Birden fazla cevap seçebilirsiniz.
After answering this question, miss question 4 and go to question 5.
Bu soruyu cevapladıktan sonra, dördüncü soruyu atlayıp beşinci soruya geçiniz.

a. It had the parts of a normal dialogue (for example, a start, questions, topic changes and a finish). □

(Normal bir diyalogtaki bölümler vardı.) (örneğin, başlangıç, sorular, konu değişimi ve bitiş)

b. I was able to speak enough. □
Yeterli konuşabildim.

c. I was able to ask questions freely. □
Serbestçe sorular sorabildim.

□

d. I was able to express my ideas and emotions.
Düşüncelerimi ve duygularımı ifade edebildim.

e. The teacher didn't tell me whether my opinion or answer was right or wrong. □
Öğretmen bana düşüncemin veya cevabımın doğru ya da yanlış olduğunu söylemedi.

f. It was mostly spontaneous and I didn’t write a script for what I would say. □
Çoğunlukta hazırlıksızdı ve ne söyleyeceğimle ilgili birşey yazmadım.

g. Other reason(s) (please write): (Diğer neden/nedenler) (Lütfen yazınız)

____

____

i. I am not sure. □
Emin değilim.

4. **Why do you think the speaking exam you took didn’t reflect the characteristics of the spoken language in real life situations?**
Sizce girdiğiniz konuşma sınavı gerçek hayattaki konuşmaya neden benzemiyordu?

Please choose any of the answers below. You can choose more than one answer
Lütfen aşağıdaki cevaplardan her hangi birisini seçiniz. Birden fazla cevap seçebilirsiniz.

a. It didn’t have the parts of a normal dialogue (for example, a start, questions, topic changes and a finish). □
(Normal bir diyalogtaki bölümler yoktu.) (örneğin, başlangıç, sorular, konu değişimi ve bitiş)

b. I could not speak enough. □
(Yeterli konuşamadım.)

c. I could not ask questions freely. □
(Serbestçe sorular soramadım.)

d. I could not express my ideas and emotions. □
(Düşüncelerimi ve duygularımı ifade edemedim.)

f. The teacher told me whether my opinion or answer was right or wrong. □
(Öğretmen bana düşüncemin veya cevabımın doğru ya da yanlış olduğunu söyledi.)

g. It was not spontaneous and I could write a script for what I would say. □
(Hazırlıksız değildi ve ne söyleyeceğimle ilgili bişeyler yazabildim.)

h. Other reason(s) (please write): (Diğer neden/nedenler) (Lütfen yazınız)

____

____

____

i. I am not sure. □
(Emin değilim.)

5. To what extent was it difficult to understand the test instructions during the test?
Sınav sırasında sınav yönergelerini anlamak ne kadar zordu?
Please choose *only one* of the answers below.
Lütfen aşağıdaki cevaplardan birini seçiniz.

a. It was very difficult.. □
(Çok zordu.)

b. It was difficult. □
(Zordu.)

c. It was neither easy nor difficult. □
(Ne çok kolay ne çok zordu.)

d. It was easy. □
(Kolaydı.)

e. It was very easy. □
(Çok kolaydı.)

f. I'm not sure. □
(Emin değilim.)

6. How did you find the teachers' attitude towards you during the exam?
Öğretmenlerin sınavdaki size karşı olan tutumunu nasıl buldunuz?

a.Excellent. □
(Mükemmel)

b. Very good. □
(Çok iyi)

c. Good. □
(İyi)

d. Neutral □
(Tarafsız)

e. Negative □
(Negatif)

f. Other(s) (Please write): ____
(Diğer) (Lütfen yazınız)

7. If you have any comments about the speaking test's procedures, please write them below.
Konuşma sınavının prosedürüyle ilgili her hangi bir yorumunuz varsa, lütfen aşağıya yazınız.
Please write your comments in Turkish. (Lütfen yorumlarınızı Türkçe yazınız.)

____

____

8. What was the aspect of the speaking exam you liked most?
Sınavın en beğendiğiniz yönü neydi?
Please write your comments in Turkish.(Lütfen yorumlarınızı Türkçe yazınız.)

____

____

9. What was the aspect of the speaking exam you disliked most?
Sınavın en beğenmediğiniz yönü neydi?
Please write your comments in Turkish. (Lütfen yorumlarınızı Türkçe yazınız.)

____

____

10. How could the test be improved? Please write your comments in Turkish if you have any.
Sınav nasıl iyileştirilebilir ?
Please write your comments in Turkish. (Lüften varsa yorumlarınızı Türkçe yazınız.)

____

11. Overall, how effective do you think the test was as a test of your speaking ability? In other words, how well were you able to reflect your knowledge of language, fluency, ideas and emotions?
Genel olarak, konuşma becerinizi ölçmede sınav ne kadar etkili bir sınavdı? Diğer bir deyişle, dilbilginizi, akıcılığınızı, düşüncelerinizi ve duygularınızı ne kadar iyi yansıtabildiniz?
Please choose only <u>one</u> of the answers below.
Lütfen aşağıdaki cevaplardan sadece birisini seçiniz.

a. Excellent. □
(Mükemmel.)

b. Very good. □
(Çok iyi.)

c. Good. □
(İyi.)

d. Adequate. □
(Yeterli.)

e. Poor □
(Zayıf.)

f. Very poor □
(Çok zayıf.)

APPENDIX F

CONTENT VALIDITY INTERVIEW QUESTIONS

1) Who is the test designed for? What is it designed for?

2) What is the basis for considering whether the test is appropriate to your students?

3) Do you have any test specifications?

4) Is test content relevant to test specifications?

5) Do the items or tasks in the test match what the test as a whole is supposed to assess?

6) Does the test produce a good sample of the contents of the syllabus of the preparatory class?

7) How well do tasks/ items of the test reflect the characteristics of speaking ability?

8) What research was conducted to determine desired test content?

9) What research was conducted to evaluate test content?

10) Are the tasks and topical contents relevant to the target language use domain namely, the potential uses, or the situations that the test taker is likely to encounter)?