



T.R.  
PAMUKKALE UNIVERSITY  
THE INSTITUTE OF EDUCATIONAL SCIENCES  
DEPARTMENT OF FOREIGN LANGUAGE TEACHING  
ENGLISH LANGUAGE TEACHING

MASTER OF ARTS THESIS

AN ANALYSIS OF ASSESSMENT AND EVALUATION  
ACTIVITIES IN THE SCHOOLS OF FOREIGN  
LANGUAGES IN TURKEY

Gülce DURSUN

June, 2014

DENİZLİ

**T.R.  
PAMUKKALE UNIVERSITY  
THE INSTITUTE OF EDUCATIONAL SCIENCES  
DEPARTMENT OF FOREIGN LANGUAGE TEACHING  
ENGLISH LANGUAGE TEACHING**

**MASTER OF ARTS THESIS**

**AN ANALYSIS OF ASSESSMENT AND EVALUATION  
ACTIVITIES IN THE SCHOOLS OF FOREIGN LANGUAGES IN  
TURKEY**

**Gülce DURSUN**

**Supervisor: Assoc. Prof. Dr. Turan PAKER**

**June, 2014**

**DENİZLİ**

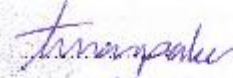
To my father İsmail Dursun,  
whose character has always illuminated all my life...

“The whole shootin’ match is for  
you”. (Arthur Miller, All My sons)

## YÜKSEK LİSANS TEZİ ONAY FORMU

Bu çalışma, İngiliz Dili Eğitimi Anabilim Dalında jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

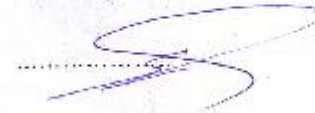
Başkan: Doç. Dr. Turan PAKER (Danışman)



Üye : Doç. Dr. Meryem AYAN




Üye: Yrd. Doç. Dr. Selami OK



Pamukkale Üniversitesi Eğitim Bilimleri Enstitüsü Yönetim Kurulu'nun  
11.07.2014 tarih ve 22/03 sayılı kararı ile onaylanmıştır.

Prof. Dr. Remazan BAŞTÜRK

Enstitü Müdürü 

## ACKNOWLEDGEMENTS

It is a real pleasure to thank to people who have contributed to this study.

I would like to express my deepest gratitude to my advisor Associate. Prof. Dr. Turan Paker without his assistance and guidance, this thesis would not have been possible. He always found time for listening to my problems and advised me on both M.A study and the workings of academic research in general

I owe special thanks to Assoc. Prof. Dr. Meryem Ayan and Asst. Prof. Dr. Selami Ok for accepting to be the members of examining committee for my thesis, and also for their constructive feedback and encouraging attitude.

I wish to express my heartfelt gratitude to Assoc. Prof. Dr. Demet Yaylı, Asst. Prof. Dr. Recep Şahin Arslan and Instructor Banu Tekingül for their suggestions and encouragement they made during my study and for what they taught during my education.

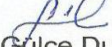
Most of all, I wish to express my heartfelt and warm thanks to my family especially my mother without her support, I would have never been able to aspire for this level of education. Without her understanding and continuous support, I would have never completed this study.

Last but not least, I would also like to express my special thanks to my brother Emrah Dursun for the patience and help for the tables.

In addition, I would like to express my special thanks to my dearest father. I can never forget how I have felt energetic by my father's support. This thesis would have never been written without him. This thesis has been dedicated to my father, İsmail Dursun without whom I wouldn't be an English teacher now.

Finally, I would like to express my warm thanks to the testing staff in the schools of Foreign Languages in the following Universities: Beykent University, Bülent Ecevit University, Eskişehir Osmangazi University, Gazi University, Istanbul Technical University, Izmir Economy University, Izmir University, Karadeniz Technical University, Muğla University and Pamukkale University

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmanın yapılması ve bulguların çözümünde bilimsel etiğe ve akademik kurallara özenle uyulduğunu; bu çalışmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atfedildiğini beyan ederim.

İmza :   
Öğrenci Adı Soyadı : Gülce Dursun

**ABSTRACT**

**AN ANALYSIS OF ASSESSMENT AND EVALUATION ACTIVITIES**

**IN THE SCHOOLS OF FOREIGN LANGUAGES IN TURKEY**

Gülce Dursun

Master of Arts Thesis in English Language Teaching Department

Supervisor: Assoc. Prof. Dr. Turan PAKER

June 2014, 119 Pages

In this study, the assessment and evaluation activities in Schools of foreign languages of various universities in Turkey have been examined. As four language skills are taught in various levels in the curriculum, we aimed to find out how these skills and their subskills have been assessed through various test types and assessment tools. The participants of the study were the Schools of Foreign Language in 10 universities in Turkey, 3 of which were private and 7 of which were state universities. As a descriptive research design, the data were collected through a questionnaire. According to the findings of the study, a certain number of various test types such as proficiency, placement, achievement are prepared and administered in schools. In addition, four language skills and subskills are assessed through various assessment tools as well as language use and vocabulary.

**Keywords:** assessment and evaluation, four skills, sub-skills, school of foreign languages

## ÖZET

### TÜRKİYE'DE YABANCI DİLLER YÜKSEKOKULLARINDA YAPILAN ÖLÇME VE DEĞERLENDİRME ETKİNLİKLERİNİN İNCELENMESİ

Gülce DURSUN

Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı

Tez Yöneticisi: Doç. Dr. Turan PAKER

Haziran, 2014, 119 Sayfa

Bu çalışmada üniversitelerin yabancı diller yüksekokullarında ölçme ve değerlendirme etkinlikleri incelenmiştir. İngilizce öğretiminde düzeylere göre dört dil becerisi öğretildiğinden bu becerilerin ve alt becerilerinin çeşitli sınav türleri ve ölçme araçlarıyla nasıl ölçüldüğü araştırılmıştır. Bu çalışmanın katılımcılarını 3 özel, 7 devlet olmak üzere toplam 10 üniversitenin yabancı diller yüksekokulları oluşturmuştur. Katılımcılarla betimleyici çalışma yapılmıştır. Bu araştırmada betimsel araştırma modeli kullanılmış ve anket yoluyla veriler toplanmıştır. Elde edilen bulgulara göre çalışmaya katılan tüm yabancı diller yüksekokullarında belirli sayıda yeterlik, yerleştirme, başarı sınavları ve quiz tipi ölçme ve değerlendirme etkinlikleri yapılmaktadır. Ayrıca her sınav türünde dört dil becerisi ve alt becerileri ölçülmekte ve bu amaçla çeşitli ölçme-değerlendirme etkinlikleri yapılmaktadır. Ek olarak, sınavlarda dil becerilerinin yanısıra kelime ve dilbilgisini ölçme ve değerlendirmeye yönelik etkinlikler de bulunmaktadır.

**Anahtar kelimeler:** ölçme değerlendirme, dört beceri, alt beceriler, yabancı diller yüksek okulu



## TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	i
ETİK SAYFASI.....	ii
ABSTRACT.....	iii
ÖZET.....	iv
TABLE OF CONTENTS .....	v
LIST OF TABLES .....	ix
LIST OF FIGURES .....	xi

### CHAPTER ONE

#### INTRODUCTION

1.1. INTRODUCTION.....	1
1.2. BACKGROUND OF THE STUDY.....	1
1.3. STATEMENT OF THE PROBLEM .....	4
1.4. THE AIM AND SCOPE OF THE STUDY .....	5
1.5. ASSUMPTIONS AND LIMITATIONS OF THE STUDY .....	6
1.6. ORGANIZATION OF THESIS.....	7
1.7. TERMS AND CONCEPTS .....	8

### CHAPTER TWO

#### REVIEW OF LITERATURE

2.1.INTRODUCTION.....	9
2.2. TESTING AND ASSESSMENT.....	9
2.2.1. The definition of testing.....	9
2.2.2. The definition of assessment.....	10
2.2.2.1. Definition of Assessment in behaviorism.....	12

2.2.2.2. How Behaviorism impacts in learning and testing.....	12
2.2.2.3. Definition of assessment in cognitivism.....	13
2.2.2.4. How cognitivism impacts in learning and testing.....	16
2.2.2.5. Definition of assessment in constructivism.....	16
2.2.2.6. Definition of assessment in humanism.....	18
2.2.3. The advantages of assessment.....	19
2.3. PRINCIPLES OF LANGUAGE ASSESSMENT.....	20
2.3.1. Reliability.....	20
2.3.2. Practicality.....	25
2.3.3. Validity.....	25
2.3.3.1. Construct- Related Evidence.....	27
2.3.3.2. Content Validity.....	28
2.3.3.3. Instructional Validity.....	29
2.3.4. Authenticity.....	30
2.3.5. Washback.....	31
2.3.5.1. Definitions of Washback.....	32
2.3.5.2. Origin of Examinations and Washback.....	34
2.3.5.3. Functions and Mechanism of Washback.....	36
2.3.5.4. Negative Washback.....	38
2.3.5.5. Positive Washback.....	40
2.3.5.6 Measurement-driven Instruction and Curriculum Alignment.....	41
2.3.5.7. Studies Investigating Washback Effects.....	43
2.3.5.8. Studies Conducted on Washback in Turkey.....	44
2.3.5.9. Washback Effect of Examinations in Overall Education.....	47
2.3.5.10. Washback Effect of Examinations in FL Classrooms and Programs..	49
2.4. SUMMATIVE ASSESSMENT.....	53
2.5. FORMATIVE ASSESSMENT.....	54
2.6. APPROACHES IN LANGUAGE TESTING.....	56
2.6.1. Integrative Approach.....	56
2.6.2. Communicative Approach.....	57

2.6.2.1. Characteristics and Types of Tests in Communicative Approach.....	57
2.6.2.2. Strengths of Communicative Approach.....	58
2.6.2.3. Weaknesses of Communicative Approach.....	58

### **CHAPTER THREE**

#### **RESEARCH METHODOLOGY**

3.1. INTRODUCTION.....	60
3.2. NATURE OF STUDY.....	60
3.3. METHODOLOGY OF THE STUDY.....	61
3.3.1. Setting.....	61
3.3.2. Participants.....	63
3.3. DATA COLLECTION AND PROCEDURES.....	65
3.4. DATA ANALYSIS AND PROCEDURES .....	65

### **CHAPTER FOUR**

#### **FINDINGS AND DISCUSSION**

4.1. INTRODUCTION.....	66
4.2. THE KIND OF ASSESSMENT AND EVALUATION ACTIVITIES CARRIED OUT IN THE SCHOOLS OF FOREIGN LANGUAGES IN TURKEY.....	67
4.3. ASSESSMENT OF LISTENING SKILL.....	71
4.4. ASSESSMENT OF READING SKILL.....	75
4.5. ASSESSMENT OF SPEAKING SKILL.....	79
4.6. ASSESSMENT OF WRITING SKILL .....	83
4.7. ASSESSMENT OF LANGUAGE USE .....	87
4.8. ASSESSMENT OF VOCABULARY.....	89

**CHAPTER FIVE****CONCLUSION**

5.1. INTRODUCTION.....	93
5.2. SUMMARY OF THE STUDY.....	93
5.3. IMPLICATION OF THE STUDY.....	94
5.4. SUGGESTIONS FOR FURTHER RESEARCH.....	95
REFERENCES.....	97
APPENDIX .....	110
CURRICULUM VITAE .....	118

**LIST OF TABLES**

Table 3.1. The number of instructors and students.....	64
Table 4.1. Frequency of the tests administered in an academic year in the School of Foreign Languages.....	67
Table 4.2. Application of a proficiency test in an academic year.....	68
Table 4.3. Application of an achievement test in an academic year.....	68
Table 4.4. Application of listening quiz.....	69
Table 4.5. Application of reading quiz.....	69
Table 4.6. Application of writing quiz.....	70
Table 4.7. Application of language use.....	70
Table 4.8. Application of vocabulary quiz.....	71
Table 4.9. Application of grammar quiz.....	71
Table 4.10. Weight of listening skill.....	72
Table 4.11. Assessment of listening subskills.....	73
Table 4.12. The item types used in listening.....	74
Table 4.13. Weight of reading skill.....	76
Table 4.14. Assessment of reading sub skills.....	77
Table 4.15. The item types used in reading .....	78
Table 4.16. The item types used in testing translation.....	78
Table 4.17. Weight of speaking skill.....	80
Table 4.18. Assessment of speaking sub skills.....	81
Table 4.19. The item types used in speaking.....	81

Table 4.20. Application of speaking test.....	82
Table 4.21. Weight of writing skill.....	83
Table 4.22. Assessment of writing sub skills at the paragraph level.....	84
Table 4.23. The item types used at the paragraph level.....	85
Table 4.24. Assessment of writing sub skills at the essay level.....	86
Table 4.25. The item types used in writing skill at the essay level.....	86
Table 4.26. Weight of language use .....	87
Table 4.27. Item types of language use.....	88
Table 4.28. Weight of vocabulary .....	90
Table 4.29. Assessment of vocabulary.....	91
Table 4.30. The item types used in vocabulary .....	91

**LIST OF FIGURES**

Figure 2.1: Factors that affect language test scores.....	24
Figure 2.2: A proposed holistic model of washback based on ideas of Hughes (1993), Bachman and Palmer (1996).....	33
Figure 2.3: Cycle of Summative Assessment.....	54
Figure 2.4: Formative Assessment Cycle.....	55
Figure 2.5: Cycle of Formative Assessment.....	55

## **CHAPTER ONE**

### **INTRODUCTION**

#### **1.1. INTRODUCTION**

Teachers all want what they teach to be learnt by their students. They have been looking for ways to make their classes more important for students. One of the strategies used is to test what they teach so as to help students learn. If there is a test at the end of instruction, they have a good reason to study. The significance of the evaluation process stems from the fact that as a result of this process, learners either pass or fail the teaching programme. Therefore, teaching and testing go hand in hand. Thus, testing is an indispensable part of second language teaching.

Teachers have to assess the knowledge of the learners, which is quite difficult due to the fact that the thing to be measured is not something tangible. There are too many variables that affect the performances of the test takers, and a direct measurement of 'foreign language knowledge' is not possible (Hughes, 2003). Accordingly, the assessor has to think and develop several techniques and procedures in order to fulfill his/her aim. Within this study, the researcher aims to deal with the present assessment and evaluation of English in terms of all skills and sub-skills in the Schools of Foreign Languages in Turkey.

#### **1.2. BACKGROUND OF THE STUDY**

Education is one of the most important and difficult issues of the society. It has been defined in many ways. Sönmez (1994) defines education "as a



period of changing behaviors” (p.18). The teaching and learning process is what the term, education, includes. While teaching, the sentence that should be remembered is that ‘to teach someone is to touch a life’ (Johnson, 2007). It is so because the effects of this process go on throughout the students’ lives. Then, each step of education should leave a positive trace on them. Assessment, an important stage of education, has a vital impact in this process. Assessment was defined as “informing and improving students’ on-going learning” (Cowie and Bell, 1999:260). Unfortunately, implementing assessment which has a positive effect on student learning is not as easy as it sounds. It is clear that assessment has an important effect on teaching and learning process. Therefore, it is of crucial importance for teachers to realize that the main purpose of assessment is to collect information about individuals or group of individuals in order to better understand them.

The purposes of assessment are providing feedback to the students and being a diagnostic and monitoring tool for the instruction (Butler and McMun, 2006). If the aim is to understand our students better, there should be an ongoing interaction between the teacher and the students, and this certainly will make a positive effect on learning and teaching process. This interaction is the important part of assessment. At this point, how we assess the students becomes more important than the assessment itself. There are two types of assessment; formative and summative. Summative assessment which is used to grade the learners’ products of learning aims to get feedback about overall judgement at the end of a course (Ciel, 2000). Tests and examinations are a classic way of measuring student progress and these are the parts of summative assessment. The aim of the students is to pass the exams or get high marks from the tests. Most of the teachers use summative assessment because it aims to record the overall achievement of a student in a systematic way (Lambert and Lines, 2000).

In contrast to summative assessment, formative assessment, which is a systematic process of continuously gathering evidence about learning, is used to identify a student’s current level of learning and helps the student reach the desired learning goal. Being active participants of the process, students share learning goals and understand how their learning is progressing. They are

informed about what next steps they need to take and how to take them (Heritage, 2007). Students become aware of their weaknesses and strengths in this way. Hence summative assessment focuses on the product obtained at the end of the teaching and learning process, formative assessment, on the other hand, focuses on the process and each step is decided and planned continuously. According to Ökten (2009) what happens with the use of summative assessment is that students cannot learn to create, analyze and learn how to learn. They only study to pass the exams so they cannot transform what they have learnt to their lives and become life-long learners.

According to the definitions above, it is clear that assessment and testing are very significant because according to Ökten (2009), the teachers determine what, when and how to teach, and they are dependent on the performance of the students. With the results acquired, students become aware of their learning in terms of what they have learned and how much they have learned. In this way they are able to take some decisions about their own learning.

According to Rudman (1989), testing and teaching are not separate entities. Teaching has always been a process of helping others to discover "new" ideas and "new" ways of organizing that which they have learned. Whether this process takes place through systematic teaching and testing, or whether it is through a discovery approach, testing is, and remains, an integral part of teaching. We can see the best example of it in what Davies (1968:5) states; "the good test is an obedient servant since it follows and apes the teaching". There are also some studies raising questions about whether improvements in test score performance actually develop improvement in learning (Cannell, 1987; Shepard, 1989).

Messick (1996:241-242) points out that:

... in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviors of listening, speaking, reading and writing of the language being learnt. Ideally, the move from learning experiences to test exercises should be seamless. As a consequence, for optimal positive washback there should be little if any difference between

activities involved in learning the language and activities involved in preparing for the test.

### **1.3. STATEMENT OF THE PROBLEM**

Testing has been used for decades, but some concerns about its influence have recently increased. According to Rudman (1989), testing and teaching are not separate entities as testing is a useful tool at the beginning of the school year, and testing can aid in having some decisions about grouping students in the class. In addition, testing can be used to diagnose what individual pupils know and can help the teacher determine the pace of classroom instruction. As Sarıçoban (2011:398) states “for decades, testing has been a neglected area in foreign language teaching (FLT) not only in our country but also other countries in that foreign language (FL) tests lack the outcomes of the language learning process.” Foreign language tests usually seem to focus on recognition rather than production skills of FL learners. In addition, Ökten (2009) states that assessment in our country is mainly based on a product approach which focuses only on what the students have learnt. This problem still exists in our context.

Assessment describes learning achieved at a certain time (Ökten, 2009) . The desired goal becomes passing the exams or getting higher marks from the standardized tests and this makes us realize that the importance of receiving a proper education with the evaluation process. The significance of the evaluation process stems from the fact that as a result of this process, learners either pass or fail. This is not as straightforward as it looks because passing or failing a particular exam may come to mean that the candidate is accepted or not. It is commonly assumed that “teachers will be influenced by the knowledge that their students are planning to take a certain test and will adapt their teaching methodology and lesson content to reflect the tests demands” (Taylor, 2005:154). In order to achieve this, teachers should create opportunities to assess how students are learning and then use this information to make beneficial changes in their teaching. This is the diagnostic use of assessment, and it provides feedback to teachers and students over the course of instruction

(Boston, 2002). It provides the learners with the opportunities to learn how to learn in order to make them more knowledgeable.

Although the studies mentioned here have contributed to the field of English Language Teaching, they have not investigated the effects of testing in terms of principles of language assessment, item types, the weight of skills and sub-skills. To fulfill this need, we attempt to focus on the recent assessment and evaluation activities and try to deal with the use of skills and sub-skills and create awareness for teachers, administrators, students and testing offices.

#### **1.4. THE AIM AND THE SCOPE OF THE STUDY**

Language testing cannot be considered apart from the teaching-learning process (Woodford, 1980). Teachers need to know about their students' progress and difficulties. In this way, they can adapt their own work to meet students' needs. This means to teach and then question whether it has worked or not. This continuous process is what formative assessment does. The teacher takes steps to close the gap between the students' current learning and the goal by modifying instruction, assessing again to give further information about learning and modifying instruction according to the students' progress (Heritage, 2007). But according to the findings of a study conducted by Köksal (2004) on 'Teachers' Testing Skills in ELT' in Turkey, most of the foreign language teachers in our schools prepare and administer language tests which are far from satisfactory. The reasons underlying this situation are; Teachers' lack of training in testing, and testing and teaching do not overlap. Teachers teach something but test something else. As Heritage (2007: 141) states; "by this way the teacher takes steps to close the gap between the students' current learning and the goal by modifying instruction, assessing again to give further information about learning and modifying instruction according to the students' progress." Moreover, Hinkel (2006:113) states "in meaningful communication, people employ incremental language skills not in isolation, but in tandem". This shows that integration of skills is important in language learning. In order to understand this, we will look at how input and output are connected in the classroom, how skills can be integrated, and how skill and language work are

connected. Therefore, it is important to be aware of its consequences. For this reason, this research focuses on the assessments in the schools of Foreign Languages and the weights of skills and sub skills in the assessment procedures.

The main purpose of this study is to describe the assessment and evaluation activities with the use of skills and sub-skills in the Schools of Foreign Languages in Turkey and create awareness for those involved to prepare and administer more valid, reliable and practical language tests by providing necessary background and theoretical knowledge about language testing. With this aim in mind, this study attempts to find answers to the following research questions:

1. What kind of assessment and evaluation activities are done in the Schools of Foreign Languages in Turkey?
2. How is the listening skill assessed?
3. How is the reading skill assessed?
4. How is the speaking skill assessed?
5. How is the writing skill assessed?
6. How is the language use assessed?
7. How is the vocabulary assessed?

## **1.5. THE ASSUMPTIONS AND LIMITATIONS THE OF THE STUDY**

The Assumptions below will be considered throughout this study:

All the data used in this study and prepared for this study are valid and reliable. Next, assessment and evaluation have been carried out in line with all skills and with their subskills, and this is supported by alternative assessment. Furthermore, even though this study has been carried out in the Schools of Foreign Languages in Turkey, and the data have been collected from the same

level of schools, generalizations can be made for the schools in the same position or for the students on the same educational level.

The limitations below will be considered throughout this study:

This study is limited to ten universities of Turkey and was carried out in 2012-2013 academic year. In this study, a questionnaire developed by the researcher was used to collect data, so the results of the study are limited to these instruments.

## **1.6. ORGANIZATION OF THE THESIS**

This thesis is composed of five chapters. Chapter One presents the background to the study. It then proposes the purpose of the study and the research questions. The first chapter also includes the significance, assumptions, and limitations of the study, and it finally describes the organization of the thesis.

Chapter Two reviews the literature on assessment and evaluation in language learning in detail. The effects of them on foreign language learning and teaching are taken into consideration in this chapter.

Chapter Three reports the methodology of the study. Survey studies, rationale for the survey research design, elements in the survey such as setting, participants, and the procedures of the pilot study and main study are described in this chapter.

Chapter Four reports and discusses the findings of this study in detail aiming to seek answers for the research questions.

Chapter Five discusses the findings of the study aims to draw conclusions through the findings. Implications and suggestions for further research are also proposed in this chapter.

## 1.7. TERMS AND CONCEPTS

**Summative Assessment:** It is designed to get feedback about overall judgement at the end of a course of learning and used to grade the learners' products of learning (Atkins, et al, 1993:7, cited in Ciel, 2000).

**Formative Assessment:** It is designed to provide feedback on the progress of learning and used to make adjustments in learning goals, teaching and learning methods, materials and so on (Atkins, et al, 1993:7, cited in Ciel, 2000).

**Life-long Learning:** The term recognizes that learning is not confined to childhood or the classroom, but takes place throughout life and in a range of situations ([www.wikipedia.com](http://www.wikipedia.com), 01.03.2014).

**Evaluation:** The term evaluation has been defined in many different ways, sometimes resulting in ambiguity in the use of the term. The term has been defined here "as the systematic attempt to gather information in order to make judgments and decisions" about the program at issue (Lynch, 1996:2).

**Washback:** Washback (Aldersen & Wall, 1993) or backwash (Biggs, 1995, 1996) refers to the influence of testing on teaching and learning. The concept is rooted in the notion that tests or examinations can and should drive teaching, and hence learning, and is also referred to as measurement-driven instruction (Popham, 1987).

## **CHAPTER TWO**

### **LITERATURE REVIEW**

#### **2.1. INTRODUCTION**

In this chapter, background information about testing and assessment, assessment types is provided. Furthermore, summative assessment, formative assessment as well as assessment in behaviorism, cognitivism and constructivism are introduced. Then, we have tried to emphasize on the principles of language assessment. As the theoretical framework of the study, approaches in language testing is explained.

#### **2.2. TESTING AND ASSESSMENT**

##### **2.2.1. The Definition of Testing**

According to Bachman (1990) the two major uses of language tests are: (1) as sources of information for making decisions within the context of educational programs; and (2) as indicators of abilities or attributes that are of interest in research on language, language acquisition, and language teaching. In educational settings the major uses of test scores are related to evaluation, or making decisions about people or programs.

Brown (2004:4) makes the distinction between testing and assessment as follows:



Tests are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated. Assessment, on the other hand, is an ongoing process that encompasses a much wider domain. Whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance. Tests, then, are a subset of assessment; they are certainly not the only form of assessment that a teacher can make. Tests can be useful devices, but they are only one among many procedures and tasks that teachers can ultimately use to assess students.

There are many other definitions of testing. Carroll states that "...a psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual" (cited in Bachman, 1990:20). "Testing is part of assessment, and it measures learner achievement" (Coombe, et al. 2007:XV). Bachman (1990:24) concludes by stating: "...then, not all measures are tests, not all tests are evaluative, and not all evaluation involves either measurement or tests."

As Upshur (1971) noted, language tests can be valuable sources of information about the effectiveness of learning and teaching. Language teachers regularly use tests to help diagnose student strengths and weaknesses, to assess student progress, and to assist in evaluating student achievement. Language tests are also frequently used as sources of information in evaluating the effectiveness of different approaches to language teaching. As sources of feedback on learning and teaching, language tests can thus provide useful input into the process of language teaching. Language tests can, thus, provide the means for more carefully focusing on the specific language abilities that are of interest.

### **2.2.2. The Definition of Assessment**

According to Coombe et al. (2007:XV) "... assessment is an umbrella term for all types of measures used to evaluate student progress." In its most general definition, assessment is the process of gathering, interpreting, recording and using information about students' responses to educational tasks (Lambert and Lines, 2000). According to Ökten (2009), assessment is one of

the most important stages of learning and teaching both for the teachers and the learners. When the teachers determine what, when and how to teach, they are dependent on the results of the students. With the results acquired, students become aware of their learning in terms of what they have learned and how much they have learned. In this way, they are able to take some decisions about their own learning.

Assessment was also defined as informing and improving students' ongoing learning (Cowie and Beverley, 1999). It is the analysis of data about the needs, interests, learning styles and achievements of students (Ming, 2002). Assessment is an ongoing process, and tests are a subset of the assessment. It seems, indeed, that each affects the other: methods of assessment may affect teaching in the classroom (Cheng 1997, Wall 1997), while new theories of language learning and teaching lead to changes in testing practices (Spolsky, 1995). By assessment, information as to the learner's language ability and achievement is collected in several ways, therefore, the assessment forms a crucial part of the evaluation process.

The purpose of assessment is providing feedback to the students and being a diagnostic and monitoring tool for the instruction (Butler and McMun, 2006). If the aim is to understand our students better, there should be an ongoing interaction between the teacher and the students, this certainly will make a positive effect on learning and teaching process. This interaction is the important part of assessment. In order to achieve this, teachers should create opportunities to assess how students are learning and then use this information to make beneficial changes in their teaching. This is the diagnostic use of assessment and it provides feedback to teachers and students over the course of instruction (Boston, 2002). In order to assess, we should bear in mind, what to assess, how to assess, who to assess, in which way to assess and how long to assess (Temel, 2007).

### **2.2.2.1. Definition of Assessment in Behaviorism**

Behaviorism is a psychological theory of learning which was very influential in the 1940s and 1950s. Traditional behaviorists believed that language learning is the result of imitation, practice, feedback on success and habit formation. According to this view the quality and quantity of the language which child hears, as well as the consistency of the reinforcement offered by others in the environment, should have an effect on child's success in language learning. (cited in Lighthbrown & Spada 2006:9 )

Learning is “a persisting change in performance or performance potential that results from experience and interaction with the world” (Driscoll, 2000:3). These two ideas—the importance of measurable and observable performance and the impact of the environment, comprise foundational principles of the behaviorist approach to learning. The basic argument is that only observable, measurable behavior is the appropriate object for psychological study. Initially, the theory contended that certain behavioral responses come to be associated with specific environmental stimuli (Driscoll, 2000). Skinner (1957) extended the concept of associations. Skinner argued that a behavior is more likely to reoccur if it has been reinforced or rewarded. Thus reinforcement can be used to strengthen existing behaviors, as well as learn new one.

### **2.2.2.2. How Behaviorism Impacts in Learning and Assessment**

Positive and negative reinforcement techniques of behaviorism can be very effective. Teachers use behaviorism when they reward or punish student behaviors. Things to remember when incorporating behaviorist principles into teaching are that; writing observable and measurable behavioral learning outcomes, specifying the desired performances in advance (the learning outcomes serve this purpose) and verifying learning with appropriate assessments, emphasizing performance and practicing in an authentic context, using instructional strategies to shape desired skills and reinforcing accomplishments with appropriate feedback (Driscoll, 2000).

As explained by Gagne (1965), “to ‘know,’ to ‘understand,’ to ‘appreciate’ are perfectly good words, but they do not yield agreement on the exemplification of tasks. On the other hand, if suitably defined, words such as to ‘write,’ to ‘identify,’ to ‘list,’ do lead to reliable descriptions” (p. 43). Thus, behaviorally-stated objectives became the required elements of both instructional sequences and closely related mastery tests.

In accordance with behaviorism, Brown (2004:29) emphasizes that “give praise for strengths and give strategic hints on how a student might improve certain elements of performance. Making the test performance an intrinsically motivating experience from which a student will gain a sense of accomplishment and challenge.”

Testing played a central role in behaviorist instructional systems. To avoid learning failures caused by incomplete mastery of prerequisites, testing was needed at the end of each lesson, with re-teaching to occur until a high level of proficiency was achieved. In order to serve this diagnostic and prescriptive purpose, test content had to be exactly matched to instructional content by means of the behavioral objective. Behavioristic assumptions also explain why, in recent years, advocates of measurement-driven instruction were willing to use test scores themselves to prove that teaching to the test improved learning (Popham, Cruse, Rankin, Sandifer, & Williams, 1985).

### **2.2.2.3. Definition of Assessment in Cognitivism:**

Cognitive theorists like Piaget and Gagne recognize that much learning involves associations established through contiguity and repetition. They also acknowledge the importance of reinforcement, although they stress its role in providing feedback about the correctness of responses over its role as a motivator. “Cognitive theorists view learning as involving the acquisition or reorganization of the cognitive structures through which humans process and store information” (Good and Brophy, 1990:187). According to Krause et al, (2003:114), learning and assessment, has a number of presumptions:

- children are committed to the goals of the teacher,

- children have self-discipline, and
- there will be an experienced “expert “ available to assist
- the teacher designs appropriate courses of action.

Cognitive psychologists asserted that meaning, understanding, and knowing were significant data for psychological study. Instead of focusing rather mechanistically on stimulus-response connections, cognitivists tried to discover psychological principles of organization and functioning. Ausubel (1965:4) noted:

From the standpoint of cognitive theorists, the attempt to ignore conscious states or to reduce cognition to mediational processes reflective of implicit behavior not only removes from the field of psychology what is most worth studying but also dangerously oversimplifies highly complex psychological phenomena.

While the cognitive approach sees its primary function as the making of meaning out of experiences with the world, and creating links with learning that had previously taken place, the presumption is that the content of the learning is valid and appropriate for each child, and also that each child has a similar learning style- a belief that is strongly contested by Maslow (1968) and Rogers (1969) and many others.

Mergel (1998) mentions the key concepts of cognitive theory below:

—*Schema*: An internal knowledge structure. New information is compared to existing cognitive structures called "schema". Schema may be combined, extended or altered to accommodate new information.

—*Three-Stage Information Processing Model*: input first enters a sensory register, then is processed in short-term memory, and then is transferred to long-term memory for storage and retrieval.

—*Sensory Register*: It receives input from senses which lasts from less than a second to four seconds and then disappears through decay or replacement. Much of the information never reaches short term memory but all information is monitored at some level and acted upon if necessary.

—*Short-Term Memory (STM)*: Sensory input that is important or interesting is transferred from the sensory register to the STM. Memory can be retained here for up to 20 seconds or more if rehearsed repeatedly. Short-term memory can hold up to 7 plus or minus 2 items. STM capacity can be increased if material is chunked into meaningful parts.

—*Long-Term Memory and Storage (LTM)*: It stores information from STM for long term use. Long-term memory has unlimited capacity. Some materials are "forced" into LTM by rote memorization and over learning. Deeper levels of processing such as generating linkages between old and new information are much better for successful retention of material.

—*Meaningful Effects*: Meaningful information is easier to learn and remember (Cofer, 1971, cited in Good and Brophy, 1990). If a learner links relatively meaningless information with prior schema, it will be easier to retain (Wittrock, Marks, & Doctorow, 1975, cited in Good and Brophy, 1990).

—*Transfer Effects*: The effects of prior learning on learning new tasks or material.

—*Interference Effects*: It occurs when prior learning interferes with the learning of new material.

—*Organization Effects*: When a learner categorizes input such as a grocery list, it is easier to remember.

—*Levels of Processing Effects*: Words may be processed at a low-level sensory analysis of their physical characteristics to high-level semantic analysis of their meaning. ( Craik and Lockhart, 1972, cited in Good and Brophy, 1990) The more deeply a word is processed, the easier it will be to remember.

—*State Dependent Effects*: If learning takes place within a certain context it will be easier to remember within that context rather than in a new context.

—*Schema Effects*: If information does not fit a person's schema it may be more difficult for them to remember and what they remember or how they conceive of it may also be affected by their prior schema.

#### **2.2.2.4. How Cognitivism Impacts in Learning and Assessment**

Cognitive theories of learning focus on the mind and attempt to model how information is received, assimilated, stored, and recalled. The implication is that by understanding the mechanics of this process, we can develop teaching methods more suited to fostering the desired learning outcome, which is a shared desire with behaviorists.

Cognitivists such as Piaget and Gagne argue that while things like the environment are important inputs to learning, learning is more than simply the collection of inputs and the production of outputs. The mind has the ability to synthesize, analyze, formulate, and extract received information and stimuli in order to produce things that cannot be directly attributed to the inputs given. Under cognitive learning theory, it is believed that learning occurs when a learner processes information. The input, processing, storage, and retrieval of information are the processes that are at the heart of learning (Cameron, 2005).

Cognitive learning theories infuse the classroom curriculum with meaningful interaction. Children grow together in intricate ways. Not all experiences can be measured equally, because everyone's experience is utterly unique. By collecting individual experiences the classroom builds a learning environment that is both deep and authentic. The assessment of such an environment may seem difficult at first glance, because the philosophy collides with standardized assessment practices. However, with practice, the teacher can realize a more artistic approach to assessment that values depth of understanding rather than test measures.

#### **2.2.2.5. Definition of Assessment in Constructivism**

According to Williams and Burden (1997), in contrast to more traditional views which see learning as the accumulation of facts or the development of skills, the main underlying assumption of constructivism is that individuals are actively involved right from birth in constructing personal meaning that is their own personal understanding from their experiences. In addition, Al-Weher

(2004) points out that learning takes place in contexts, where learners construct what they learn and understand their learning as a function of their experiences in situation. The teacher leads the student to construct new understanding and acquire new skills (Brooks and Brooks, 2001). From a constructivist perspective, formative assessments are more valuable to the learner (Lamon, 2007).

Brooks and Brooks (2001) describe what assessment, in a constructivist classroom, looks like: Below is a list of the important principles that guide the work of a constructivist teacher:

- Constructivist teachers encourage and accept student autonomy and initiative.
- Constructivist teachers use raw data and primary sources along with manipulative, interactive, and physical materials.
- Constructivist teachers use cognitive terminology such as "classify," "analyze," "predict," and "create" when framing tasks.
- Constructivist teachers allow student responses to drive lessons, shift instructional strategies, and alter content.
- Constructivist teachers inquire about students' understandings of concepts before sharing their own understandings of those concepts.
- Constructivist teachers encourage students to engage in dialogue both with the teacher and with one another.
- Constructivist teachers encourage student inquiry by asking thoughtful, open-ended questions and encouraging students to ask questions of each other.
- Constructivist teachers seek elaboration of students' initial responses.
- Constructivist teachers engage students in experiences that might engender contradictions to their initial hypotheses and then encourage discussion.
- Constructivist teachers allow a waiting time after posing questions.
- Constructivist teachers provide time for students to construct relationships and create metaphors.
- Constructivist teachers nurture students' natural curiosity through frequent use of the learning cycle model.



### 2.2.2.6. Definition of Assessment in Humanism

Humanistic approaches to teaching, learning and assessment take a totally different belief system as a beginning point than behaviorist and cognitive approaches. Humanism is “any system of thought that is predominantly concerned with the human experience of reasoning rather than with the spiritual aspects of life” (Krause et al., 2003:172). Humanism is also described as the “belief that individual human beings are the fundamental source of all value and have the ability to understand—and perhaps even to control—the natural world by careful application of their own rational faculties” (Dictionary of Philosophical Terms and Names [Online]).

Maslow believes that unless children’s basic needs are met, they may not find other learning worth engaging in (cited in Dembo, 1944:206). Rogers (1983:21) was adamant that “...prescribed curriculum, similar assignments for all student, lecturing as almost the only mode of instruction, standard tests by which all students are externally evaluated and instructor-chosen grades as the measure of learning...” was a flawed approach. He saw humanism as the alternative “freedom to learn”, where teachers and parents were to take on the role of facilitator, who “actively listen” to children, and guide them in their own endeavors by really engaging in children’s thinking and problem solving with them and developing a good and positive relationship with the learner. He also highlights another crucial component of a teacher’s repertoire: they must be truly human, and that their human qualities are a crucial part of the teaching learning equation (Dembo, 1944:209).

The humanistic approach is a broad term that encompasses three main approaches (Kirschenbaum, 2003: 64):

- Humanistic content curricula - Teaching topics that are directly relevant to the students' lives (e.g. drugs awareness)
- Humanistic process curricula - Focuses on the whole student and can include teaching assertiveness training, for example.

- Humanistic school and group structures - restructuring the whole timetable and school environment in order to facilitate humanistic teaching or just individual classes.

### **2.2.3. The Advantages of Assessment**

The main aim of testing and assessment is to identify how much of the targets are attained. As a result of the assessment, if there is no relation between the results and targets, the system should be renewed. By the help of testing and assessment not only it is easy to state the achievements and failures but also every target can be planned according to the level of students. By this way students can be guided with feedbacks.

As Temel (2007:20), suggests, the advantages of assessment are listed below:

- The teacher knows her students.
- The student knows her teacher.
- The teacher knows herself better in terms of techniques and methods.
- It motivates the students.
- The parents will know the student's failure or success.
- It will help for the improvement of education.

In the study conducted by Steadman (1998), the advantages of the assessment are emphasized as follows:

—tuning into students' voices and as a result having students who are more satisfied.

—the opportunity to engage in reflection on and systematic change of their teaching

—student improvement and involvement in learning, because according to her assessment is done to obtain feedback on the effectiveness of and

student satisfaction with teaching and classroom activities, To improve teaching, to monitor students' learning, to improve students' learning (in terms of retention or learning skills), to improve communication and collaboration with students.

Besides these, tests help teachers diagnose students' strengths and weaknesses, assess students' progress, and assist in evaluating students' achievement (Bachman, 1990:3). During language teaching period, from students' perspectives, this helps teachers to teach effectively and motivate students and trigger them to learn English more eagerly by providing constructive feedback. The students can evaluate both themselves and their peers. From teachers' perspectives, this helps teachers plan the schedule according to the unattained goals and revise it properly, to evaluate their teaching skills, methods , ways and to evaluate the students in order to understand how well the teacher has taught or not taught so far. Moreover, as it helps to state the strengths and weaknesses of the students, it is like a SWOT analysis. It states strengths, weaknesses, creates opportunity to use the language and threat accordingly. Language teachers should determine the success levels of their students in acquiring the intended behavior, and the success levels of the students can only be determined via the process of measurement and the assessment procedure including measurable objectives, decision-making, setting tasks, and scoring (Weigle, 2007). As this assessment provides constructive feedback, this promotes autonomous learners (Tambini, 1999).

## **2.3. PRINCIPLES OF LANGUAGE ASSESSMENT**

### **2.3.1. Reliability**

As Bachman (1990) points out, the investigation of reliability is concerned with answering the question, 'How much of an individual's test performance is due to measurement error, or to factors than the language ability we want to measure?' and with minimizing the effects of these factors on test scores. Bachman (1990:161) emphasizes that:

The investigation of reliability involves both logical analysis and empirical research; we must identify sources of error and estimate the magnitude of their effects on test scores. In order to identify sources of error, we need to distinguish the effects of the language abilities we want to measure from the effects of other factors, and this is a particularly complex problem.

Reliability simply refers to consistency and dependability (Gatewood & Field, 2001). Reliability is the consistency of the measurement or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. That is, a test is considered reliable if we get the same result after administering it twice to the same subject group. A same test delivered to a same student across time administration must yield same results. This means consistency and a reliable test means a dependable test (McBride, 2010)(retrieved from: [http://www.sagepub.com/upm-data/58460\\_Chapter\\_4.pdf](http://www.sagepub.com/upm-data/58460_Chapter_4.pdf)) . As Brown (2000:386) suggests: “If you give the same test to the same subject or matched subjects on two different occasions, the test itself should yield similar results, it should have test reliability.”

In a test there must be consistency related with scorers, test takers, the time for testing. As Bachman (1990:24) points out: “reliability thus has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context.” According to Henning (1987), reliability is a measure of accuracy, consistency, dependability, or fairness of scores resulting from the administration of a particular examination e.g. 75% on a test today, 83% tomorrow – problem with reliability.

Factors affecting reliability are (Heaton, 1990: 155-156; Brown, 2004:21-22):

1. student-related reliability: students’ personal factors such as motivation, illness, anxiety can hinder from their ‘real’ performance,

2. rater reliability: either intra-rater or inter-rater leads to subjectivity, error, bias during scoring tests. As Brown (2004) pointed out, the careful specification of an analytical scoring instrument can increase rater reliability. Reliability of a test can be determined both by estimating the *rater reliability* and *instrument*

*reliability*. Rater reliability can be done either by *inter-rater reliability* which refers to “a measure of whether two or more raters judge the same set of data in the same way” (Mackey & Gass, 2005:129) or by *intra-rater reliability* which means that the rater judge the data the same at different times.

3. test administration reliability: when the same test administered in different occasion, it can result differently.

4. test reliability: dealing with duration of the test and test instruction. If a test takes a long time to do, it may affect the test takers performance such as fatigue, confusion, or exhaustion. Some test takers do not perform well in the timed test. Test instruction must be clear for all of test takers since they are affected by mental pressures.

On the other hand, Hughes (2003:8) suggests some ideas as to how to make tests more reliable. These are listed below:

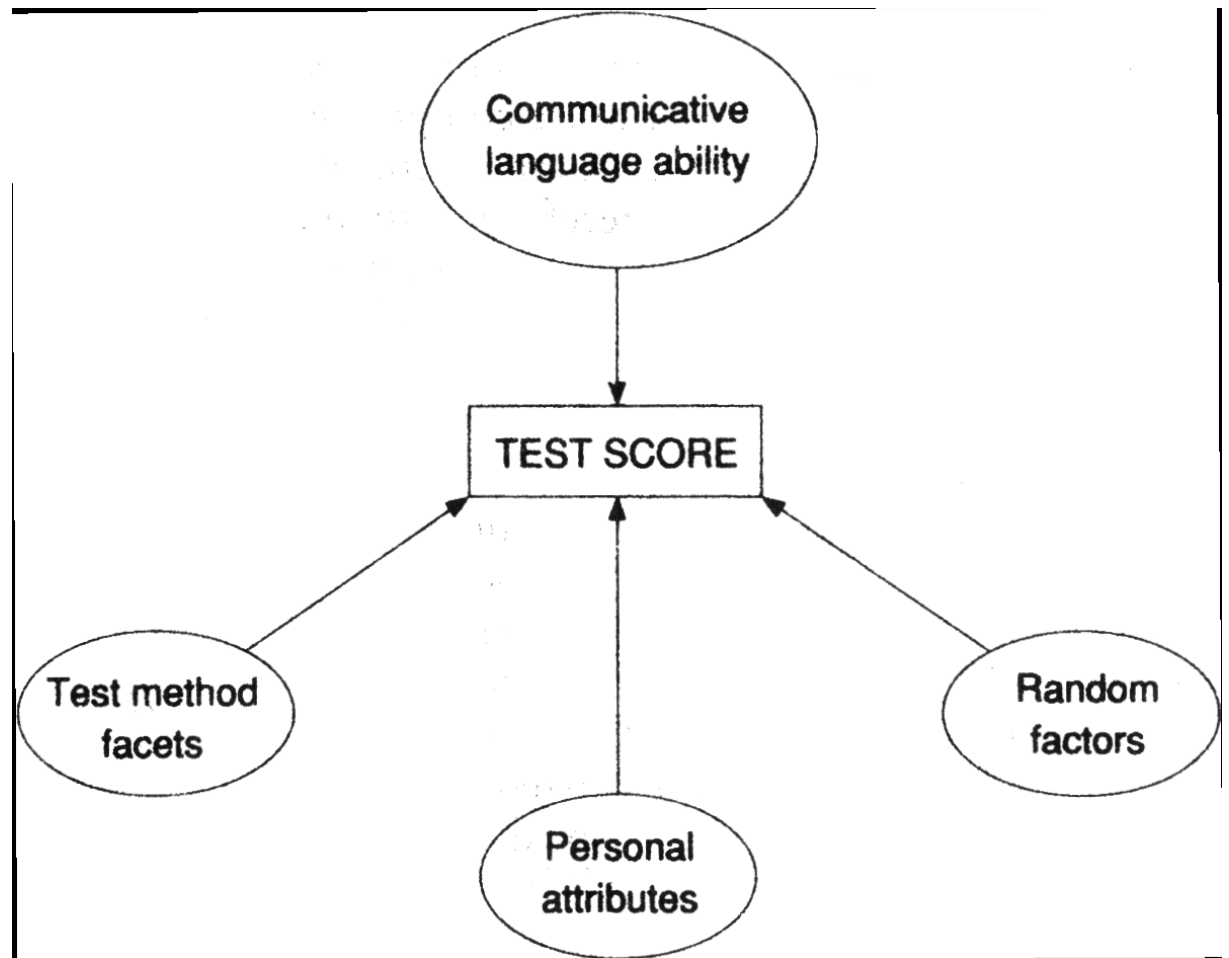
- Take enough samples of behavior
- Do not allow candidates too much freedom
- Write unambiguous items
- Provide clear and explicit instructions
- Ensure that the tests are well laid out and perfectly legible
- Candidates should be familiar with format and testing techniques
- Provide uniform and non-distracting conditions of administration
- Use items that permit scoring which is as objective as possible
- Make comparisons between candidates as direct as possible
- Provide a detailed scoring key
- Train scorers
- Agree respectable responses and appropriate scores at the outset of scoring
- Identify candidates by number , not name
- Employ multiple, independent scoring

Some methods are employed to gain reliability of assessment (Heaton, 1975:156; Weir 1990:32; Gronlund and Waugh, 2009:59-64). They are:

1. test-retest/re-administer: the same test is administered after a lapse of time. Two gained scores are then correlated. Then, “in order to arrive at a score by which reliability can be established, one determines the correlation coefficient between the two test administrations” (Mackey & Gass, 2005:129). This type of reliability differs from *mark/re-mark reliability* in the sense that the latter indicates the marking of the same test papers is done by either two or more different testers or by the same tester on different occasions and we still get the same grades or marks.
2. parallel form/equivalent-forms method: administering two cloned tests at the same time to the same test takers. Results of the tests are then correlated.
3. split-half method: a test is divided into two, corresponding scores obtained, the extent to which estimates reliability by grouping questions in a questionnaire that measure the same concept. For example, you could write two sets of three questions that measure the same concept (say class participation) and after collecting the responses, run a correlation between those two groups of three questions to determine if your instrument is reliably measuring that concept. Split-half, Kuder-Richardson 20 and 21, and Cronbach’s are some of the statistical methods to determine reliability. They correlate with each other governing the reliability of the test as a whole.
4. test-retest with equivalent forms: mixed method of test-retest and parallel form. Two cloned tests are administered to the same test takers in different occasion.
5. intra-rater and inter-rater: employing one person to score the same test in different time is called intra-rater. Some hits to minimize unreliability are employing rubric, avoiding fatigue, giving score on the same numbers, and suggesting students write their names at the back of test paper. When two people score the same test, it is inter-rater. The tests done by test takers are divided into two. A rubric and discussion must be developed first in order to have the same perception. Two scores either from intra- or inter-rater are correlated.

Quite naturally, there are some factors that might affect the reliability of a test (Heaton, 1990:162). These are:

- (1) *The Size*: The larger the sample, the greater the reliability,
- (2) *The Administration*: Is the same test administered to different groups under different conditions or at different times?
- (3) *Test Instructions*: Are the test instruction simple and clear enough?
- (4) *Personal Factors*: Motivation, illness, etc.,
- (5) *Scoring the test*: Subjective or objective? (cited in Sariçoban, 2011:399)



**Figure 1:** Factors that affect language test scores

### **2.3.2. Practicality**

Practicality is a primary issue. Validity and reliability are not enough to build a test. Instead, the test should be practical across time, cost, and energy. Dealing with time and energy, tests should be efficient in terms of making, doing, and evaluating. That means a test which is not expensive and easy to administer, stays within appropriate time constraints. Then, the tests must be affordable. It is quite useless if a valid and reliable test cannot be done in remote areas because it requires an inexpensive computer to do it (Heaton, 1975; Weir, 1990; Brown, 2004).

Brown (2001:386) points out that an effective test is practical provided that the value and quality of a test depend on practical considerations. For example, a test which is expensive is impractical. A language proficiency test which requires ten hours to complete is impractical. Sometimes the extent to which a test is practical hinges on whether it is norm-referenced or criterion-referenced. In norm-referenced tests, each test taker's score is interpreted in relation to a mean, median and standard deviation. In criterion referenced tests, lesson objectives are criteria. As Brown and Hudson (2002) suggest, these tests emphasize on teaching and testing matches, focus on instructional sensitivity, curricular relevance, absence of normal distribution restrictions, no item discrimination restriction.

### **2.3.3. Validity**

The test must test what it is intended to test. In other words, test items must be representative of what we intend to test (Köksal, 2004). In short, "the validity of a test is the extent to which it measures what it is supposed to measure and nothing else" (Heaton, 1990:159).

Bachman (1990) asks a crucial question as to "how much of an individual's test performance is due to the language ability we want to measure?" It is validity. Validity links to accuracy. A good test should be valid or accurate. Some experts have defined the term of validity in various ways. Heaton (1975:153), for example, points out that "the validity of a test is the



extent to which it measures what it is supposed to measure.” Bachman (1990:236) also emphasizes that “in examining validity, the relationship between test performance and other types of performance in other contexts is considered.” Messick (1989), for example, describes validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores” (p. 13). (cited in Bachman, 1990). “Validity is not a characteristic of a test, but a feature of the inferences made on the basis of test scores and the uses to which a test is put” as pointed out by Alderson (2002:5). As Gronlund emphasized (1998:226) “it is the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment.” It is on the basis of test scores meaningful, appropriate and useful. In examining validity, we must also be concerned with the appropriateness and usefulness of the test score for a given purpose. It must be reliable at all, however, a reliable test may not be valid at all. Brown (2004:22) defines validity as “the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment.” Similarly, Gronlund and Waugh (2009:46) state that “validity is concerned with the interpretation and use of assessment results.” From these definitions, it can be inferred that when a test is valid, it can elicit students’ certain abilities as it is intended to. The valid test can also measure what it is supposed to measure.

The validity can be measured as non-empirically, involving inspection, intuition and common sense and empirically, involving the collection and analysis of qualitative and quantitative data (Henning, 1987). Validity is a unitary concept (Bachman, 1990; Gronlund and Waugh, 2009). To gain valid inferences from test scores, a test should have some kinds of evidence. The evidence of validity includes face validity, content-related evidence, criterion-related evidence, construct-related evidence, and consequential validity.

### 2.3.3.1. Construct-related Evidence

Messick (1980:1015) defines construct validity as “the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships.” A construct-related evidence, so called construct validity, is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions. Constructs may or may not be directly or empirically measured. Their verification often requires inferential data (Brown, 2004: 25). Messick (1975:957) points out that “a measure estimates how much of something an individual displays or possesses. The basic question [of construct validation] is “what is the nature of that something?” In attempting to answer this question, we must identify and define what the ‘something’ is that we want to measure, and when we define what this is, we are, in effect, defining a construct.

For Carroll (1968), a construct of ‘mental ability’ is defined in terms of a particular set of mental tasks that an individual is required to perform on a given test. Similarly, Cronbach and Meehl (1955) define a construct as “a postulated attribute of people, assumed to be reflected in test performance” (p. 283); further, a construct is defined in terms of a theory that specifies how it relates to other constructs and to observable performance. Thus, constructs can be viewed as definitions of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behavior. Another way of viewing constructs is as a way of classifying behavior. Whenever one classifies situations, persons, or responses, he uses constructs. The term concepts might be used rather than constructs, but the latter term emphasizes that categories are deliberate creations to organize experience into general law-like statements (Cronbach, 1955).

Before an assessment is built, the creator must review some theories about content of it. He then will get new concept related to the content of the items. In language assessment, test makers believe on existence of several characteristics related to language behavior and learning. When the test makers

interpret the results of assessment on basis of psychological constructs, they deal with construct-related evidence (Heaton, 1975; Gronlund and Waugh, 2009). Although it is endless to obtain construct-related evidence, test makers should list from the most relevant ones.” According to Brown (2004), construct validity is a major issue in validating large-scale standardized tests of proficiency. Because such tests must adhere to the principle of practicality, and because they must sample a limited number of domains of language, they may not be able to contain all the content of a particular field or skill.

### **2.3.3.2. Content Validity**

The investigation of content relevance requires “the specification of the behavioral domain in question and the attendant specification of the task or test domain” (Messick 1980:1017). While it is generally recognized that this involves the specification of the ability domain, what is often ignored is that examining content relevance also requires the specification of the test method facets. The importance of also specifying the test method facets that define the measurement procedures is clear from Cronbach’s description of validation:

A validation study examines the procedure as a whole. Every aspect of the setting in which the test is given and every detail of the procedure may have an influence on performance and hence on what is measured. Are the examiner’s sex, status, and ethnic group the same as those of the examinee? Does he put the examinee at ease? Does he suggest that the test will affect the examinee’s future, or does he explain that he is merely checking out the effectiveness of the instructional method? Changes in procedure such as these lead to substantial changes in ability- and personality-test performance, and hence in the appropriate interpretation of test scores. . . . The measurement procedure being validated needs to be described with such clarity that other investigators could reproduce the significant aspects of the procedure themselves.

(Cronbach 1971:449)

The test can have content-related evidence if it represents the whole materials taught before so that the students can draw conclusions from the materials (Weir, 1990; Brown, 2004; Gronlund and Waugh, 2009). In addition, “the test should also reflect objectives of the course” (Heaton, 1975:154). According to Heaton (1975), if the objective of the test is to

enable students to speak, the test should make the students speak communicatively. If the objective of the test is to enable students to read, the test should make them read something. A speaking test which appears in paper-and pencil multiple-choice test cannot be claimed as containing content-related evidence. In relation of curriculum, a test which has content-related evidence represents basic competencies. “Establishing content-related evidence is problematic especially dealing with portion of items representing the larger domain. To build an assessment which provides valid results, a guideline below can be applied” (Gronlund and Waugh, 2009:48-49).

According to Gronlund and Waugh, (2009) the guideline should:

1. Identify the learning outcomes to be assessed (objective of the course),
2. prepare a plan that specifies the sample of tasks to be used (blueprint),
3. prepare an assessment procedure that closely fits the set of blueprint (rubric).

As Bachman (1990) emphasizes, the examination of content relevance and content coverage is a necessary part of the validation process, since the domain specification upon which a test is based provides the means for examining other relationships, such as those between test performance and performance in other contexts. By itself, however, content relevance is not sufficient evidence for validity, since it does not permit inferences about abilities and does not take into consideration how test takers perform.

### **2.3.3.3. Instructional Validity**

Instructional validity is used to refer to the extent to which an assessment is systematically sensitive to the nature of instruction offered. It registers differences in the amount and kind of instruction to which students have been exposed. It is an aspect of larger consequential validity (Yoon, 1996). Messick (1989) introduced the instructional validity. Consequences are a logical part of an evaluation of test use; therefore, examination of effects following from the test use is essential in evaluating test validity (Shephard, 1997). A test that is

instructionally valid, in the sense of being systematically sensitive to differences in opportunity to learn can be further evaluated in terms of its consequential validity- that is, its effectiveness in leading teachers to spend time on classroom activities helpful to learning goals and responsive to individual student learning styles and needs (Darling,1996; Glaser, 1990).

Consequential validity encompasses all the consequences of a test. Weir (1990: 27) calls this evidence as washback validity. It focuses on the effect of tests with regard to specific uses, e.g. its impact to preparation of test- takers, the effect on the learners (positive or adverse effects), or social consequences of test interpretation and use. According to Weir (1990), for teachers, consequential evidence is important. They can judge test scores and use the judgment to improve learning. For stakeholders, this evidence leads to the development of curriculum.

Hublely & Zumbo (1996) point out that “of all the concepts in testing and measurement, it may be argued, validity is the most basic and far-reaching, for without validity, a test, measure or observation and any inferences made from it are meaningless.” They also believe that an observation can be reliable without being valid, but cannot be valid without first being reliable. In other words, reliability is a necessary, but not sufficient, condition for validity while Henning (1987) believes that even an ideal test which is perfectly reliable and possessing perfect criterion-related validity will be invalid for some purposes.

#### **2.3.4. Authenticity**

Bachman and Palmer (1996:23) define authenticity as “the degree of correspondence of the characteristics of a given language test task to the features of a target language task.” Brown (2000) states that if the language in a test is natural, if the items in the test are contextualized, if the tasks in the test are real world tasks, this means that the test is authentic. According to Bachman (1996) authenticity provides a link for investigating the extent to which score interpretations generalize beyond performance on the test, thus, it is linked with construct validity which is an important part of the validation. Besides

this, authenticity is important because it has effect on test taker's perceptions and performance. Authenticity is also pivotal to Douglas' (1997) consideration of specific purpose tests in that it is one of two features which distinguishes such tests from more general purpose tests of language (the other feature being the interaction between language knowledge and specific purpose content knowledge).

According to Carroll (1968:114):

The issue of authenticity must always be an important aspect of any discussion on language testing. A full application of the principle of authenticity would mean that all the tasks undertaken should be real-life, interactive communicative operations and not the typical routine examination responses to the tester's 'stimuli', or part of a stimulus-response relationship; that the language of the test should be day-to-day discourse, not edited or doctored-in the interests of simplification but presented with all its expected irregularities; that the contexts of the interchanges are realistic, with the ordinary interruptions, background noises and irrelevancies found in the airport or lecture-room; and that the rating of a performance, based on its effectiveness and adequacy as a communicative response, will rely on non-verbal as well as verbal criteria.

### **2.3.5. Washback**

Washback is a term generally used in language testing. Washback, commonly used in the field of applied linguistics, refers to "the impact of a test on teaching" (Wall & Alderson, 1993). It refers to the extent to which a test influences language teachers and learners to do things "they would not necessarily otherwise do because of the test" (Alderson & Wall, 1993). The effects of tests on teaching and learning are called washback. Teachers must be able to create classroom tests that serve as learning devices through which washback is achieved. Washback enhances intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment in the students. Instead of giving letter grades and numerical scores which give no information to the students' performance, giving generous and specific

comments is a way to enhance washback (Brown 2004: 29). As Pearson (1988) points out that 'public examinations influence the attitudes, behaviors, and motivation of teachers, learners and parents. And as examinations often come at the end of a course, this influence is seen working in a backward direction, hence the term 'washback'.

### **2.3.5.1. Definitions of Washback**

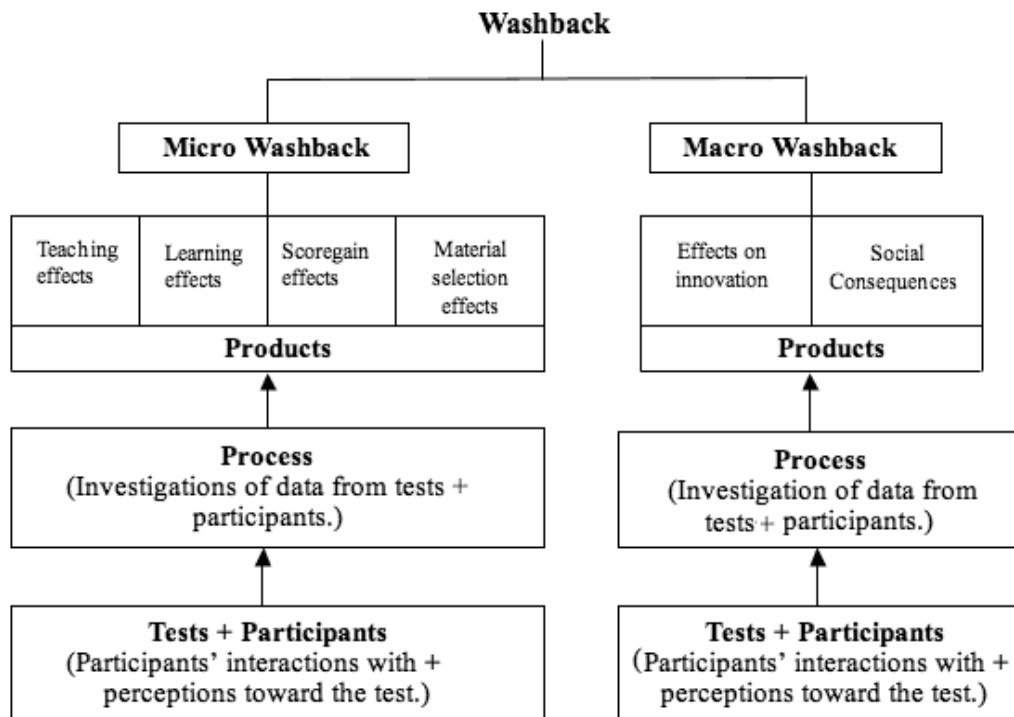
The definition of the word 'washback' is often given as "the effects of tests on teaching and learning." Hughes (2003: 53) defines washback as "the effect of testing on teaching and learning." It generally refers to the effects of the tests on instruction in terms of how students prepare for the tests. (Brown, 2004:28). There is also another definition for washback as Buck describes below:

There is a natural tendency for both teachers and students to tailor their classroom activities to the demands of the test, especially when the test is very important to the future of the students, and pass rates are used as a measure of teacher success. This influence of the test on the classroom (referred to as washback by language testers) is, of course, very important; this washback effect can be either beneficial or harmful (p.17.)

Thus, Buck's definition stresses the impact of a test on what teachers and students do in classrooms (Fullilove, 1992:131). Cohen (1994) also describes the washback as "how assessment instruments affect educational practices and beliefs" while Messick (1996) defines it as "not simply good or bad teaching or learning practice that might occur with or without the test, but rather good or bad practice that is evidentially linked to the introduction and use of the test" (p. 254). Berry (1994) also notes an increased interest in washback with her definition: "one of the major issues within the field of assessment in the 1990s has been a concern with the systemic validity of tests the so-called 'washback effect' or the effect a test has on classroom practice".

Bachman and Palmer (1996) argued that the washback effect of tests operates at two levels: the micro level, which means the effect of tests on teachers and individual students in classroom settings, and the macro level, which refers to the effect of tests on the educational system and society as a

whole. Bachman and Palmer (1996: 35) note that washback is a more complex phenomenon than simply the effect of a test on teaching and learning and they consider washback to be a subset of a test's impact on society, educational systems and individuals. They believe that test impact operates at micro level (i.e. the effect of the test on individual students and teachers); and macro level or the impact the test may have on society and the educational system.



**Figure 2:** A proposed holistic model of washback based on ideas of Hughes (1993), Bachman and Palmer (1996).

Alderson and Wall (1993) also restrict the use of the term 'washback' to classroom behavior of teachers and students and explain that tests are held to be powerful determiners of what happens in classrooms. The washback effect should, therefore, refer to the effects of the test itself on aspects of teaching and learning. Pierce (1992: 687), on the other hand, uses the term 'washback' on the macro level to indicate "the impact of a test on classroom pedagogy, curriculum development, and educational policy". Cohen (1994: 41) also views the macro aspects of washback with regard to "how assessment instruments affect educational practices and beliefs". However, the following studies on,



'washback' cover both the micro level and the macro level: Biggs (1995) uses the term, 'washback' to indicate that testing drives not only curriculum, but also teaching methods and students' approaches to learning. Shohamy, Donita-Schmidt, and Ferman (1996 : 299) explain that "the power and authority of tests enable policymakers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results administrators, teachers, and students". In general, Bailey (1996 : 259) outlines the definition of washback as follows:

- 1) washback is defined as the influence of testing on teaching and learning;
- 2) it is widely held to exist and to be important; but
- 3) relatively little empirical research has been done to document its exact nature or mechanisms by which it works.

### **2.3.5.2. The Origin of Examinations and Washback**

Examinations have long been used as a means of control (Arnové, Altback, & Kelly, 1992; Lai, 1970). Those examinations were probably the first civil service examinations ever developed. Although the goal of the examination was to select civil servants, its washback effect was to establish and control an educational program, as prospective mandarins set out to prepare themselves for the examination that would decide not only their personal fate but also influence the future of the Empire (Spolsky, 1995). The use of examinations to select for education and employment has also existed for a long time. Linn (2000: 4) classified the use of tests and assessments as key elements in relation to five ways of educational reform over the past 50 years: their tracking and selecting role in the 1950s; their program accountability role in the 1960s; minimum competency testing in the 1970s; school and district accountability in the 1980s; and the standards-based accountability systems in the 1990s.

Furthermore, it is clear that tests and assessments are continuing to play a crucial and critical role in education into the new millennium. In spite of this long and well-established place in educational history, the use of tests has,

constantly, been subject to criticism. The researchers such as Baker, 1991; Calder, 1997; Cannell, 1987; Cheng, 1997, 1998a; Heyneman, 1987; Heyneman & Ransom, 1990; Kehaghan & Greaney, 1992; Li, 1990; Shohamy, 1993a; Shohamy, Donitsa-Schmidt, & Ferman, 1996; Widen et al., 1997; and others have, over many years, documented the impact of testing on school and classroom practices, and on the personal and professional lives and experiences of principals, teachers, students, and other educational stakeholders. Aware of the power of tests, policymakers in many parts of the world continue to use them to manipulate their local educational systems, to control curricula and to impose (or promote) new textbooks and new teaching methods. Testing and assessment is “the darling of the policy-makers” (Madaus, 1985) despite the fact that they have been the focus of controversy for as long as they have existed. Shohamy (1992: 513) originally noted that “this phenomenon [washback] is the result of the strong authority of external testing and the major impact it has on the lives of test takers”. High stakes test results are used as an engine to introduce desirable changes in teaching and learning around the world. Davies (1990: 24) asserted that, ‘Testing is always used in teaching, in the sense that much teaching is related to the testing which is demanded of the students’. Hence, tests become an integral part of teaching and learning.

Focusing on the importance of testing, Cheng, L. (1997) wrote that, ‘Traditionally, tests come at the end of teaching and learning process. However, with the advent of high stakes public examinations testing nowadays, the direction seems to be reversed. Testing usually comes before the teaching and learning processes’. Madaus (1988:84, as cited by Spratt, 2005:05) asserted that, ‘it is testing not the official stated curriculum that is increasingly determining what is taught, how it is taught, what is learnt, and how it is learnt’. The teachers and students have been reported to change their teaching and learning strategies according to the demands of tests. Buck (1988: 17 as quoted in Bailey, 1996: 257) asserted that:

“There is a natural tendency for both teachers and students to tailor their classroom activities to the demands of the test, especially when the test is very important to the future of the students, the pass rates are used as

a measure of teacher success. This influence of the test on the classroom is, of course, very important, which is termed as washback”

Petrie (1987:175) concluded that “it would not be too much of an exaggeration to say that evaluation and testing have become the engine for implementing educational policy”. Although washback has only been identified relatively recently, it is likely that washback effects have been occurring for an equally long time. It is also likely that these teaching-testing relationships are likely to become closer and more complex in the future. It is, therefore, essential that the education community work together to understand and evaluate the effects of the use of testing on all of the interconnected aspects of teaching and learning within different education systems.

### **2.3.5.3. Functions and Mechanism of Washback**

Traditionally, tests have come at the end of the teaching and learning process for evaluative purposes. However, with the widespread expansion and proliferation of high-stakes public examination systems, the direction seems to have been largely reversed. There is often a distinction in the literature on assessment between high- and low-stake tests (Madaus, 1988): ‘high’ is defined as situations when admission, promotion, placement or graduation are directly dependent on test scores while ‘low’ implies the opposite. Testing can come first in the teaching and learning process. In addition to these changes, many more changes in the teaching and learning context can occur as the result of a new test. Such influences were linked to test validity by Shohamy (1993a: 2), who pointed out that “the need to include aspects of test use in construct validation originates in the fact that testing is not an isolated event; rather, it is connected to a whole set of variables that interact in the educational process”. Similarly, Linn (1992: 29) encouraged the measurement research community “to make the case that the introduction of any new high-stakes examination system should pay greater attention to investigations of both the intended and unintended consequences of the system than was typical of previous test-based reform efforts”. As a result of this complexity, Messick

(1989) recommended a unified validity concept, which requires that when an assessment model is designed to make inferences about a certain construct, the inferences drawn from that model should not only derive from test score interpretation, but also from other variables operating within the social context (Bracey, 1989; Cooley, 1991; Cronbach, 1988; Gardner, 1992; Gifford & O'Connor, 1992; Linn, Baker, & Dunbar, 1991; Messick, 1992). The importance of collaboration was also highlighted by Messick (1975: 959):

Researchers, other educators, and policy makers must work together to develop means of evaluating educational effectiveness that accurately represent a school or district's progress toward a broad range of important educational goals.

In exploring the mechanism of such an assessment function, Bailey (1996: 262-264) cited Hughes' trichotomy (1993) to illustrate the complex mechanisms through which washback occurs in actual teaching and learning environments. Hughes (1993:2) explained his model as follows: a) the nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks; b) these perceptions and attitudes in turn may affect what "the participants" do in carrying out their work "the process", including practicing the kind of items that are to be found in the test; c) these, in turn, will affect the learning outcomes, "the product" of the work. Wall (1996:334) stressed the difficulties in finding explanations of how tests exert influence on teaching. Wall (1999, 2000) used the innovation literature and incorporated findings from this literature into her research areas to propose ways of exploring the complex aspect of washback: - The writing of detailed baseline studies to identify important characteristics in the target system and the environment, including an analysis of the current testing practices (Shohamy et al., 1996), current teaching practices, resources (Bailey, 1996; Stevenson & Riewe, 1981), and attitudes of key stakeholders (Bailey, 1996; Hughes, 1993).

Fullan explained that the "subjective reality" which teachers' experience would always contrast with the "objective reality" that the proponents of change had originally imagined. According to Fullan, teachers work on their own, with

little reference to experts or consultation with colleagues. They are forced to make on-the-spot decisions, with little time to reflect on better solutions. They are pressured to accomplish a great deal, but are given far too little time to achieve their goals. This may help to explain why intended washback does or does not occur in teaching and learning. If educational change is imposed upon those parties most directly affected by the change, that is learners and teachers, without consultation of those parties, resistance is likely to be the natural response (Curtis,2000). Andrews (1994,1995) highlighted the complexity of the relationship between washback and curriculum innovation, and summarized three possible responses of educators in response to washback: fight it, ignore it, or use it (cited in Heyneman, 1987:260). By “fight it,” Heyneman referred to the effort to replace examinations with other sorts of selection processes and criteria, on the grounds that examinations have encouraged rote memorization at the expense of more desirable educational practices. In terms of “ignoring it,” Andrews (1994: 51-52) used the metaphor of the ostrich pretending that oncoming danger does not really exist by hiding its head in the sand. The third response, “use it,” is now perhaps the most common of the three, and using washback to promote particular pedagogical goals is now a well-established approach in education.

#### **2.3.5.4. Negative Washback**

Washback can also be negative and positive (Saehu, 2012: 124-127). It is easy to find negative washback such as narrowing down language competencies only on those involve in tests and neglecting the rest. While language is a tool of communication, most students and teachers in language class only focus on language competencies in the test. It is usually thought that language tests have negative influence on teaching and learning, that’s why, this influence is called negative washback. This has been considered as a potential problem. It is thought that teachers showed tendency in ignoring the subjects and activities that did not help passing the exam and that examinations or tests ‘distort the curriculum’ (p.166). Negative washback is commonly described as the phenomenon in which teachers drop curriculum and teach

toward tests. To explain situations of negative washback, Wall (1997) describes 'principles' that Madaus (1988) presents about the impact of testing as follows:

The power of tests is a perceptual phenomenon, the higher the stakes attached to a test the more it will distort the teaching process, past exam papers eventually become the teaching curriculum, teachers adjust their teaching to fit the form of exam questions, test results become the major goal of schooling, and the agencies which set or control examinations eventually assume control over the curriculum (cited in Wall, 1997: 292).

Fish (1988) discovers that "teachers reacted negatively to pressure created by public displays of classroom scores" (cited in Cheng, 2000:9). Noble and Smith (1994a:6) also found that high-stakes testing could affect teachers directly and negatively, and that "teaching test-taking skills and drilling on multiple-choice worksheets is likely to boost the scores but unlikely to promote general understanding". Buck (1992: 141) expresses his opinion about the negative effects of tests on teaching when he states that "it seems likely that translation tests could have very negative washback indeed, and lead to activities which would not be beneficial to second language learners" (cited in Watanabe, 1996:319). However, his opinion is criticized as mere self-report without results of systematic empirical research (Watanabe, 1996).

In order to explain 'negative washback', Alderson and Wall (1993:115) cite Vernon's (1956:166) comment that "teachers tend to ignore subject and activities which are not directly related to passing the exam so that examinations distort the curriculum". Alderson and Wall (1993) referred to negative washback as the undesirable effect on teaching and learning of a particular test seemed to be "poor" (p. 5). Alderson and Wall meant by saying poor "something that the teacher or learner does not wish to teach or learn." Alderson and Wall (1993), stressed that the quality of the washback effect might be independent of the quality of a test (pp. 117–118). Noble and Smith (1994a) also found that "teaching test-taking skills and drilling on multiple-choice worksheets is likely to boost the scores but unlikely to promote general understanding" (1994, p. 6). According to Pearson (1988), a test's washback effect will be negative if it fails to reflect the learning principles and course objectives to which the test supposedly relates, and it will be positive if the

effects are beneficial and “encourage the whole range of desired changes” (p. 101). As Heyneman (1987) put it: “Testing is a profession, but it is highly susceptible to political interference. To a large extent, the quality of tests relies on the ability of a test agency to pursue professional ends autonomous” (p. 262).

#### **2.3.5.5. Positive Washback**

If the effects of the tests on teaching and learning process can be considered as ‘positive’ then we can talk about ‘positive washback’. Positive washback is, according to Messick, linked to authentic and direct assessments and to the need to minimize construct under-representation and construct irrelevance in the test. Tests are encouraged to promote the idea of lifelong learning and encourage people to learn English (Language Testing and Training Centre, 2008). Good tests can be utilized and designed as beneficial teaching-learning activities so as to encourage a positive teaching-learning process (Pearson, 1988:107). According to Messick (1996), “a poor test may be associated with positive effects and a good test with negative effects because of other things that are done or not done in the education system” (p. 242). Wall and Horak (2008) focus on the role of communication in creating positive washback. They found that teachers usually do not understand the nature of tests and encourage testers to communicate their intentions so that teachers and learners can prepare for new kinds of assessment.

Davies (1985) takes the view that a good test should be “an obedient servant of teaching; and this is especially true in the case of achievement testing” (cited in Cheng, 2000:9). Pearson (1988:107) considers that good tests will be more or less directly usable as teaching-learning activities. Similarly, good teaching-learning tasks will be more or less directly usable for testing purposes, even though practical or financial constraints limit the possibilities (cited in Alderson & Wall, 1993). Crooks (1988) discusses the influence that evaluation activities in class can have on students, proposing possible situations in that testing can have a positive effect on them as follows: teachers stress the need for ‘deep learning’ rather than ‘surface learning’, use evaluation

to assist students rather than to judge them, use feedback to focus students' attention on their progress set high but attainable standards, and select evaluation tasks to suit the goals being assessed (cited in Wall, 1997:292).

According to Hughes (1989:2), backwash-washback can be harmful or beneficial; however, "if testing always had a beneficial effect on teaching, it would have a much better reputation amongst teachers". For this reason, he suggests seven ways to achieve beneficial backwash:

1. Test the abilities whose development you want to encourage.
2. Sample widely and unpredictably.
3. Use direct testing.
4. Make testing criterion-referenced.
5. Base achievement tests on objectives.
6. Ensure test is known and understood by students and teachers.
7. When necessary, provide assistance to teachers.

Bailey (1996) also suggests some factors which might promote beneficial washback such as language learning goals, authenticity, learner autonomy and self-assessment, and detailed score reporting.

### **2.3.5.6. Measurement-driven Instruction and Curriculum Alignment**

Washback (Alderson and Wall, 1993), with other similar terms such as backwash (Biggs, 1995, 1996), *test impact* (Bachman and Palmer, 1996), *systemic validity* (Fredericksen and Collins, 1989), *consequential validity* (Messick, 1989, 1996), *measurement-driven instruction* (Popham, 1983, 1987), and other possible terms refer to the influence of testing on teaching and learning. As tests and testing in language learning have gained much more importance than ever, washback has been used widely in language testing and applied linguistics. Measurement-driven instruction (Shohamy 1992:15). Shohamy contended that "the use of external tests as a device for creating the



educational process is often referred to as the washback effect or measurement-driven instruction".(Shohamy et al 1996:6) Shohamy et al. defined curriculum alignment as "the curriculum is modified according to test results". High-stakes testing to achieve the goals of teaching and learning, such as the introduction of new textbooks and new curricula (Shohamy, 1992:15; Wall & Alderson 1993 :1; Cheng; 2005 :8). Tests are encouraged to promote the idea of lifelong learning and encourage people to learn English (Language Testing and Training Centre, 2008).

Although the number of studies on washback have increased considerably since the seminal work of Alderson and Wall (1993) it is still not clear how testing influences teaching and learning. When we look at the history of empirical studies on washback, as a first study the seminal work of Alderson and Wall (1993) which took place in Sri- Lankan secondary schools for the Sri-Lankan O-Level Evaluation exam is seen. The effects of changing the O-level examinations had been searched in this study. Alderson and Wall (1993), in their Sri Lankan study, dealt with the aspects of teaching and learning and came up with 15 hypotheses thinking that this process may be influenced by the examinations. These hypotheses, regarding washback are listed below:

1. A test will influence teaching.
2. A test will influence learning
3. A test will influence what teachers teach;
4. A test will influence how teachers teach;
5. A test will influence what learners learn;
6. A test will influence how learners learn
7. A test will influence the rate and sequence of teaching; and
8. A test will influence the rate and sequence of learning.
9. A test will influence the degree and depth of teaching;
10. A test will influence the degree and depth of learning.
11. A test will influence attitudes to the content, method, etc. of teaching and learning.
12. Tests that have important consequences will have washback; and conversely

13. Tests that do not have important consequences will have no washback
14. Tests will have washback on all learners and teachers.
15. Tests will have washback effects for some learners and some teachers, but not for others.

After this study and the hypothesis, Alderson and Wall (1993) concluded that further research on washback is needed to be able to understand the washback better. And then, Shohamy (1996) in secondary schools in Israel researched English foreign language test and Arabic second language test to see the impact of the tests on teaching and learning. She wrote in her report that these tests served as "an effective tool for changing the behaviours of teachers and students because of their power and high stakes". The same year Alderson and Hamp Lyons did some researches in a language school where TOEFL was studied for university entrants. And Watanabe, (1996; 2000) did a research in private institutions where the students were being prepared for the university entrance exams in Japan. Madaus,(1988 p.22) also concluded that Measurement-driven instruction will definitely result in cramming, narrowing the curriculum, focus of attention on those skills that are most relevant to testing, placement of constraints on teachers' and students' creativity and spontaneity, and disparage the Professional judgment of educators.

#### **2.3.5.7. Studies Investigating Washback Effects**

It is a common belief that testing affects teaching and learning, as stated by Alderson and Wall (1993:1) that "tests are held to be powerful determiners of what happens in classroom". Tests are encouraged to promote the idea of lifelong learning and encourage people to learn English (Language Testing and Training Centre, 2008). The way in which examinations influence teaching and learning is commonly described as "washback" or "backwash". The aim of the research may be to investigate how tests influence teachers' internal factors such as personal beliefs about teaching, motivation or how they influence students, their learning or their personal feelings, or how they influence both. Also, the research may investigate the effects of the examination on materials such as course books. Decisional makers use the authority power of high-

stakes testing to achieve the goals of teaching and learning, such as the introduction of new textbooks and new curricula (Shohamy, 1992:15; Wall & Alderson 1993:1; Cheng; 2005:8).

In order to gather data from teachers and students, it may be possible to administrate interviews or questionnaires. In addition, classroom observation is significant at this point because an attempt should be made to establish credibility or to demonstrate “that the research was conducted in a way that maximizes the accuracy of identifying and describing the object(s) of study” (Brown, 2001:225). To carry out an observation study, a set of data-gathering instruments, such as observation instruments, pre observation instruments recording classroom events, and post observation interviews,, needs to be constructed.

Another way to gather data is interview with teachers. The researcher may have pre-observation interview before recording classroom events and then have post-observation interviews. A valuable piece of information, such as teachers’ personal beliefs about education, may also be obtained through casual conversations with teachers (Watanabe, 2001: 30). Spratt (2005) has stated that the teacher plays a significant role in determining the types and intensity of washback, and thus, teachers have become the sources of promoting positive washback. Chapman and Snyder (2000:462) have expressed a similar view by stating that “washback is not the examination itself that influences teachers’ behavior, but teachers’ beliefs about those changes”. As Watanbee (2005) suggested, teachers should be provided with in-service training and be familiar with a wide range of teaching methods.

#### **2.3.5.8. Studies Conducted on Washback in Turkey**

Although there are not many, there are still a few studies conducted on washback effects of language examinations within English as a Foreign Language context, in Turkey.

Osken (1999) investigated the content validity and backwash effect of the end-of-term Oral Assessment Test (OAT) administered at Hacettepe University,

Department of Basic English. The end-of-term OAT is a final achievement test used to measure students' oral language abilities. The content validity of the OAT was investigated in terms of consistency between the learning goals set for the students in the course book content and taught in the language program and the content of the OAT. A related issue to the content validity was the backwash effect of the OAT, which is the effect of the test on teaching and learning in the classroom. This study included three groups of subjects: 14 B-level subject teachers and two testers, 62 B-level students and three administrators. To gather data, questionnaires were given to the three groups of subjects mainly to obtain their opinions about the course book content and the content of the OAT. The results of the documentary analysis of the types of speaking tasks both in the course book content and content of the OAT showed that although there were 13 types of speaking tasks occurring in the course book, only three of them were on the OAT. This resulted in a low degree of the content validity of the OAT. The results of the questionnaires supported the findings of the documentary analysis above indicating that the majority of the speaking task types in the course book were not included and tested in the OAT, which proved inconsistency to a certain extent.

Ari (2002) carried out a study examining the effects of changes made in university examination system on the education in chemistry department in faculty of science and arts.

Boylug (2003) investigated the agreement between the opinions of the teachers and students related to the reading activities practised in the English as a Foreign Language classes at the Foreign Language Track of Foreign Language Oriented High Schools in Gaziantep, Turkey. It also aimed to see how efficiently the teachers prepared their students for the Foreign Language Examination, a reading-based examination, by employing EFL reading activities. The teachers and the students of the high schools were administered questionnaires to gather their opinions. The results indicated that although there were no statistically significant differences between the teachers and students' opinions for most of the items, the classroom application frequencies for almost all the items were quite low. The interpretation of these results revealed that these activities were not conducted efficiently, and even more important, that

the students were not taught strategies which were expected to help them to study independently.

Sevimli (2008) carried out a study to investigate whether there was a washback effect from the FLE (the Foreign Language Examination) on the teaching and learning of FLE classrooms in three types of high schools; three Anatolian High Schools, two Private high Schools and one Super High School in Gaziantep, Turkey. These schools were only school types with FLE groups. The first step of the study was to find out whether the FLE had washback effects on teaching/learning activities in FLE classrooms and also whether its effect would show differences among three types of high schools. The results indicated that there was a negative washback effect of the FLE on EFL teaching and learning in secondary schools.

Duran (2011) carried out a study to investigate teachers and students perceptions of the washback effects of classroom-based speaking tests with teachers' and students' attitudes towards and beliefs about teaching and testing speaking. The data were collected through teacher and student questionnaires and teacher and student interviews. The results revealed that teachers stated that they were not influenced by the speaking tests in terms of what they did in classes, but they had positive attitudes towards teaching and testing speaking and they believed that speaking tests had a positive effect on their students' speaking ability. Teachers and students believed that getting ready for speaking tests improved the general speaking skills of students.

Şentürk (2013) carried out a study to show the washback effect of international exams in learning English as a foreign language and to seek answers for the research questions: 1. what is the nature and the scope of washback effects of the KET preparation program on classroom practices? 2. What differences can be seen in KET preparation programs regarding classroom interactions? 3. What do the teachers think about preparing the students for an international exam? 4. What do the students think about getting prepared for an international exam? The results of this study aimed to show the nature and the scope of washback effects of the KET preparation program on classroom practices and interactions including the teachers' and students'

thoughts about this preparation process. It was found that the washback effect varied in different situations. The washback effect on classroom interactions and practices was observed because the teacher changed her teaching method taking the exam format into consideration. As the exam format was based on the communicative skills, this effect could be considered as positive.

### **2.3.5.9. Washback Effect of Examinations in Overall Education**

Many educationalists have written about the power of examinations over what takes place in the classroom (Vernon 1956; Davies 1968; Madaus and Airasian 1982; Alderson 1986; Morrow 1986). Pearson for example says “it is generally accepted that public examinations influence the attitudes, behavior, and motivation of teachers, learners and parents (1988:98) This influence is often seen negative. Vernon (1956:166) claimed that examinations “distort the curriculum” but see the washback in a more positive way.

Elton and Laurillard (1979) summarized the strategy very succinctly: “The quickest way to change student learning is to change the assessment system” (p. 100, as cited in Tang & Biggs, 1996, p.159). Washback and the impact of tests more generally have become a major area of study within educational research, and language testing in particular. Therefore, most of the studies conducted on washback are on language examinations. However, there are still some studies conducted on education in general as in the following:

In his study of teachers’ beliefs about the influence of testing on the classroom practices, Madaus (1988) compared the content of the actual tests and the content of tests in the textbook in order to examine whether or not both reflected what the curriculum said. It was found that both failed to measure what the curriculum indicated such that students should be able to know and do at certain levels.

Haas, Haladyna, and Nolen (1989: 8) conducted research into the effects of external testing on teachers in junior high schools. They collected data through questionnaires and teacher interviews. The study revealed that teachers believed the test scores were “routinely inappropriately used” to

evaluate teachers and that such inappropriate uses had harmful effects on their teaching.

In a qualitative study about the effect of external testing in elementary schools in Arizona, Smith (1991) reported that teachers had negative feelings such as great anxiety, shame, and embarrassment related to their students' test results and believe that the test scores were used against them, despite the perceived invalidity of the scores. In addition, Cheng's (1997) study embodied both teacher and student opinions. She used questionnaires for teachers and students, teacher interviews, and classroom observations to examine how the revised Hong Kong Certificate of Education Examination (HKCEE) influenced secondary school teaching. She reported that the examination had the most 'intensive' washback effect on the contents of teaching so that fast changes occurred in teaching materials, which was due largely to the commercial characteristics of the Hong Kong society and washback effect worked slowly and reluctantly and with difficulties in the methods teachers employed.

Cheng (1998, 1999) conducted a follow-up study that focused on how the revised HKCEE influenced secondary school teaching. She (1998) reported the impact of the examination change on student perceptions and attitudes toward their learning. The findings from the questionnaires indicated that although more teaching and learning activities were similar to the examination activities over two years, in which the follow-up study was conducted, student perceptions and attitudes toward the aspects of the examination remained unchanged. Cheng (1999) also reported washback on teacher perceptions and actions by observing three teachers over the two years.

In Sri Lanka, Wall and Alderson (1993) investigated the effects on Sri Lankan classes of changing the O level English examination. These changes were intended to reinforce innovations in textbook materials and teacher training courses. They conducted a two year longitudinal observational study.

### **2.3.5.10. Washback Effect of Examinations in Foreign Language**

#### **Classrooms and Programs**

The study of wash back has resulted in recent developments in language testing, and measurement-driven reform of instruction in general education. Research in language testing has centered on whether and how we assess the specific characteristics of a given group of test takers and whether and how we can incorporate such information into the ways in which we design language tests. Language test scores cannot be interpreted simplistically as an indicator of the particular language ability we think we are measuring. The scores are also affected by the characteristics and contents of the test takers, the characteristics of the test takers, the strategies the test takers employ in attempting to complete the test tasks, as well as the inferences we draw from the test results. These factors undoubtedly interact with each other (Cheng, Watanabe, Curtis, 2000:4-5).

Alderson (1986) identified washback as a distinct area within language testing, to which researchers needed to turn our attention. Alderson (1986:104) discussed the “potentially powerful influence offsets” and argued for innovations in the language curriculum through innovations in language testing. Davies (1985) stated that tests should necessarily follow the curriculum, and suggested that perhaps tests ought to lead and influence the curriculum. Morrow (1986: 6) extended the use of washback to include the notion of washback validity, which describes the relationship between testing, and teaching and learning.

Alderson and Wall (1993: 120-121), in their Sri Lankan study, attempted to do in establishing baseline data through observations of English classes in Sri Lankan secondary schools prior to the implementation of an innovative test. Alderson and Wall concluded that further research on washback is needed, and that such research must entail “increasing specification of the Washback Hypothesis”. These hypotheses regarding washback from their review of the literature on language testing and their own experience of discussing with teachers about their teaching and testing are as follows:

1. A test will influence teaching.



2. A test will influence learning.
3. A test will influence what teachers teach; and
4. A test will influence how teachers teach; and by extension from above,
5. A test will influence what learners learn; and
6. A test will influence how learners learn.
7. A test will influence the rate and sequence of teaching; and
8. A test will influence the rate and sequence of learning.
9. A test will influence the degree and depth of teaching; and
10. A test will influence the degree and depth of learning.
11. A test will influence attitudes to the content, method, etc. of teaching and learning.
12. Tests that have important consequences will have washback; and conversely.
13. Tests that do not have important consequences will have no washback.
14. Tests will have washback on all learners and teachers.
15. Tests will have washback effects for some learners and some teachers, but not for others.

Spratt (2005) reviews the empirical studies of washback from external examinations and tests that have been carried out in the field of English language teaching from the point of view of the teacher so as to provide teachers with a clearer idea of the roles they can play and the decisions they can make concerning washback. What intervening factors the studies have indicated influence whether and to what degree washback occurs are examined. This examination highlights how much washback cannot be considered an automatic or direct effect of examinations. As a result, this study shows how crucial a role the teacher plays in determining types and intensity of washback, and how much teachers can therefore become agents for promoting positive washback.

Watanabe (1996) observes the classroom practice of two different English exam-preparation classes taught by two experienced teachers: one of each teacher's exam-preparation classes is grammar-translation oriented and the other is not. From the classroom observations, it is found that translation

oriented university entrance examinations do not influence the two teachers in the same way; that is, the examinations induce washback on one teacher, but not on the other.

In a study conducted by Li (1990), again teachers and also administrators were participants, but students' views and opinions were not involved. The Matriculation English Test (MET; the reformed English test for entrance to all universities in China) is an example that undoubtedly shows the existence of washback effects on the teaching of English throughout China.

The studies above include information provided by teachers, in which only teachers' views and beliefs are considered but do not encompass student views and beliefs. However, the research conducted in Israel by Shohamy, Donitsa-Schmidt, and Freeman (1996) on the long-term washback effect includes both teacher and student perceptions. Through document analysis, questionnaires, and interviews with teachers, students, and language inspectors, they investigate the long-term impact of two national tests that have been implemented in the late 1980's. One is Arabic as a second language (ASL) and the other is English as a foreign language (EFL). Results show that there are different washback patterns for the two tests: whereas the impact of the EFL test, which is a high-stakes test, has increased, the washback effect of the ASL test, which is a low-stake test, has significantly decreased over the years.

Alderson and Hamp-Lyons (1996), in a washback study of TOEFL preparation courses in the United States, also consider both teacher as well as student views. They compare TOEFL preparation classes and non-TOEFL preparation classes by the same teachers as well as the teachers' behaviors in both types of classes through the use of three kinds of instruments: student interviews, teacher interviews, and classroom observations. This study shows that the TOEFL test affects both what and how teachers teach, but the degree and kind of influence vary from teacher to teacher.

Wall and Alderson (1993:41-68) investigate the impact of a secondary school English examination in Sri Lanka on language teaching. In order to determine whether the examination has an effect on teaching, they focus on the relationship between the examination and the textbook, that is whether the examination is intended to reinforce the textbook. The findings from the study indicate that the examination impacts on what teachers teach but not on how they teach.

In the studies above the examinations whose wash back effects have been investigated are language examinations in general. However, there are also some other studies that handle the examinations evaluating only one specific skill of the students on English language. Two studies below are examples of this type of examinations.

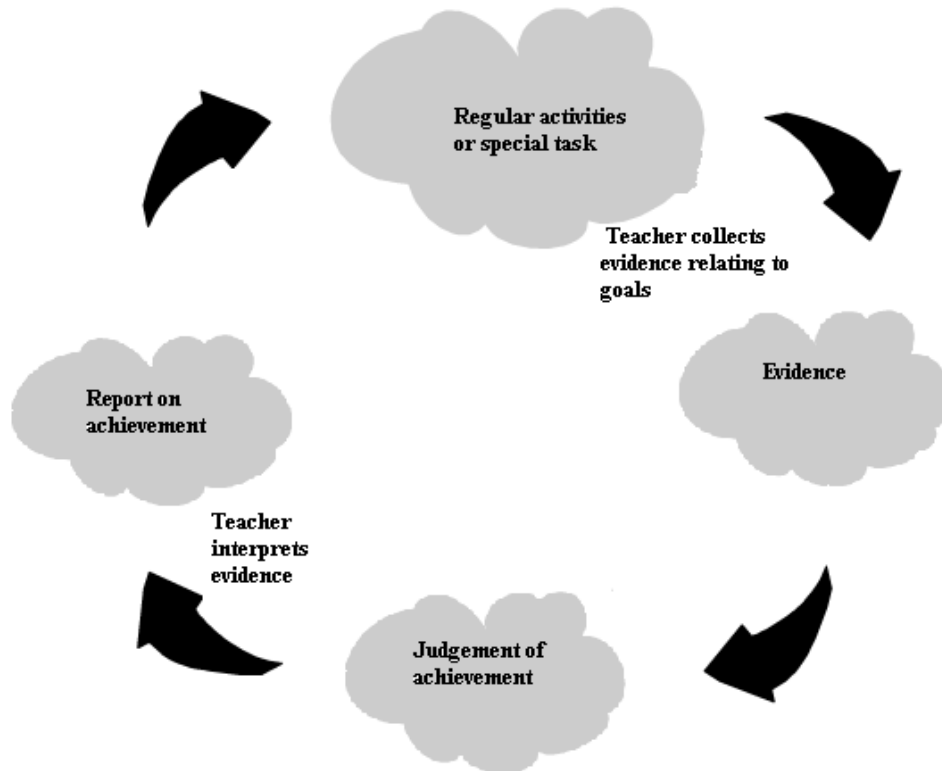
Stecher, Chun, and Barron (1999) conduct two statewide surveys-of Washington principals and teachers- to study the impact of the Washington educational reform on school and classroom practice. The teachers' reports about writing curriculum and instruction and data about school practices from the principal surveys when trying to model the impact of the reform on Washington Assessment of Student Learning (WASL) scores. The WASL test in writing achieves more than a multiple-choice test of writing would do, because students must produce an essay, not merely fill in blanks, identify mistakes, or complete other writing- related tasks that can be assessed using a multiple-choice format. Although the standards-based, test-driven reform adopted in Washington has reduced the extent of the "washback" effect of testing on instruction, it has not eliminated the effect altogether.

Freeman (1996) examines the washback effects of a national EFL oral matriculation test, introduced by the Ministry of Education into the Israeli educational system. The study attempted to find whether this high-stakes test affected the educational processes, the participants, and the products of teaching and learning, and if so, how; it attempted to find whether the washback of the examination innovation corresponded very closely to the effect intended by the policymakers. Following four types of instruments were used: structured

questionnaires completed by students, structured interviews held with teachers, open interviews held with three regional inspectors, and Document Analyses of the Director General Bulletins and instructions issued by the Chief Inspector for English were performed to investigate the intentions of the test designers. In conclusion, the EFL oral matriculation test resulted in strong washback on the educational processes, the participants and the products of teaching and learning in Israeli high schools.

## **2.4. SUMMATIVE ASSESSMENT**

According to Ciel, (2000), summative assessment is designed to get feedback about overall judgement at the end of a course of learning and used to grade the learners' products of learning. It is an assessment activity which results in a mark or grade to judge the student performance (Irons, 2008). Summative assessment is conducted to monitor and record student achievement (McMillan, 2007). Since summative assessment aims to measure or summarize what a student have learnt, it occurs at the end of a course or unit of instruction but it does not focus on the future progress (Brown, 2004). It only helps the teachers in organizing their courses because summative assessment shows whether program goals and objectives have been met or not as well.

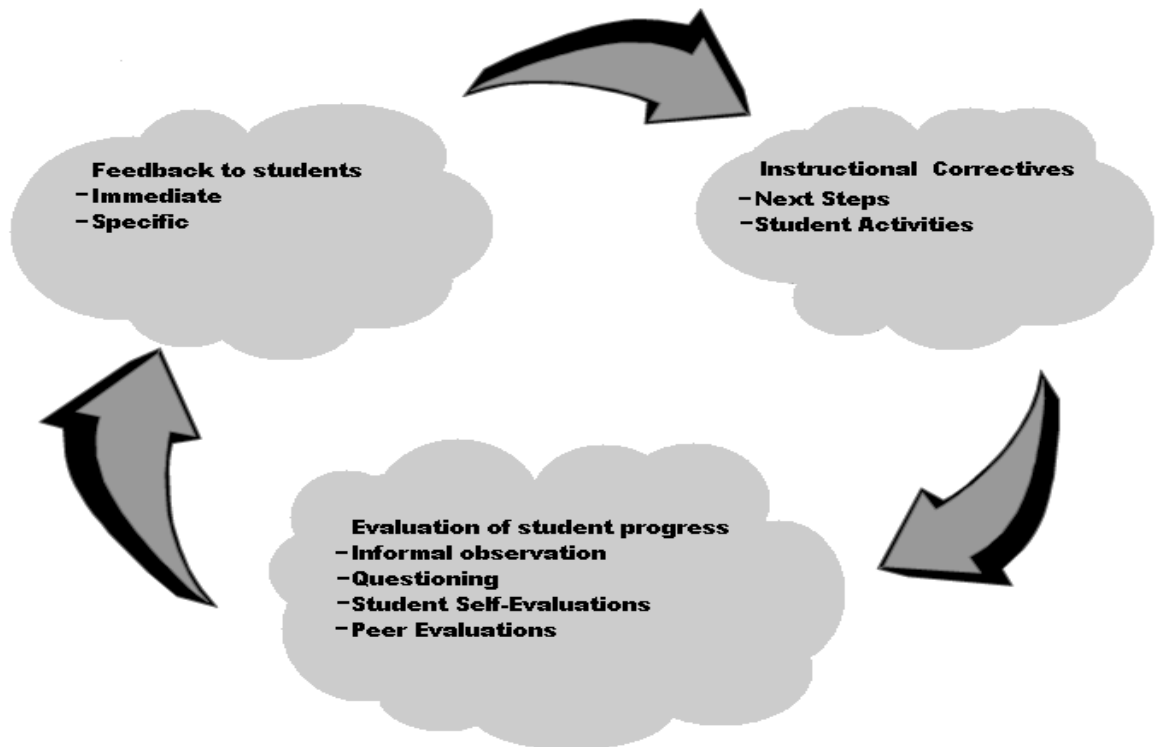


**Figure 3:** Cycle of Summative Assessment (Harlen, 2003:87)

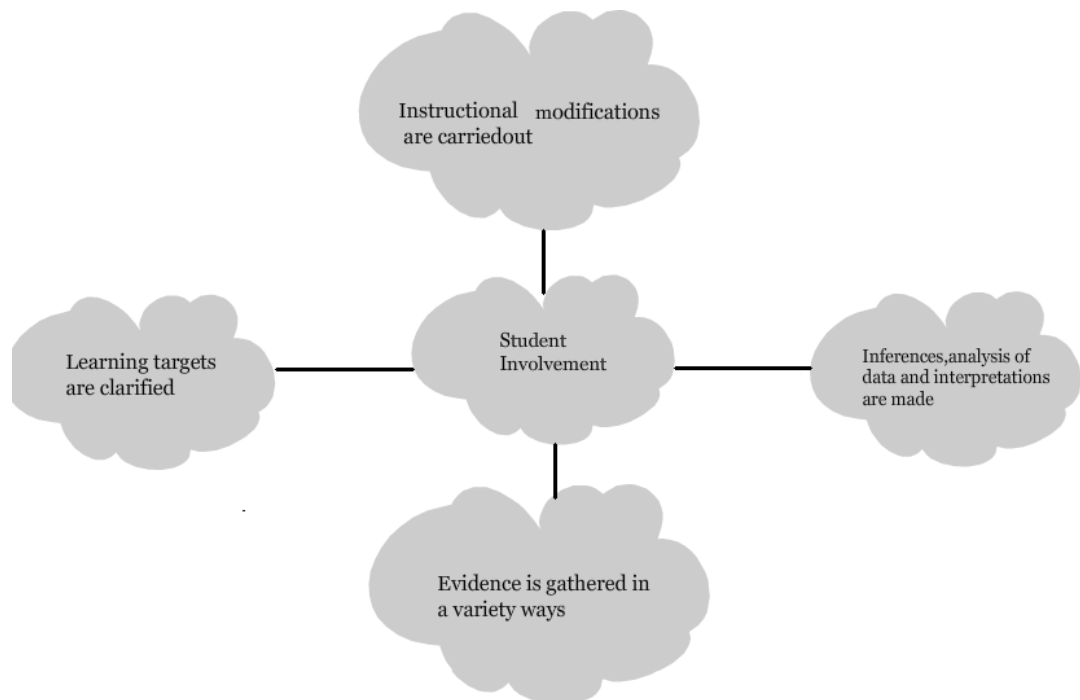
## 2.5. FORMATIVE ASSESSMENT

Formative assessment not only gives teachers information for instructional decisions but also gives students information for improvement (Brookhart, 2001). Formative assessments are used to improve instructional methods and provide student feedback throughout the teaching and learning process are ongoing assessments, reviews, and observations in a classroom (Frey and Fisher, 2007). They are used to check where the students are and what should be done for the better. By actively involving the students in this process, the teachers are able to meet individual students' needs and help them reach high standards (Policy Brief, 2005). Taking place during the course of learning, formative assessment involves the teacher in gathering evidence about students' skills, concepts, and attitudes relevant to the goals of learning (Harlen, 2003). With the help of formative assessment, teachers know how

students are progressing and where they are having trouble (Carol, 2002). It also helps the students to close the gap between where they are now and where they should be (Stiggins, 2005).



**Figure 4:** Formative Assessment Cycle (McMillan, 2007:7 )



**Figure 5:** Cycle of Formative Assessment (Harlen, 2003: 507)

## **2.6. APPROACHES IN LANGUAGE TESTING**

### **2.6.1. Integrative Approach**

The tests which were designed on the basis of this approach were described as holistic, integrative, pragmatic, sociological, subjective, yet and more importantly were characterized by two features: 1) assessing practical language skills of foreign learners who wish to be enrolled in the English speaking community universities-say UK and USA, and 2) both skills and sub-skills of language were integrated to achieve the communication purpose which was not achieved by the earlier approaches especially the psychometric-structuralists approach, (Madsen, 1983, McNamara, 2000, McNamara & Roever (unknown) and Weir, 2005). A major proponent of this approach is Oller who has introduced his influential hypothesis (Unitary Competence Hypothesis), (Madsen, 1983, McNamara, 2000, McNamara & Roever (n.d.) and Weir, 2005).

Integrative approach involves the testing of language in context and is thus concerned primarily with meaning and the total communicative effect of discourse. Integrative testing involves functional language but not the use of functional language. It is designed to assess the learner's ability to use two or more skills simultaneously. It is best characterized by the use of cloze testing as it is a good indicator of general linguistic ability, including the ability to use language appropriately according to particular linguistic and situational contexts and of dictation. The use of cloze test, dictation, oral interview, translation and essay writing are included in many integrative tests. This approach involves the testing of language in context and is thus concerned primarily with meaning and the total communicative effect of discourse. Integrative tests are concerned with a global view of proficiency.

One more drawback of this approach is its unreliability and invalidity in the case of translation tests, because many skills and sub-skills are required not only from the target language but also from the native tongue language, that is,

the source language, (Madsen, 1983, McNamara, 2000, McNamara & Roever (n.d.) and Weir, 2005).

### **2.6.2. Communicative Approach**

Communicative approach emphasizes the importance of the meaning of utterances rather than their form and structure. Communicative tests are concerned primarily, if not totally, with how language is used in communication. Language use is often emphasized to the exclusion of language usage. Use is concerned with how people actually use language for a multitude of different purposes; usage concerns the formal patterns of language (described in prescriptive grammars and lexicons). According to this theory and approach, knowing language is more than knowing its rules, (Madsen, 1983, McNamara, 2000, McNamara & Roever (n.d.) and Weir, 2005). Above all, tests which are designed following this approach are characterized by two features: 1) learners are assessed with the use of performance tests on the basis of communication acts they perform be it receptive or productive, 2) social roles must be integrated in any test, (Madsen, 1983, McNamara, 2000, McNamara & Roever (n.d.) and Weir, 2005). Bachman is a clear example and a proponent of such an approach, (Bachman, 1990; Bachman & Palmer 1996). Language testing constantly involves making compromises between what is ideal and what is practicable in a certain situation. As a communicative test can measure all language skills, it can help students in getting the score.

#### **2.6.2.1. Characteristics and Types of Tests in Communicative Approach**

- Communicative tests are concerned primarily with how language is used in communication
- Language use is often emphasized to the exclusion of language usage.
- The attempt to measure different language skills in communicative tests is based on a view of language referred to as divisibility hypothesis



- The test content should totally be relevant for a particular group of examinees and the tasks set should relate to real-life situation
- Communicative testing introduces the concept of qualitative modes of assessment in preference to quantitative modes of assessment (Tunçel, 1995:34-35; Çaykan, 2001:36-37; Davies, 1995:9-10).

### **2.6.2.2. Strengths of Communicative Approach**

Communicative tests are able to measure all integrated skills of students. The tests using this approach face students in real life so it will be very useful for them. In such tests, meaningful and realistic interaction is required. At least two participants, comprehension, feedback and use of all four skills are all essential. Such tests are criterion-referenced; that is, the testees are assessed according to their performances; whether or not they have performed a certain task properly, not in accordance with other testees' performances.

This testing approach should go hand in hand with a communicative language teaching program. The learners should be provided with a communication oriented foreign language program in order that they can be tested communicatively. Via this program, the learners can have the opportunity to practise and to have experience required to meet their needs in the TL. (Tunçel, 1995, pp.34-35; Çaykan, 2001, pp. 36-37; Davies, 1995, pp. 9-10).

### **2.6.2.3. Weaknesses of Communicative Approach**

Unlike the structuralist approach, this approach does not emphasize learning structural grammar, yet it may be difficult to achieve communicative competence without a considerable mastery of the grammar of a language. It is possible for cultural bias to affect the reliability of the tests being administered.

Communicative testing is a challenge for test designers. One reason is the issue of predictive validity. When designing a test of communicative ability, identifying test takers' needs based on communicative encounters that they are

likely to experience is one of the basic principles. However, it is not certain if test makers can guarantee that testees performing well on a test in class are also able to do well outside the classroom in a real life situation. One reason for this is that real life communication is characterized by unpredictability. Studies have proved that test designers have tried to make real-world tasks, but encountered difficulties from the varied or diverse nature of contexts (Katsumasa, 1997; Brown, 2003)

No matter what kind of approach is followed since it achieves a positive backwash, washback is useful to both language teaching and language learning, (Cheng, Watanabe, & Curtis 2004 and Hughes, 1989).

## **CHAPTER THREE**

### **METHODOLOGY**

#### **3.1. INTRODUCTION**

This chapter presents information about the research design, the methodology; the selection of the participants, the instruments, the data collection procedures, and the methods used for data analysis.

#### **3.2. NATURE OF STUDY**

In this study, the descriptive study design was adopted because in the study the researcher wanted to examine the recent assessment and evaluation activities carried out in the schools of foreign languages. Descriptive study does not answer questions about how, when, why the characteristics occurred. Rather it addresses the "what" question. Gay (1976) defines descriptive method involving collection of data in order to test hypothesis or to answer questions concerning the current status of the subject of the study. The purpose of using the descriptive research method is to acquire accurate, factual, systematic data that can provide you with an actual picture of the data set that you are reviewing (Williams, 2011). Descriptive studies intend to describe or explain relationships among phenomena, situations, and events as they occur. The major purpose of descriptive research is to provide an overall "picture" of a population or phenomenon by describing situations or events (Rubin & Babbie, 1993).

According to Travers (1978), the aim is to describe the nature of situation as it exists at the time of the study and to explore the causes of particular phenomena. The description is used for frequencies, averages and other statistical calculations. Often the best approach, prior to writing descriptive research, is to conduct a survey investigation. Qualitative research often has the aim of description, and researchers may follow-up with examinations of why the observations exist and what the implications of the findings are.

### **3.3. METHODOLOGY OF THE STUDY**

#### **3.3.1. Setting**

First of all, in order to inform about the context, some information about the participant schools have been presented:

Beykent University has got 5 different level courses from the beginning to the advanced, and these levels are in line with Common European Framework. It has got seven different modules, and each one lasts for seven weeks. Each student has to pass each module; otherwise, they cannot continue their education in their departments. All of the contexts are supplied with technological devices.

Bülent Ecevit University has got the compulsory preparatory education for the students for a year. Their aim is helping students to use English in four skills efficiently and for each skill they also define subskills to teach. While teaching, they try to teach it with student oriented techniques and make use of technological devices.

Eskişehir Osmangazi University has the aim of the knowledge of English at the equal level for the students who have different knowledge before. For this reason, they give instruction by using the four skills. In order to provide effective and efficient instruction, there are four offices helping them. These are testing and assessment, Curriculum development, Supplementary material and Student affairs.

Gazi University has the aim of improving students' four skills with the updated courses, effective ways of teaching with the most modern technological devices. For them, it is very important for students to have necessary English knowledge to be successful in the life.

Istanbul Technical University has been giving compulsory preparatory program for all architecture and engineering students since 1999. This lasts for one year and their aim is to teach English. They give importance to four skills equally with modern technological devices.

Izmir Economy University has got compulsory preparatory programme for all the new comers. Their aim is creating teaching atmosphere for the students by using two languages effectively. The courses are among the small groups of students with technological equipment.

Izmir University aims to equip incoming students enrolled to English-medium programs with the English and academic skills necessary for them to fulfill the requirements of the courses. In addition, it aims at helping students to attain B1 level competence in a second foreign language. The aim of instruction is to produce autonomous life-long learners with strong language, academic and critical thinking skills. Therefore, the courses reflect an active as well as a collaborative/cooperative approach to learning. To create a realistic context in which learners make use of various language skills i.e. a simulation of what is done within academic contexts, instruction is theme based and the skills are practiced in an integrated manner.

Karadeniz Technical University has got 25 students in each class. They give the compulsory preparatory education to 981 students in bachelor degree and 200 students in master degree. There is an interactive atmosphere in the classes. They give importance to each skill as well as translation and business English. They make the classes interactive with the presentation of the students.

Muğla University has the aim of helping students to have the necessary strategies, and using them efficiently with their knowledge of English in professional and educational life. In the preparatory programme, there are three

levels. These are beginner, elementary and pre-intermediate levels. In beginner level, students have 24 hours of English in a week, 20 hours of English in elementary and pre-intermediate levels.

Pamukkale University has got 31 classes in the School of Foreign Languages. They give preparatory education for one year for the departments having compulsory preparatory programme. Their aim is giving the students the knowledge of foreign language as much as they need. In order to provide it in the courses, they use on-line materials, computer, projection, internet connection with the appropriate level of text books.

The goals of the Foreign Language schools may be summarized as follows:

Students will be able to learn and use four language skills communicatively; namely, listening, speaking, reading and writing, and improve them for academic purposes in order to carry out their education in their related field of study. For this purpose, the schools try to make use of the recent methods and techniques, the most effective teaching methods and learning strategies, and necessary technical equipment, resources and library.

### **3.3.2. Participants**

The present study investigated the assessment and evaluation activities in the School of Foreign Languages in 10 universities in Turkey. Three of them are private whereas the rest of them are state universities (see Table 3.1.). Here are the short descriptions of each university:

Beykent University is a private university and located in Ankara. The School of Foreign Languages has got 40 instructors and 900 students. They give A2, B1, C2, C1 level of tests.

Bülent Ecevit University is a state university and located in Zonguldak. The School of Foreign Languages has got 55 instructors and 915 students. They give just A1 and A2 level of tests.

Eskişehir Osmangazi University is a state university and located in Eskişehir. The school of Foreign Languages has got 72 instructors and 886 students. They give A1, A2 and B1 level of tests.

Gazi University is a state university and located in Ankara. The School of Foreign Languages has got 100 instructors and 1400 students. They give A1, A2, B1 and B2 level of tests.

Istanbul Technical University is a state university and located in İstanbul. The School of Foreign Languages has got 150 instructors and 2650 students. They give A1, A2, B1 and B2 level of tests.

Izmir Economy University is a private university and also located in İzmir. There are 117 instructors and 1172 students. They give A1, A2, B1 and B2 level of tests.

Izmir University is a private university and located in İzmir. There are 54 instructors and 431 students. They give A1, A2, B1 and B2 level of tests.

Karadeniz Technical University is a state university and located in Trabzon. In the School of Foreign Languages, there are 58 instructors and 1800 students. They give A1, A2, and B1 level of tests.

Muğla University is a state university and located in Muğla. The School of Foreign Languages has got 70 instructors and 750 students. They give A1, A2, B1, B2 level of tests.

**Table 3.1.** The participant universities, and the number of instructors and Students in the school of Foreign Languages

The universities	The number of instructors	The number of students
Beykent University *	67	1500
Bülent Ecevit University	58	1800
Eskişehir Osmangazi University	54	431
Gazi University	117	1172
Istanbul Technical University	150	2650
Izmir Economy University *	72	886
Izmir University*	100	1400
Karadeniz Technical University	55	915
Muğla University	40	900
Pamukkale University	67	750

\* private universities

Pamukkale University is a state university and located in Denizli. The School of Foreign Languages has 67 instructors and 1500 students. They give A1, A2, B1 and B2 level of tests.

### **3.3. DATA COLLECTION AND PROCEDURES**

In order to collect our data, we designed a questionnaire. Our questionnaire had two sections. First section was about the demographic information about the participant school of foreign languages such as the number of instructors, the number of students, the level of tests, the number of placement tests, the number of proficiency tests, the number of achievement tests, the number of quizzes administered in a semester depending on the skills and the weight of skills. In the second section of the questionnaire, we tried to focus on the skills and the sub-skills assessed in various achievement tests and the item types they used depending on the frequency. The piloting was carried out with other three universities, and the Cronbach Alpha of the questionnaire was .91. On the other hand, the reliability of the questionnaire after the application was .96.

### **3.4. DATA ANALYSIS AND PROCEDURES**

The data gathered through the questionnaire were analyzed by using quantitative analysis techniques. In evaluating quantitative data, numeric data obtained from questionnaires were calculated through the SPSS for Windows-Version 16.0 software. The questionnaire results in the tables were presented in terms of means, standard deviations, percentages and frequencies. Depending on the type and content of the data gathered, either mean scores and standard deviations or percentages and frequencies were presented in the tables or in charts.



## **CHAPTER FOUR**

### **FINDINGS AND DISCUSSIONS**

#### **4.1. INTRODUCTION**

In this section, the results of our data based on the research questions are presented and discussed. The findings are presented as tables and are interpreted by comparing and contrasting with previous research.

The main purpose of this study is to describe the assessment and evaluation activities in the Schools of Foreign Languages in Turkey and create awareness for those involved in preparing and administering communicative language assessment tools in which four language skills are tested by providing necessary background and theoretical knowledge about language testing. With this aim in mind, this study attempts to find answers to the following research questions:

1. What kinds of assessment and evaluation activities are done in the School of Foreign Languages in Turkey?
2. How is the listening skill assessed?
3. How is the reading skill assessed?
4. How is the speaking skill assessed?
5. How is the writing skill assessed?
6. How is the language use assessed?
7. How is the vocabulary assessed?

#### 4.2. THE KIND OF ASSESSMENT AND EVALUATION ACTIVITIES CARRIED OUT IN THE SCHOOLS OF FOREIGN LANGUAGES IN TURKEY

Our first research question was “what kind of assessment and evaluation activities are done in the School of Foreign Languages in Turkey?” Our data revealed that a proficiency test, a placement test, an achievement test, and different quizzes on various language skills and subskills were administered with various frequencies in the participant schools in an academic year.

According to the findings, all of the participant schools administered A2 level tests. However, nine of them administered A1 and B1 level tests while only one of them administered C1 level tests. On the other hand, none of the schools administered C2 level tests. This showed that most of the participant schools ended up their curriculum in teaching English at B2 level (see Table 4.1.). The levels, mentioned here, were specified according to Common European Framework of References for Languages (CEFR).

**Table 4.1.** Frequency of the tests administered in an academic year in the School of Foreign Languages

Levels		Percentage %
A1	YES	90
	NO	10
A2	YES	100
	NO	-
B1	YES	90
	NO	10
B2	YES	70
	NO	30
C1	YES	10
	NO	90
C2	YES	-
	NO	100

As the participant schools stated, almost all of them administered a placement test once in an academic year; however, just two schools administered a placement test twice a year. This showed that all of the participant schools placed their students at an appropriate level in their program according to the placement test administered at least once a year.

On the other hand, the application of a proficiency test in an academic year varied. One of the participant schools administered a proficiency test three times a year whereas another one administered a proficiency test eight times a year. Two of the participant schools administered a proficiency test four times a year. Although there are some different applications of proficiency tests, the majority of the participant schools (six of them) administered a proficiency test once or twice a year (see Table 4.2.).

**Table 4.2.** Application of a proficiency test in an academic year

	Number	Percentage %
Once a year	3	30
Twice a year	3	30
Three times a year	1	10
Four times a year	2	20
Eight times a year	1	10

It should be kept in mind that a proficiency test was an also exemption test, and it is a criterion-referenced assessment. When the students achieve this test, they are considered proficient in English so as to carry on their academic studies in their own departments.

**Table 4.3.** Application of an achievement test in an academic year

	Number	Frequency %
Four times in a year	6	60
Five times in a year	1	10
Six times in a year	1	10
Seven times in a year	1	10
Ten times in a year	1	10

Our data revealed that six of the participant schools administered an achievement test four times in an academic year; however, the frequency of the other universities varied such as five times, six times, seven times and ten times in an academic year. Table 4.3. show that all of the schools applied achievement tests with various frequencies.

As for the quizzes administered, all of the participant schools administered quizzes on various skills in an academic year. They administered quizzes on listening, watching, reading, writing, language use, vocabulary and

grammar. Among the schools, four of them administered a listening quiz twice a year, and two of them five times a year. On the other hand, one of them administered a listening quiz seven times a year, which was the most frequent (see Table 4.4). This showed that listening skill was important for the participant schools.

**Table 4.4.** Application of listening quiz

	Number	Frequency %
Once a year	1	10
Twice a year	4	40
Three times a year	1	10
Four times a year	1	10
Five times a year	2	20
Seven times a year	1	10

Among the participant schools, only one of them administered a watching quiz. It showed that watching as an assessment tool was ignored by the participants although most of the text books were equipped with CD ROMs or on-line materials to watch some pedagogic films.

Of the schools which participated in the study, two of them administered a reading quiz once a year; three of them administered a reading quiz twice a year, one of them three times a year, one of them four times a year, two of them five times a year. On the other hand, one of them administered a reading quiz eight times a year, which was the most frequent (see Table 4.5). This shows that reading skill is important for the participant schools

**Table 4.5.** Application of reading quiz

	Number	Percentage %
Once	2	20
Twice	3	30
Three times	1	10
Four times	1	10
Five times	2	20
Eight times	1	10

Of the schools, one of them administered a writing quiz once a year, four of them administered a writing quiz twice a year, one of them three times a year and three of them four times a year .On the other hand, one of them administered a writing quiz ten times a year, which was the most frequent (see Table 4.6.). This showed that writing skill was important for the participant schools.

**Table 4.6.** Application of writing quiz

	Number	Percentage %
Once	1	10
Twice	4	40
Three times	1	10
Four times	3	30
Ten times	1	10

Of the schools, one of them administered a language use quiz once a year, two of them administered a listening quiz twice a year, one of them three times a year, one of them four times a year, and one of them six times a year. On the other hand, three of them never administered a language use quiz (see Table 4.7.). This showed that testing language use skill separately was not very common among the schools.

**Table 4.7.** Application of language use

	Number	Percentage %
Never	3	30
Once	1	10
Twice	2	20
Three times	1	10
Four times	1	10
Six times	2	20

Of the schools, two of them administered a vocabulary quiz once a year, one of them three times a year, one of them four times a year, one of them five times a year, one of them six times a year and two of them eight times a year. On the other hand, two of them never administered a vocabulary quiz (see Table 4.8.). This showed that for most of the schools testing vocabulary skill separately was not very usual.

**Table 4.8.** Application of vocabulary quiz

	Number	Percentage %
Never	2	20
Once	2	20
Three times	1	10
Four times	1	10
Five times	1	10
Six times	1	10
Eight times	2	20

Of the schools, one of them administered a grammar quiz once a year, two of them three times a year, two of them four times a year, one of them five times a year, and one of them eight times a year. On the other hand, three of them never administered a grammar quiz (see Table 4.9.). This showed that grammar skill was not tested very frequently by the participant schools.

**Table 4.9.** Application of grammar quiz

	Number	Frequency %
Never	3	30
Once	1	10
Three times	2	20
Four times	2	20
Five times	1	10
Eight times	1	10

### 4.3. ASSESSMENT OF LISTENING SKILL

The second research question was as to how the listening skill was assessed. The weight of listening skill in proficiency, placement and achievement tests was presented below in Table 4.10.

Our data revealed that listening skill was very important for the participant schools except for one in the achievement test. All of the participant schools gave equal importance to listening skill in the achievement test. But the weight of them varies such 15%, 20%, 25%. However, we could not state the same result for the placement test. For all the participant schools, listening skill was not important in the placement test. Only four of the participant schools test

listening skill in the placement test with the variety in weight such as 20% and 25%. It was also the same in the proficiency exam. Only five of the schools tested listening skill in the proficiency exam and although one of the universities administers listening skill in the proficiency exam, there was no information provided for it. The weight of listening skill varied between 10% and 25%.

**Table 4.10.** Weight of listening skill in various tests

The Universities	Achievement Test %	Placement Test %	Proficiency Test %
Beykent University	20	25	No information
Bülent Ecevit University	28	-	-
Eskişehir Osmangazi University	10	20	-
Gazi University	25	25	-
Istanbul Technical University	15	-	20
Izmir Economy University	20	-	15
Izmir University	15	20	10
Karadeniz Technical University	-	-	-
Muğla University	15	-	15
Pamukkale University	25	-	25

In the assessment of listening skill, the participant schools prepared items for the following subskills of listening: skimming, scanning, guessing the title, understanding the main idea, referencing, dictation/note taking (at word/phrase level), guessing the meaning of unknown words/phrases, information transfer, inferencing, speaker's attitude or opinion, identifying facts or opinions, recognizing discourse markers and patterns at different levels. Among the subskills, understanding the main idea, skimming, information transfer, inferencing, scanning, dictation/note taking (at word/phrase level) and recognizing discourse markers, patterns were the most frequent subskills tested in the participant schools (see Table 4.11).

**Table 4.11.** Assessment of listening subskills

Subskills	Mean	Sd	Participation Level
1. Understanding the main idea	3.60	1.17	Sometimes
2. Skimming	3.50	1.26	Sometimes
3. Information transfer	3.20	1.62	Sometimes
4. Inferencing	3.10	1.62	Sometimes
5. Dictation/note taking (word/phrase level)	3.10	1.44	Sometimes
6. Recognizing discourse markers, patterns	3.00	1.33	Sometimes
7. Scanning	3.00	3.14	Sometimes
8. Note taking (guided)	2.80	1.39	Rarely
9. Referencing	2.80	1.22	Rarely
10. Identifying facts or opinions	2.80	0.79	Rarely
11. Guessing the meaning of unknown words/ phrases	2.60	1.43	Rarely
12. Identifying speaker's attitude or opinion	2.60	1.35	Rarely
13. Guessing the title	2.00	1.24	Rarely

However, referencing, guided note taking, identifying facts or opinions, guessing the meaning of unknown words/phrases, and identifying speaker's attitude or opinion were sometimes tested. On the other hand, guessing the title was a subskill which was rarely tested in listening part of the achievement tests.

When we analyzed the item types used to assess listening subskills, we found out that the most frequently used item type was multiple choice at word/phrase level. This item type was often used in the listening parts of achievement tests ( $m=4.10$ ;  $sd=0.73$ ). In addition, the following item types were sometimes used ones: multiple choice at sentence level, filling in the blanks at word/phrase level, True/False, filling out a form/a table, matching at word/phrase level, open-ended items, filling in the blanks at sentence level. On the other hand, multiple choice in a cloze test, sentence completion, sequencing sentences to make a summary, sequencing the sentences to put in a correct order, and completing the dialogue with multiple choice item types were rarely used. However, sequencing the paragraphs, writing a response to a given situation, choosing the irrelevant statement in a paragraph, finding the paraphrased statement and summarizing the text are the item types which were never used in testing the listening skill (see Table 4.12).



**Table 4.12.** The item types used in testing listening subskill

Item Types	Mean	Sd	Participation Level
1. Multiple Choice at word/phrase level	4.10	0.73	Often
2. Multiple Choice at sentence level	3.70	1.25	Sometimes
3. Filling in the blanks at word/phrase level	3.50	2.27	Sometimes
4. True/False	3.40	1.50	Sometimes
5. Filling out a form/a table	3.30	0.82	Sometimes
6. Matching at word/phrase level	3.20	1.62	Sometimes
7. Filling in the blanks at sentence level	3.10	1.45	Sometimes
8. Open-ended items	3.10	1.45	Sometimes
9. Matching at sentence level	3.00	1.56	Sometimes
10. Multiple Choice in a cloze test	2.90	1.79	Rarely
11. Sentence completion	2.80	1.13	Rarely
12. Completing the dialogue with MC	2.40	0.96	Rarely
13. Sequencing the sentences to put in a correct order	2.30	1.33	Rarely
14. Matching at paragraph level	2.10	1.19	Rarely
15. Sequencing sentences to make a summary	2.10	1.19	Rarely
16. Placing the appropriate sentence in a paragraph	2.10	1.45	Rarely
17. Choosing the irrelevant statement in a paragraph	1.80	1.13	Never
18. Sequencing the paragraphs	1.70	1.06	Never
19. Finding the paraphrased statement	1.70	1.06	Never
20. Writing a response to a given situation	1.60	1.07	Never
21. Summarizing the text	1.40	0.51	Never

In Turkey, there was one similar study related to our study. Kırık (2008) studied the attitudes of English teachers' of Anatolian High Schools in İstanbul in terms of testing and assessment, whose aim was describing the current assessment activities. In the study, she described the current assessment and evaluation activities. The most frequently used item type was dictation; however it was not the same in our study. In our study multiple choice was the most frequent used item type. The studies on testing listening, such as Kitao, Chastain's studies, were about the general assessment of testing listening. The main concern in these studies was that creating reliable and valid L2 listening tests was not an easy process. But because of the importance of listening in

language learning and communication, it was imperative that teachers and testers invested the resources needed to make quality tests. Ur (1984: 35-46) stated the subskills for listening. According to her, the subskills should be distinguishing different words, identifying words, working out the spelling, relating pronouns, identifying important points. In our study we could see that our subskills covered what Ur mentioned.

Hughes (2003) emphasized that testing listening was done best with multiple choice test item and in our study the most frequent test item was multiple choice. According to Chastain (1976:287-293), with listening comprehension one must be able to: 1) discriminate between the significant sound and intonation patterns of the language; 2) perceive an oral message; 3) keep the communication in mind while it is being processed; and finally, 4) understand the contained message and the data from the questionnaire. Our data revealed that the item types used for listening were parallel with his ideas, and the most frequently tested subskills were related to the suggestions of Chastain (1976).

#### **4.3. ASSESSMENT OF READING SKILL**

The third research question was how the reading skill was assessed. The weight of reading skill in proficiency, placement and achievement tests was presented below in Table 4.13.

Our data revealed that reading skill was very important for all of the participant schools in the achievement test. They gave equal importance to reading skill in the achievement test. It was also the same for the placement test except for one participant school because there was no information about it. For all of the participant schools, reading skill was important in the placement test, too. The average weight was 22.5 % in the placement test. It was also the same in the proficiency exam. They all tested reading skill in the proficiency exam and the average weight was 24 %.

**Table 4.13.** Weight of reading skill in various tests

The Universities	Achievement Test %	Placement Test %	Proficiency Test %
Beykent University	20	25	No information
Bülent Ecevit University	9.2	10	10
Eskişehir Osmangazi University	20	20	20
Gazi University	25	25	25
Istanbul Technical University	30	40	45
Izmir Economy University	25	30	30
Izmir University	30	30	30
Karadeniz Technical University	25	25	25
Muğla University	30	15	30
Pamukkale University	15	30	25

Our data revealed that reading skill was very important for all the participant schools in the achievement, placement and proficiency test. All of the participant schools gave equal importance to reading skill in the achievement test. But the weight of them varies such 15%, 20%, 25%. In the placement test, the weight varied between 10% and 40%. In the proficiency exam, for one of the universities there was no information provided for it. The weight of reading skill varied between 10% and 45%.

In the assessment of reading skill, the participant schools prepared items for the following subskills of reading: skimming, scanning, guessing the title, understanding the main idea, referencing, guessing the meaning of unknown words/phrases, identifying facts or opinions, inferencing, guessing the title, information transfer, recognizing discourse markers and patterns at different levels and outlining (paragraph and test level). Among the subskills, understanding the main idea, skimming and referencing were the most frequent subskills tested in the participant schools (see Table 4.14.). However, identifying facts or opinions, guessing the meaning of unknown words/phrases, and identifying speaker's attitude or opinion were sometimes tested. On the other hand, information transfer speaker's attitude or opinion, recognizing discourse markers, patterns, outlining at different levels were the sub skills which were rarely tested in reading part of the achievement tests.

**Table 4.14.** Assessment of reading subskills

Subskills tested in reading	Mean	Sd	Participation Level
1.Skimming	4.30	0.50	Often
2.Understanding the main idea	4.20	0.42	Often
3.Referencing	3.90	0.74	Often
4.Scanning	3.80	1.23	Sometimes
5.Guessing the meaning of unknown words/phrases	3.70	1.33	Sometimes
6.Identifying facts or opinions	3.50	0.85	Sometimes
7.Inferencing	3.40	1.43	Sometimes
8.Guessing the title	3.00	1.63	Sometimes
9.Information transfer	2.90	1.28	Rarely
10.Speaker's attitude or opinion	2.80	1.47	Rarely
11.Recognizing discourse markers, patterns	2.70	1.25	Rarely
12.Outlining (paragraph level)	1.60	0.84	Rarely
13.Outlining (text level)	1.60	0.84	Rarely

When we analyzed the item types used to assess reading subskills, we found out that the most frequently used item type was multiple choice at word/phrase level. This item type was often used in the reading parts of achievement tests ( $m=4.30$ ;  $sd=0.67$ ). In addition, the following item types were sometimes used: filling the blanks, sentence completion, matching at paragraph level, finding the paraphrased statement, multiple choice in a cloze test. However, sequencing the paragraphs, sequencing the sentences to put in a correct order, completing the dialogue with mc, writing a response to a given situation, summarizing the text were the item types which were rarely used in testing the reading skill. On the other hand outlining was the item type which was never used in testing reading skill (see Table 4.15.).

**Table 4.15.** The item types used in testing reading subskills

Item types	Mean	Sd	Participation level
1. Multiple Choice at word/phrase level	4.30	0.67	Often
2. Multiple Choice at sentence level	4.30	0.67	Often
3. True/False	4.00	1.33	Often
4. Matching at word/phrase level	3.20	1.23	Often
5. Open-ended items	3.10	1.52	Often
6. Matching at sentence level	3.10	1.37	Often
7. Placing the appropriate sentence in a paragraph	3.00	1.56	Often
8. Filling in the blanks at word/phrase level	2.80	1.47	Sometimes
9. Filling in the blanks at sentence level	2.60	1.50	Sometimes
10. Sentence completion	2.50	1.43	Sometimes
11. Matching at paragraph level	2.20	1.03	Sometimes
12. Finding the paraphrased statement	2.20	1.31	Sometimes
13. Multiple Choice in a cloze test	2.10	1.59	Sometimes
14. Sequencing sentences to make a summary	2.10	0.99	Sometimes
15. Choosing the irrelevant statement in a paragraph	2.10	1.19	Sometimes
16. Filling out a form/a table	2.00	1.41	Sometimes
17. Sequencing the paragraphs	1.90	1.10	Rarely
18. Sequencing the sentences to put in a correct order	1.80	1.13	Rarely
19. Completing the dialogue with MC	1.80	1.31	Rarely
20. Writing a response to a given situation	1.60	1.07	Rarely
21. Summarizing the text	1.30	0.67	Rarely

As part of the assessment, the participant schools never tested translation except one. In that school, the usual item types used in the assessment were multiple choice (sentence level), paragraph translation and essay translation (see Table 4.16.)

**Table 4.16.** The item types used in testing translation

Item Types	Mean	Standard Deviation
1. Multiple Choice (sentence level)	1.70	1.25
2. Paragraph translation	1.10	0.31
3. Essay translation	1.10	0.31

In Kırık's study (2008), the attitudes of English teachers' in terms of testing and assessment, it was found out that reading aloud was the most frequent used subskill. However, in our study there was no sub skill called like

that and the most frequent one was skimming. According to Kitao (1996) reading should be assessed in line with the reading levels. When we examined the subskills used or reading we could see that the participants focus on high level of readers. Kitao (1996) also suggested that in testing middle and higher level students, the item types were generally true false, multiple choice, short answer. In our study our data revealed that these item types were often preferred by the participant universities.

In the area of reading research specifically, Lunzer, Waite and Dolan (1974) constructed reading tests, intended to measure different reading skills, but failed to find evidence an implicational scale. However, in our study we could easily observe that there was a scale among the skills. For example in our study skimming was the first skill but outlining was the last skill. According to Brown (2004), the assessment of reading could imply the assessment of reading strategies like skimming, scanning, deducing the meaning, and in our study we could easily observe that these were the most frequently used subskills. According to Hughes (2003) these subskills could be assessed easily with multiple choice and short answer questions. When we analyzed our data, we saw that the participant universities made use of the multiple choice items the most. Schreiner (1977) suggested that the ideal measurement instrument in reading was reading tasks that had been created to reflect about cognitive processing which meant understanding the main idea. When we looked at the data we got, we could easily observe that this was one of the most frequently used subskill in testing reading among the participant schools.

#### **4.4. ASSESSMENT OF SPEAKING SKILL**

The fourth research question was how the speaking skill was assessed. The weight of speaking skill in proficiency, placement and achievement tests was presented below in Table 4.17.

**Table 4.17.** Weight of speaking skill in various tests

The Universities	Achievement Test %	Placement Test %	Proficiency Test %
Beykent University	-	-	-
Bülent Ecevit University	19	-	-
Eskişehir Osmangazi University	-	-	-
Gazi University	25	25	-
Istanbul Technical University	100( separate exam)	-	-
Izmir Economy University	20	-	20
Izmir University	15	-	15
Karadeniz Technical University	-	-	-
Muğla University	-	-	-
Pamukkale University	100(separate exam)	-	-

Our data revealed that speaking skill was not very important for all of the participant schools. Six of the participant schools gave equal importance to speaking skill in the achievement test. Of all two of the participant universities administer separate speaking skill. But the weight of them varied such 15% and 20%. However, we could not state the same result for the placement test. For all the participant schools, speaking skill was not important in the placement test. Only two of the participant universities gave place to speaking skill in the placement test with the weight 25%. It was also the same in the proficiency exam. Only two of the schools tested speaking skill in the proficiency exam and the weight of speaking skill was 15% and 20%.

Our data revealed that speaking skill was not as important as the other language skills in the achievement test. Six of the participant universities gave equal importance to speaking skill in the achievement test. In addition, among the participant schools, two of them applied a separate speaking part of the achievement exam. It was also the same for the placement test that just one of the participant schools tested speaking skill in the placement test. None of the other participant schools tested speaking skill in the placement test.

**Table 4.18.** Assessment of speaking subskills

Subskills	Mean	Sd	Participation level
1. Description	4.20	0.63	Often
2. Having a dialogue on a topic	3.90	1.19	Often
3. Narration	3.60	1.07	Often
4. Problem solution	3.10	0.87	Often
5. Giving a presentation on a topic	3.10	1.37	Often
6. Comparison/Contrast	3.00	1.05	Sometimes
7. Cause and effect	3.00	1.05	Sometimes
8. Talking about a process (how to make a cake, etc.)	2.80	1.13	Sometimes
9. Persuasive Talk	2.30	0.95	Sometimes
10. Argumentative Talk	2.20	1.03	Sometimes

In the assessment of speaking skill, the participant schools prepared items for the following subskills of speaking: descriptive, having a dialogue on a topic, narrative, problem solution, giving a presentation on a topic, comparison contrast, cause and effect, talking about a process, persuasive and argumentative talk. Among the subskills, descriptive, having a dialogue on a topic, narrative, problem solution, giving a presentation on a topic were the most frequent subskills tested in the participant schools (see Table 4.18). However, comparison contrast, cause and effect, talking about a process, persuasive talk sometimes tested.

When we analyzed the item types used to assess speaking subskills, we found out that the most frequently used item type was having a dialogue on a topic. This item type was often used in the speaking part of achievement tests ( $m=3.70$ ;  $sd=1.33$ ). In addition, the following item type was sometimes used: summarizing a film/a story/a novel (see Table 4.19).

**Table 4.19.** The item types used in speaking

Item types	Mean	Sd	Participation level
1. Having a dialogue on a topic	3.70	1.33	Often
2. Picture Talk	3.60	1.17	Often
3. Choosing a topic and talking about it	3.60	1.17	Often
4. Summarizing a film/a story/a novel	2.70	1.06	Sometimes

When we analyzed the application of the speaking test, we found that the most frequently used application type was testing one student at a time. This



was often used in speaking part of achievement tests ( $m= 3.70$ ;  $sd= 1.33$ ). However, testing a group of 4 or 5 students at a time was rarely used (see Table 4.20).

**Table 4.20.** Application of the speaking test

Application Types	Mean	Sd	Participation level
1. Testing one student at a time	3.70	1.33	Often
2. Testing two students at a time	3.40	1.77	Often
3. Testing three students at a time	2.00	1.05	Sometimes
4. Recording all the students	2.30	1.89	Sometimes
5. Testing a group of 4 or 5 students at a time	1.20	0.42	Rarely

Kitao (1999) suggested that communicative language tests were intended to be a measure of how the testees were able to use language in real life situations. In testing productive skills, emphasis was placed on appropriateness rather than on ability to form grammatically correct sentences. From the data we got, we could easily observe that the aim of participant universities was having the student produce in the target language. Like Güllüoğlu's study (2004), whose aim was to determine whether there was a lack of teaching speaking skills and speaking tests at Gazi University Preparatory School of English, our study also determined the lack of speaking tests. In our study it was seen that there was a positive attitude towards the speaking tests, so there was no lack of speaking tests.

As Brown (2004) stated speaking was a productive skill that could be directly observed, our data revealed that it was directly observed because at one time one student was assessed. According to Hughes (2003) elicitation techniques like discussion, picture talk must be used and our data revealed that these were the most frequent used item types. In Kırık's study (2008), the attitudes of English teachers' in terms of testing and assessment, it was found out that the most frequent used item type was question answer, however in our study we found out that having a dialogue on a topic was the most frequent used item type . There were several speaking-assessment tasks often used by teachers for assessing learners' oral communication skills, for instance, responding orally to question slips, describing and reacting to visual prompts,

story-telling, and giving an oral presentation (Chinda, 2009; Khamkhien, 2010; Sook, 2003). For them, of all the direct performance-based assessment tasks, face-to-face interview and especially role-play apparently were the most popular choices among teachers for assessing speaking skills of Thai EFL learners. However in our study we could see that descriptive speaking was the most frequent one and argumentative talk was the least frequent one.

#### 4.5. ASSESSMENT OF WRITING SKILL

In the fifth question, we tried to find out how the writing skill was assessed. The weight of writing skill in proficiency, placement and achievement tests was presented below in Table 4.21.

**Table 4.21.** Weight of writing skill in various tests

The Universities	Achievement Test %	Placement Test %	Proficiency Test %
Beykent University	-	-	-
Bülent Ecevit University	15	-	-
Eskişehir Osmangazi University	-	-	-
Gazi University	25	25	25
Istanbul Technical University	15-20	-	25
Izmir Economy University	20	-	30
Izmir University	15	20	15
Karadeniz Technical University	-	-	-
Muğla University	15	-	15
Pamukkale University	25	-	25

Our data revealed that writing skill was very important for the participant schools except for two in the achievement test. All of the participant schools gave equal importance to writing skill in the achievement test. But the weight of them varies such 15%, 20%, and 25%. However, we could not state the same result for the placement test. For all the participant schools, writing skill was not important in the placement test. Only two of the participant universities tested writing skill in the placement test with the variety in weight such as 20% and 25%. It was also the same in the proficiency exam. Only six of the schools

tested writing skill in the proficiency exam and the weight of writing skill varied between 15% and 30%.

Our data revealed that writing skill was frequently tested in the seven participant schools in the achievement test. Four of the participant schools gave equal importance to writing skill in the achievement test. However, It was not the same for the placement test that just three of the participant schools tested writing skill in the placement test. In the proficiency exam, four of the participant schools did not administer writing skill.

In the assessment of writing skill at the paragraph level, the participant schools prepared items for the following subskills of writing: description, comparison and contrast, opinion and cause and effect. Among the subskills, description was the most frequent subskill tested (see Table 4.22.). However, comparison/contrast, cause and effect, opinion were sometimes tested.

**Table 4.22.** Assessment of writing subskills at the paragraph level

Subskills	Mean	Sd	Participation level
1. Description	4.20	0.79	Always
2. Comparison/contrast	3.90	1.19	Often
3. Opinion	3.80	1.23	Often
4. Cause and effect	3.60	1.50	Often

When we analyzed the item types used to assess writing subskills at the paragraph level, we found out that the most frequently used item type was guided writing task. This item type was often used in the writing part of achievement tests ( $m=3.90$ ;  $sd=1.19$ ). In addition, writing a topic sentence in a text, controlled writing task, free writing task, writing an email and writing notes about daily issues were often used. In addition, the following item types were sometimes used: putting the cohesive words in appropriate place, re-writing some sentences, writing a letter of advice, writing an application of letter, making an outline of given text, finding and correcting the misspelled words (see Table 4.23.).

**Table 4.23.** The item types used at the paragraph level

<b>Subskills</b>	<b>Mean</b>	<b>Sd</b>	<b>Participation level</b>
1.Guided writing task	3.90	1.19	Often
2.Writing a topic sentence/ conclusion sentence in a text	3.60	1.50	Often
3.Controlled writing task	3.50	1.27	Often
4.Free writing task	3.10	1.19	Often
5.Writing an e-mail	3.10	0.99	Often
6.Writing notes about daily issues	3.00	1.33	Often
7.Putting the cohesive words in appropriate place	2.90	1.37	Sometimes
8.Re-writing some sentences (paraphrase)	2.90	1.45	Sometimes
9.Writing a letter of advice	2.80	1.13	Sometimes
10.Writing an application letter	2.70	1.16	Sometimes
11.Making an outline of the given text	2.60	1.71	Sometimes
12.Finding and Correct the misspelled words	2.20	1.47	Sometimes

In the assessment of writing skill at the essay level, the participant schools prepared items for the following subskills of writing: focus on the cause, focus on the effect, comparison and contrast, opinion, descriptive, narrative, classification, problem solution, and argumentation. Among the subskills, focus on the cause and the effect was the most frequent subskills tested in the participant schools (see Table 4.24). However, description, narration, classification, problem solution, argumentative subskills were sometimes tested.

**Table 4.24.** Assessment of writing subskills at the essay level

Subskills	Mean	Sd	Participation level
1.Focus on the cause	3.80	1.13	Often
2.Focus on the effect	3.80	1.13	Often
3.Comparison/Contrast	3.60	1.07	Often
4.Opinion	3.40	1.43	Often
5.Description	2.70	1.56	Sometimes
6.Narration	2.60	1.43	Sometimes
7.Classification	2.50	1.50	Sometimes
8.Problem solution	2.50	1.35	Sometimes
9.Argumentation	2.30	1.33	Sometimes

When we analyzed the item types used to assess writing subskills at the essay level, we found out that the most frequently used item type was guided writing task. This item type was often used in the writing part of achievement tests ( $m=3.80$ ;  $sd=1.31$ ). In addition, controlled writing task was often used. In addition, free writing item type was sometimes used (see Table 4.25.).

**Table 4.25.** The item types used in writing skill at the essay level

Item types	Mean	Sd	Participation level
1.Guided writing task	3.80	1.31	Often
2.Controlled writing task	3.40	1.43	Often
3.Free writing task	2.70	1.56	Sometimes

In her study, Kirik (2008) revealed that the most frequent used sub skill item was guided questions and answers. However, it was not the same in our study. In essay writing, the most frequent one was focused on the cause, while in paragraph writing it was descriptive writing. According to Kitao (1999), testing each skill was uniquely difficult, but testing writing presented two particular problems. The first was making decisions about the matter of control, objectivity of the evaluation, and naturalness in the writing test and the second one was if the test was done in a way that it could not be graded objectively, it was necessary to develop a scale that allows it to be graded as objectively as possible. That's why Hughes (2003) stated that we should just test the writing

ability and nothing else. We could infer that all the participants applied writing test accordingly. According to Breland (1983) assessment of writing varied depending on the reliability, validity and other influencing things. This could be narrative, descriptive, argumentative, expressive, role playing. In our study our data revealed that description was the most frequent one at the paragraph level and in the essay level the most frequent one was the focus on the cause. According to Brown (2004) it was important to design writing tasks. This could be either controlled or guided. From the data we got, we saw that all the participants gave place to either controlled or guided writing activities.

#### 4.6. ASSESSMENT OF LANGUAGE USE

In the sixth question, we tried to find out how the language use was assessed. The weight of language use in proficiency, placement and achievement tests was presented below in Table 4.26.

**Table 4.26.** Weight of language use in various tests

The Universities	Achievement Test %	Placement Test %	Proficiency Test %
Beykent University	-	25	No information
Bülent Ecevit University	27.6	75	80
Eskişehir Osmangazi University	40	30	55
Gazi University	-	-	35
Istanbul Technical University	25-35	60	15
Izmir Economy University	10	60	25
Izmir University	15	25	25
Karadeniz Technical University	45	45	45
Muğla University	40	70	40
Pamukkale University	20	65	20

Our data revealed that language use was very important for the participant schools except for two of them in the achievement test. Eight of the participant schools gave equal importance to language use in the achievement test. But the weight of them varied such as 40%, 20% and 10%. However, we could not state the same result for the placement test. For nine of the participant

schools, language use was important in the placement test with the variety between 75% and 25%. In the proficiency exam, all of the schools tested language use, and although one of the universities administered language use in the proficiency exam, there was no information provided for it. The weight of language use varied between 15% and 80%.

When we analyzed the item types used to assess language use, we found out that the most frequently used item type was multiple choice. This item type was often used in the language use part of achievement tests ( $m=4.20$ ;  $sd=1.23$ ). In addition, the following item types were sometimes used: writing appropriate form of the verbs in a cloze test, re-writing the sentence, completing the dialogue, filling the blanks, matching, sentence completion, find the mistake and correcting it, making sentences, making statements. However, odd one out type was rarely used (see Table 4.27).

**Table 4.27.** The item types used in testing language use

Item types	Mean	Sd	Participation level
1. Multiple Choice	4.20	1.23	Often
2. Writing appropriate form of verbs in a cloze test	3.40	1.50	Sometimes
3. Re-writing the sentence (transformation)	3.40	1.35	Sometimes
4. Completing the dialogue	3.40	1.35	Sometimes
5. Filling in the blanks	3.30	1.41	Sometimes
6. Matching	3.00	1.24	Sometimes
7. Sentence completion	2.90	1.37	Sometimes
8. Find the mistake and correcting it	2.50	1.35	Sometimes
9. Making sentences	2.40	1.35	Sometimes
10. Making statements	2.30	1.25	Sometimes
11. Odd one out	1.90	0.87	Rarely

Kitao (1996) suggested that testing grammar was one of the main strays of language testing because this underlied the ability to use language and the most common way was the multiple choice technique. In the data we got, we could easily see that the most frequently used item type was multiple choice. According to Hughes (2003) testing language use could be done with paraphrase, completion and modified cloze. However, these were not preferred in the participant universities. It was also important to test grammar integrated

with skills. Rea-Dickins (1997) pointed out that “if we didn’t consider it necessary to test grammar as distinct from the mentioned skills, this would raise negative washback on teaching and further lack of respect for teaching of grammar. If grammar were eliminated from testing what effect would this have on teaching?” From the questionnaire we administered, we could easily observe the language use in line with skills.

In Kırık’s study (2008), the attitudes of English teachers’ in terms of testing and assessment, the most frequently used item type was gap filling. Kırık (2008) also stated that this type of item was traditional item type. However in our study we found out that the most frequent one was multiple choice. Mc Marmara and Roever (2006) stated that in 1970s testing was typically done by means of decontextualized, discrete-point items such as sentence unscrambling, fill-in-the-blanks, error correction, sentence completion, sentence combining, picture description, elicited imitation, judging grammatical correctness, and modified cloze passages. Such formats tested grammar knowledge, but they did not assess whether test takers can use grammar correctly in real-life speaking or writing. Now it should be more open-ended, but they were subject to possible inconsistencies as Purpura (2006) stated. In our study from the data we got we could easily observe that, the participant universities applied multiple choice and secondly error correction which Roever called decontextualized.

#### **4.7. ASSESSMENT OF VOCABULARY**

In the seventh question, we tried to find out how vocabulary was assessed. The weight of vocabulary in proficiency, placement and achievement tests was presented below in Table 4.28.



**Table 4.28.** Weight of vocabulary in various tests

The Universities	Achievement Test %	Placement Test %	Proficiency Test %
Beykent University	20	25	No information
Bülent Ecevit University	9.2	15	10
Eskişehir Osmangazi University	30	30	25
Gazi University	-	-	15
Istanbul Technical University	10	-	-
Izmir Economy University	-	10	-
Izmir University	10	5	5
Karadeniz Technical University	30	30	30
Muğla University	-	15	-
Pamukkale University	15	5	5

Our data revealed that vocabulary was very important for the participant schools except for three in the achievement test. The weight of them varied such as 15%, 20% and 30%. However, we could not state the same result for the placement test. For all the participant schools, listening skill was not important in the placement test. Only eight of the participant universities give place to vocabulary in the placement test with the variety in weight such as 5% and 30%. It was also the same in the proficiency exam. Only seven of the schools tested listening skill in the proficiency exam and although one of the universities administered vocabulary in the proficiency exam, there was no information provided for it. The weight of vocabulary varied between 5% and 30%.

In the assessment of vocabulary, the participant schools prepared items for the following sub skills: deducing the meaning of new words from context, using the appropriate form of the words, finding synonym of a given word in a context, finding the function a word. Among the sub skills, deducing the meaning of new words from the context is the most frequent sub skills tested in the participant schools (see Table 4.29.). However, using the appropriate form of the words, finding synonym of a given word in a context, finding the function a word were sometimes tested.

**Table 4.29.** Assessment of vocabulary

Subskills	Mean	Sd	Participation level
1.Deducing the meaning of new words from context	4.00	1.15	Often
2.Using the appropriate form of the words	3.70	1.16	Sometimes
3.Finding synonym/antonym of a given word in context	3.20	1.40	Sometimes
4.Finding the function of the word in a context such as noun, adj. or verb,adv. etc.	3.00	1.33	Sometimes

When we analyzed the item types used to assess vocabulary, we found out that the most frequently used item type was cloze test with multiple choice. This item type was often used in the language use part of achievement tests ( $m=4.00$ ;  $sd=0.66$ ). In addition, the following item types were sometimes used: cloze test by choosing the appropriate word, filling in the blanks with appropriate words, matching words with their antonyms or synonyms, matching words with their definitions and using the words in a sentence. However, matching words with pictures, writing definition of word in English, writing Turkish definition or meaning of a given word types were rarely used. (see Table 4.30).

**Table 4.30.** The item types used in testing vocabulary

Item types	Mean	Sd	Participation Level
1.Cloze test with multiple choice	4.00	0.66	Often
2. Multiple Choice	4.00	1.24	Often
3.Cloze test by choosing the appropriate word from a word box.	3.80	1.13	Sometimes
4. Filling in the blanks with appropriate words	3.60	1.09	Sometimes
5. Matching words with their synonyms or antonyms	2.60	1.26	Sometimes
6. Matching words with their simple definitions	2.50	1.18	Sometimes
7. Using the words in a sentence	2.10	1.37	Sometimes
8. Matching words with pictures	1.70	0.95	Rarely
9. Writing definition of a word in English	1.40	0.70	Rarely
10. Writing Turkish definition or meaning of the given word	1.10	0.31	Rarely

In the assessment of vocabulary, our data revealed that the most frequently used test item was multiple choice and cloze procedure, which was parallel to the study of Kırık (2008). In both studies, matching with definitions was also popular. According to Hughes (2003), knowledge of vocabulary was

essential to the development and demonstration of language skills. But that did not mean that it should be tested separately. Our data revealed that the participant universities did not apply a separate vocabulary test.

Hughes (2003) pointed out that the item types used in vocabulary testing could be synonyms and definitions. These types were sometimes used in the participant universities. In analyzing assessment tasks in archival vocabulary studies, Scott et al. (2006) reported that most researchers had devised assessments that tested knowledge of the specific words that had been taught in an instructional intervention. The only common construct underlying word selection across a majority of studies was students' prior knowledge. That is, it was assumed that the words taught, or at least the majority of them, were unknown to the target students. This assumption was validated in one of three ways: (a) by using a pretest that tested each word directly, (b) by selecting words with a low p value (percent correct) from a source such as *The Living Word Vocabulary* (Dale & O'Rourke, 1981), or (c) by asking teachers or researchers to select words not likely to be known by the target population. The result of our study was similar to that of their study. Like this study, deducing the meaning from the context was the most frequent used subskill.

## **CHAPTER FIVE**

### **CONCLUSION**

#### **5.1. INTRODUCTION**

In this chapter, a summary of the study has been presented with general conclusions. The implications of the study on school of foreign languages, testing offices, students and any other parties have been discussed. Finally, suggestions for further studies have been presented.

#### **5.2. SUMMARY OF THE STUDY**

The study aimed to find out assessment and evaluation activities in schools of foreign language in Turkey. The data have been gathered through a questionnaire. Analysis of our data revealed that students have often been assessed through written exams. The results showed that all the participant schools give all level of the tests except C2 level according to CEFR. They administered placement, proficiency and achievement tests at least once a year. The listening skill was equally important for all participants, and they usually preferred understanding main idea, skimming and information transfer subskills with multiple choice question types. On the other hand, watching was almost never used as part of any test types.

Reading skill was also tested through skimming, scanning, understanding main idea and referencing subskills. All these sub skills were assessed with multiple choice and true false items. However, translation was ignored by all of the participant universities.

Our results revealed that the schools give importance to production skills as they assess both speaking and writing skills. Speaking was assessed in all institutions. In order to assess this skill, generally descriptive talk and a dialogue on a topic in an interview format were preferred, and students were usually tested one by one.

Writing skill was also assessed in terms of paragraph level and essay level. While assessing the skills, the participant universities did not ignore the language level. They administered writing skill either at paragraph or essay level.

Language use and vocabulary was not ignored and in the assessment the participant schools assessed vocabulary in terms of synonym, antonym activities, finding the function of the word etc. Deducing the meaning of new words from the context was the most frequent sub skills tested in vocabulary skill. Language use was assessed mostly with multiple choice item type. They did not give separate vocabulary or language skill test. However, from the data we got, we learned that three universities did not apply language use test.

From this point of view, it has been seen that all the skills are interrelated to each other. Our results showed that assessment was very important and all skills should be assessed in line with their subskills, and this revealed the importance of the backwash effect because any skill which was not tested was ignored by the students.

### **5.3. IMPLICATION OF THE STUDY**

The aim of this study is to bring about positive washback effects on teaching and learning in schools while describing the assessment and evaluation activities. As a result of this study, we realize that skills should be tested equally because there is a washback effect, which is the effect of test items on teaching and learning, the educational system and the various stakeholders in the education process.

The testing coordinators and the staff in the testing office should be careful about the content of the tests. The tests which are going to be applied should cover all skills equally because students need all skills in their academic life. All skills should be taught and assessed equally. Thus, they will create a positive backwash effect on students and teachers. On the other hand, test designers should give importance to the principles of language assessment such as validity, reliability, washback, practicality, and authenticity. All of these are the essential parts of any type of test. Ignorance of even one of these features will create some problems and result in negative backwash effect on both students and teachers.

Teachers will focus on four language skills in their teaching. They will design their lesson plan, teaching activities and materials in line with the test items asked in the exams. They will assign homework and provide feedback in accordance with the assessment items. This approach in assessment will create a positive backwash effect on teachers because the weight of each skill and the items representing each subskill will influence how much time they spend, how many activities or tasks they do, and how much importance they give them in detail during the classes or when they assign outdoor tasks and activities.

Students will see the benefits of learning and using four language skills rather than having the knowledge about the foreign language. Students who score well on the tests will feel a sense of pride and accomplishment. Schools, teachers and parents often publicly praise these students for their achievement. Rather than pressure, they will have the pleasure of learning English practically as they usually have instrumental motivation towards learning English.

#### **5.4. SUGGESTIONS FOR FURTHER RESEARCH**

While the present study describes the assessment and evaluation activities in the school of foreign languages in Turkey, more studies are needed to see the long-term effect of this application on students' learning English and

their effective use in their academic life. Besides, it might be interesting to observe the effects of such an application on larger groups. For that reason, the study could be replicated with larger and more diverse participants in different universities.

## REFERENCES

- Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, 13, 280-297
- Alderson, J.C. (2002). *Language testing and assessment*. Lancaster University, UK.
- Alderson, J. C. and Wall, D. (1993). Does washback exist? *Applied Linguistics* 14, (2), June 1993.
- Al-Weher, M. (2004). The effect of a training course based on constructivism on student teachers' perceptions of the teaching/learning process *Asia-Pacific Journal of Teacher Education*, vol: 32, 169- 184.
- Andrews, S. (1994). The washback effect of examinations: Its impact upon curriculum innovation in English language teaching. *Curriculum Forum*, 4(1), 44-58.
- Andrews, S. (1995). *Washback or washout? The relationship between examination reform and curriculum innovation*. In D. Nunan, V. Berry & R. Berry (Eds.), *Bringing about change in language education* (pp.67-81). Hong Kong: University of Hong Kong.
- Ari, E. (2002). The effects of university examination which was performed on the education in chemistry department in faculty of science and arts. *Celal Bayar University. Chemistry Department*.
- Arnove, R.F., Altback, P. G., & Kelly, G. P. (Eds.). (1992). *Emergent issue in education: Comparative perspectives*. Albany, NY: State University of New York Press.
- Ausubel, D. (1965). Introduction to part one. In R. Anderson & D. Ausubel (Eds.), *Readings in the psychology of cognition* (pp. 3-17). New York: Holt, Rinehart & Winston.
- Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F., Lynch, B.K. and Mason, M. (1995): Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing* 12, 238-57.



- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bailey, K. M. (1996). Working for wash back: A review of the wash back concept in language testing. *Language Testing*, 13, 3, 257-279.
- Baker, E. L. (1991). Issues in policy, assessment, and equity. *Paper presented at the national research symposium on limited English proficient students' Issues: Focus on evaluation and measurement*. Washington, DC.
- Berry, V. (1994). Current assessment issues and practices in Hong Kong: A preview. In D. Nunan, R. Berry, & V. Berry, (Eds.), *Bringing about change in language education: Proceedings of the International Language in Education Conference 1994* (pp. 31-34). Hong Kong: University of Hong Kong.
- Biggs, J. B. (1995). Assumptions underlying new approaches to educational assessment. *Curriculum Forum*, 4 (2), 1-22.
- Biggs, J. B. (Ed). (1996). *Testing: To educate or to select? Education in Hong Kong at the cross-roads*. Hong Kong: Hong Kong Educational Publishing.
- Boston, C. (2002). The Concept of Formative Assessment. *Eric Digest*, vol: 8, pp. 101- 105.
- Boyluğ, M. (2003). An analysis of the compatibility of the reading activities in the FLE classes at high schools in Gaziantep and the foreign language section of the student selection examination (OSS). *Unpublished MA Dissertation, University of Gaziantep, Gaziantep*.
- Bracey, G. W. (1989). *The \$150 million redundancy*. Phi Delta Kapa, 70, 698-702.
- Breland, H. M. (1983). *Linear Models of Writing Assessments*. ETS Research Report, Educational Testing Service, Princeton, N.J.
- Brookhart, S.M. (2001). Successful Students' Formative and Summative Uses of Assessment Information. *Assessment in Education: Principles, Policy and Practice*, Vol. 8, No. 2, pp.153-169.
- Brooks, J.G. & Brooks, M.G. (2001). *In search of understanding the case for constructivist classrooms*, New Jersey: Merrill Prentice Hall.
- Brown, D.H.(2001). *Teaching by principles*. San Francisco: Longman.

- Brown, D. H. (2004). *Language Assessment: Principles and Classroom Practices*. USA: Pearson Education Ltd.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- Buck, G. (1992). Translation as a language testing procedure: does it work? *Language Testing*, 9, 123-148.
- Butler, S.M. & Mcmunn, N.D. (2006). *Understanding and Using Assessment to Improve Student Learning*. PB Printing: United States of America.
- Calder, M. C. (1997). *Juveniles and children who sexually abuse: a guide to risk assessment*. Lyme Regis, Dorset: Russell House Publishing.
- Cameron, L.(2005). *Teaching Languages to Young Children*. Cambridge: Cambridge University Press.
- Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: how all 50 states are above the national average*. Second edition. Daniels, west Virginia: Friends for Education.
- Carroll, J. B. (1968). *The psychology of language testing* in Davies 1968a. pp: 46-69.
- Çaykan, I. (2001). *Material development for testing oral skills communicatively for intermediate students*. Unpublished M.A. dissertation, Hacettepe University.
- Chastain, K. (1976). *Developing second- language skills: Theory and practice*. Chicago: Rand McNally College Publishing.
- Cheng, L. (1997). *How does washback influence teaching? Implications for Hong Kong*. *Language Education*, 11, 38–54.
- Chinda, B. (2009). Professional development in language testing and assessment: A case study of supporting change in assessment practice in in-service EFL teachers in Thailand (Doctoral dissertation). *University of Nottingham, UK*.
- Ciel Language Support Network. (2000). *Assessment and independent language learning*. Available at <http://www.llas.ac.uk/resourceid=1407>

- Cohen, A. D. (1994). *Assessing language ability in the classroom (Second edition)*. New York: Heinle & Heinle.
- Cooley, W. W. (1991). *State-wide student assessment*. *Educational Measurement Issues and Practice*, 10, 3-6.
- Coombe, C., Folse, N. and Hubley, N. (2007). *A Practical Guide to Assessing English Language Learners*. USA: The University of Michigan Press.
- Cowie, B. & Beverley, B. (1999). Model of Formative Assessment in Science Education. *Assessment in Education*, Vol. 6, No. 1, pp.101-116.
- Cronbach, L. J., and Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52, 281-302.
- Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research*, 58 (4), 438–481.
- Curtis, A. (2000). A problem-solving approach to the management of change in language education. *Korea TESOL Journal*, 3(1), 1-12.
- Darling, H. (1996). *A license to teach : The foundation for 21st Century Schools*.
- Davies, A.(1988). *Communicative language testing*. In Hughes 1988b.
- Davies, A. (1968) . *Language Testing Symposium: A Psycholinguistic Approach*. London: Oxford University Press.
- Davies, A. (1995). *Communicative Language Testing*. In Sheldon, Leslie E. (Ed.), *Testing English for university study: ELT document 127*. (pp.5-15). United Kingdom: Modern English Publications in association with the British Council.
- Dembo, M. H. (1994). *Applying Educational Psychology (5Th edition)*. Longman.
- Douglas, D. (1997). Language for specific purposes testing. In Clapham, C. and Carson, D., editors, *Encyclopedia of language in education. Volume 7: Language testing and assessment*. Dordrecht: Kluwer Academic, 111–20.
- Driscoll, M. (2000). *Psychology of Learning for Instruction*. Needham Heights, MA, Allyn & Bacon.

- Duran ,O. (2011). Teachers' and Students' Perceptions about Classroom-Based Speaking Tests and Their Washback. *Unpublished MA Dissertation, University of Bilkent, Ankara.*
- Elton, L., & Laurillard, D. (1979). Trends in student learning. *Studies in Higher Education, 4*, 87–102.
- Fish, J. (1988). Responses to mandated standardized testing .*Unpublished Doctoral dissertation, University of California, Los Angeles.*
- Frederiksen, J.R. & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher, 18, (9), 27-32.*
- Freeman, D. (1996). Redefining the relationship between research and what teachers know. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 88-115). Cambridge: Cambridge University Press.
- Fisher, D. & Frey, N. (2007). *Checking for Understanding, Formative Assessment Techniques for Your Classroom.* ASCD Publication: United States of America.
- Fullilove, J. (1992). The tail that wags. *Institute of Language in Education Journal, 9*, 131-147.
- Gagne, R. M. (1965). *The conditions of learning.* New York: Rinehard & Winston.
- Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In *Changing assessments: Alternative views of aptitude, achievement and instruction* B. R. Gifford & M. C. O'Connor (Eds), (pp.77-119). London: Kluwer Academic.
- Gatewood, R. D., Field, H. S. (2001). *Human Resource Selection.* South-Western: United States.
- Gay, L.R. (1992). *Educational research competencies for analysis and application.* Colombus, Chlo: publishing company and A.Bell & Hohwelt Company.
- Gifford, B. R., & O'Connor, M. C. (Eds). (1992). *Changing assessments: Alternative views of aptitude, achievement and instruction.* London: Kluwer Academic.

- Glaser L.R. (1990). *Testing and assessment Pittsburgh, P.A.* Learning Research and Development Center.
- Good, T. L., Brophy, J. E. (1990). *Educational psychology: A realistic approach.* (4th ed.). White Plains, NY: Longman.
- Gronlund, N. E. (1998). *Assessment of student achievement.* Sixth edition. Boston: Allyn and Bacon.
- Gronlund, N.E., & Waugh, K. C. (2009). *Assessment of Student Achievement.* Upper Saddle River, New Jersey: Pearson, Education, Inc.
- Güllüoğlu, Ö. (2004). Attitudes towards testing speaking at Gazi University Preparatory school of English and suggested speaking tests. *Unpublished M.A. dissertation. University of Gazi. Ankara.*
- Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20(5), 2-7.*
- Harlen, W. (2003). *Enhancing Inquiry Through Formative Assessment.* Exploratorium: USA.
- Heyneman, S. P. (1987). Use of examinations in developing countries: Selection, research, and education sector management. *International Journal of Education Development, 7, 251-263.*
- Heyneman, S. P., & Ranson, A. W. (1990). Using examinations and testing to improve educational quality. *Educational Policy, 177-192.*
- Heaton, J. B. (1990). *Writing English Language Tests.* (New Edition). London.
- Heaton, J. B. (1975). *Writing English Language Tests.* London.
- Henning, G. (1987). *A guide to language testing: Development, evaluation and research* Cambridge, Mass: Newbury House.
- Heritage, M. (2007). Formative Assessment: What Do Teachers Need to Know and Do? *Phi Delta Kappa, vol.89, pp.140.*
- Hinkel, E. (2006). Current perspectives in teaching the four skills. *Tesol Quaterly 40/1.*
- Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

- Hubley, A. M. & Zumbo, B. D. (1996). A dialectic on validity: where we have been and where we are going. *The Journal of General Psychology* 123,3, 207-215.
- Irons, A. (2008), *Enhancing Learning Through Formative Assessment and Feedback*, Routledge: U.S.A and Canada.
- Johnson, L.(2007) . *Education World* retrieved from :  
[https://www.educationworld.com/a\\_curr/profdev/profdev114w.shtml](https://www.educationworld.com/a_curr/profdev/profdev114w.shtml)
- Katsumasa, S. (1997). Communicative language testing: Principles and problems. *English Review*, 12, 3-24. Retrieved December 1, 2007, from <http://ci.nii.ac.jp/naid/110004693443/en/>.
- Kehaghan T. & Greaney, V. (1992). *Using examinations to improve education: A study of fourteen African countries*. Washington DC: The World Bank.
- Khamkhien, A. (2010). Teaching English speaking and English speaking tests in The Thai context: A reflection from Thai perspective. *English Language Teaching*, 3(1), 184-190.
- Kırık, M.(2008). Yabancı dil olarak İngilizce öğretmenlerinin ölçme değerlendirme bağlamında tutum ve yaklaşımları. *Unpublished doctorate dissertation. University of Istanbul, Istanbul*.
- Kirschenbaum, H. (2003). *Carl Rogers and the Person-Centered Approach. A 60- minute videotape/DVD presentation*. Webster, NY: Values Associates.
- Kitao , K. (1996). Testing Listening. *Internet Tefl Journal*.  
 (retrieved from : <http://iteslj.org/Articles/Kitao-TestingListening.html>)
- Kitao , K. (1996). Testing Reading. *Internet Tefl Journal*.  
 (from: <http://www.cis.doshisha.ac.jp/kkitao/library/article/test/reading.htm>)
- Kitao , K. (1996). Testing Speaking. *Internet Tefl Journal*.  
 (from:<http://www.cis.doshisha.ac.jp/kkitao/library/article/test/speaking.htm>)
- Kitao , K. (1996). Testing Writing. *Internet Tefl Journal*.  
 (<http://www.cis.doshisha.ac.jp/kkitao/library/article/test/writing.htm>)
- Kitao , K. (1996). Testing Grammar. *Internet Tefl Journal*.  
 (retrieved from:<http://iteslj.org/Articles/Kitao-TestingGrammar.html>)

- Krause K., Bochner, S., Duchesne, S. (2003). *Educational Psychology for Learning and Teaching. Teaching, Learning and Assessment: The Road to Democracy*. Nelson: Australia.
- Köksal, D. (2004). Assessing Teachers' Testing Skills in ELT and Enhancing their Professional Development through Distance Learning on the Net. *Turkish Online Journal of Distance Education-TOJDE*. Volume: 5 Number: 1. [http://tojde.anadolu.edu.tr/tojde13/pdf/koksal\\_pdf.pdf](http://tojde.anadolu.edu.tr/tojde13/pdf/koksal_pdf.pdf)
- Lambert, D. & Lines, D. (2000). *Understanding Assessment ,Purposes, Perceptions and Assessment*. Routledge Falmer: U.S.A and Canada.
- Lamon, M. (2007). Learning Theory - Constructivist Approach, "Life-long Learning." Available at <http://education.stateuniversity.com/pages/2174/Learning-Theory> www.wikipedia.com, 25.01.2013.
- Li, X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development*, 11, 393-404.
- Linn, M. C. (2000). *Assessment and accountability: what kinds of assessment are used and for what purposes?* The National Academies Press.
- Linn, R. L. (1992). *Educational assessment: Expanded expectations and challenges*(Tech. Rep. 351). Boulder: University of Colorado at Boulder, Center for the Study of Evaluation.
- Lightbown, P., & Spada, N. M. (2006). *How languages are learned*. Oxford England:Oxford University Press.
- Lunzer, E., Waite,M. and Dolan,T.(1979) Comprehension and comprehension tests in Lunzer, E. and Gardner, K. (eds) *The Effective Use of Reading Heinemann Educational Books*, pp 37-71.
- Lynch, B. K. (1996). *Language program evaluation: theory and practice*. Cambridge: Cambridge Univ. Press.
- Mackey, A. & Gass, S. M. (2005). *Second Language Research: Methodology and Design*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Madaus, G. F. (1988). The influence of Testing on the Curriculum. *In L. N. Tanner (ed), Critical Issues in Curriculum: 87th Yearbook for the National Society for the Study of Education*, (pp. 83-121). Chicago: University of Chicago Press.

- Madaus, G. F. (1985). The Irish experience in competency testing: Implications for American education. *American Journal of Education* 93(2), 268-94.
- Maslow, A.H. (1968). *Towards a Psychology of Being*. Van Nostrand, Princeton :New Jersey.
- McNamara, T. and Roever, C.(2006). *Language testing: The social dimension*.(Language Learning and Monograph Series). Malden, MA and Oxford, UK: Blackwell Publishing, pp. 291.
- Mcmillan, J.H. (2007). *Formative Classroom Assessment, Theory Into Practice*. Teachers College Press: United States of America.
- Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing* 13, pp. 241-256.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*(3rd ed., pp. 13–103). NewYork: Macmillan.
- Messick, S. (1992, April). *The interplay between evidence and consequences in the validation of performance assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Mergel, B. (1998). *Instructional design and learning theory*. Educational Communications and Technology. University of Saskatchewan.
- Ming, T.Y.K. (2002). Assessment Strategies that Maximize Our EPCL Students' Potential. *Student Assessment and Feedback*, vol: 3, pp.11-12.
- Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing: Proceedings of the IUS/NFER conference* (pp. 1–13). London: NFER/Nelson.
- Noble, A. J., & Smith, M. L. (1994a). *Measurement-driven reform: Research on policy, practice, repercussion* (Tech. Rep. 381). Tempe, AZ: Arizona State University, Center for the Study of Evaluation.
- Oksen, H. (1999). An investigation of the content validity and backwash effect of The end-of-term oral assessment test administered at Hacettepe University, Department of Basic English. *Unpublished MA Dissertation, Hacettepe University, Ankara*.



- Ökten, A. (2009). Effects of formative assessment application on students' language Proficiency in language learning in e.f.l context: a case study. *Unpublished MA Dissertation, University of Çukurova, Adana.*
- Pearson, I. (1988). Tests as levers for change. *In D. Chamberlain and R.J. Baumgardner (eds.)ESP in the classroom: practice and evaluation* (pp.98-107) Modern English Publications.
- Petrie, S. (1987). *The Power to Shape our Future. Social Work Today*, BASW, London, April 1987, p. 1.
- Pierce, B. N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly* 26(4), 665-691.
- Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, R.D., Willaims, P.L., (1985) Measurement driven instruction: It is on the road. *Phi Delta Kappan* 66, 628-634.
- Popham W.J. (1983). Measurement as an instructional catalyst *New Directions for Testing and measurement : Measurement, technology and individualization in education.* pp:19-30.
- Purpura, J. (2006). Issues and challenges in measuring SLA. *Paper presented At the American Association for Applied Linguistics Conference, June, Montreal.*
- Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing* 14,,3 pp: 304-314.
- Rogers, Carl. (1969). *Freedom to Learn: a view of what Education Might Become.* C.E.Merrill: Collumbus, Ohio.
- Rubin, A., and Babbie, E. (1993) *Research Methods for Social Work.* (second edition) California: Cole Publishing.
- Rudman, H. (1989). Integrating testing with teaching. *Practical Assessment, Research & Evaluation*, 1(6), 1-4.
- Saehu, A. (2012). Testing and Its Potential Washback. *In Bambang Y. Cahyono. And Rohmani N. Indah (Eds.), Second Language Research and Pedagogy.* pp. 119-132. Malang: State University of Malang Press.

- Sarıçoban, A. (2011). A study on the English Language teachers' preparation of tests. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education)* 41: 398-410.
- Schreiner, R. (1977). *Assessing reading comprehension*. Miami Beach: Florida.
- Scott, J.A., Lubliner, S., & Hiebert, E.H. (2006). Constructs underlying word selection and assessment tasks [plato4]in the archival research on vocabulary instruction. In J.V. Hoffman, D.L. Schallert, C.M. Fairbanks, J. Worthy, & B. Maloch (Eds.), *55th yearbook of the National Reading Conference* (pp. 264–275). Oak Creek, WI: National Reading Conference.
- Sevimli, S. (2007). The washback effects of foreign language component of the University entrance examination on the teaching and learning context of Language groups in secondary education ( a case study). *Unpublished MA Dissertation, University of Gaziantep, Gaziantep*.
- Sentürk, F. (2013). Washback effect of Ket exam in learning English as an Foreign Language. *Unpublished MA Dissertation, University of Çağ, Mersin*.
- Shepard, L.A. (1989). *Inflated test score gains: is it old norms or teaching the test?* Center for the study evaluation, university of California, Los Angeles.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8, 13, 24.
- Shohamy, E. (1993a). *The power of test: The impact of language testing on Teaching and learning*. Washington, DC: National Foreign Language Center Occasional Papers. The National Foreign Language Center, Washington, DC.
- Shohamy, E., Donitsa-Schmidt, S.,& Ferman, I. (1996). Test Impact revisited: Washback effect over time. *Language Testing*, 13, 298-317.
- Skinner , B.F.(1957). *Verbal behavior*. New York: Appleton-Century-Crofts.
- Smith, M. L. (1991). Put to test: The effects of external testing on teachers. *Educational Researcher*, 20 (5), 8-11.
- Sook, K. H. (2003). The types of speaking assessment tasks used by Korean Junior Secondary school English teachers. *Asian EFL Journal*. Retrieved from [http://www.asian-efl-journal.com/dec\\_03\\_gl.pdf](http://www.asian-efl-journal.com/dec_03_gl.pdf).

- Sönmez, V. (1994). *Program geliştirmede öğretmen el kitabı*, Ankara: PEGEM Yayınları. No :12, 7. Basım.
- Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.
- Spratt, M. (2005). Washback and classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research* 9, 1(pp. 5- 29).
- Steadman, M. (1998). Using classroom assessment to change both teaching and learning new directions for teaching and learning. Jossey-Bass Publishers no:75.
- Stecher, B., Chun, T.,& Barron, S. (1999). *Quadrennial milepost accountability Testing in Kentucky (Tech. Rep. 505)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student testing.
- Stevenson, D. K., Riewe, U. (1981). Teachers' attitudes towards language tests and testing. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing. Occasional Papers*, 26 (pp. 146-155). Essex, UK: University of Essex.
- Stiggins, R. (2005). *From Formative Assessment to Assessment for Learning: A path .to Success in Standards-Based Schools*. Available at [www.assessmentinst.com](http://www.assessmentinst.com)
- Tambini, R. (1999). Aligning Learning Activities and Assessment Strategies in the ESL Classroom. *Tesl Internet journal*.
- Taylor, L. (2005) . Washback and impact. *ELT Journal*, 59, 2, pp 154-155.
- Temel, A. (2007). *Eğitimde Ölçme ve Değerlendirme*. Maltepe Üniversitesi.
- Travers, R.M.W. (1978). An introduction to educational research. The Mcmillan Company. New York.
- Tunçel, E. (1995). Communicative language testing. *Unpublished M.A. dissertation, Gazi University*.
- Upshur, J. A. (1971). *Productive communication testing: a progress report in G, Perren and J.L. M. Trim (eds.): Applications in Linguistics*. Cambridge: Cambridge University Press:43542.
- Ur, P. (1984). *Teaching Listening Comprehension*. Cambridge: Cambridge University Press.

- Vernon, P.E. (1956). (2nd edition) *The measurement of abilities*.  
London : University of london Press.
- Wall, D. (1997). Impact and washback in language. In C. Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7*. The encyclopedia of language and education. Dordrecht, Holland: Kluwer Academic. 291–302.
- Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe—Phase 2, coping with change*. Princeton, NJ: ETS.
- Watanabe, Y. (1996). Does grammar translation come from the entrance examination Preliminary findings from classroom-based research. *Language Testing, v. 13*, p. 318- 332.
- Watanabe, Y. (2001). Does the university entrance examination motivate learners? A case study of learner interviews. In Akita Association of English Studies (Eds.), *Trans-equator exchanges: A collection of academic papers in honor of Professor David Ingram*(pp. 100–110). Akita, Japan: Author.
- Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*, 194-209.
- Weir, C. J. (1990). *Communicative language testing*. Wiltshire: Prentice Hall.
- Williams, M. & Burden, R.L. (1997). *Psychology for Language Teachers: A Social Constructivist Approach*, Cambridge: Cambridge University Press.
- Williams, C. (2011). Research Methods. *Journal of Business & Economic Research – March 2007 Volume 5, Number 3*.
- Woodford, P.E. (1980). Foreign Language Testing. *The Modern Language Journal, 64*: 97-102.
- Yoon, B. (1996). *Instructional validity*. University of California, Los Angeles, CA.

**APPENDIX**  
**QUESTIONNAIRE**

**Dear Director / Head of the Department,**

I am carrying out a research with my MA student regarding assessment types, contents, frequency and procedures in the School of Foreign Languages in Turkish Universities. We would be very happy if the testing coordinator could fill out the questionnaire below and send it to the e-mail address of the student: [gulcedursun@hotmail.com](mailto:gulcedursun@hotmail.com). Thank you very much for your interest and contribution.

Asst. Prof. Dr. Turan Paker

Gülce Dursun

Pamukkale University

Pamukkale University

PART I. The School of Foreign Languages/Foreign Languages Department:

1) is in the university of : \_\_\_\_\_

**In your school:**

2) the number of instructors (full time + part time): \_\_\_\_\_

3) the number of students in the 2012-2013academic year: \_\_\_\_\_

4) the levels which you give test during the year: A1\_\_ A2\_\_ B1\_\_ B2\_\_ C1\_\_C2\_\_

5) the number of a placement test administered in an academic year: \_\_\_\_\_

6) the number of a proficiency test administered in an academic year: \_\_\_\_\_

7) the number of an achievement test administered in an academic year: \_\_\_\_\_

8) the number of quizzes administered in a semester depending on the skills

Skills	the number of the Quiz
1. Listening	_____
2. Watching	_____
3. Reading	_____
4. Writing	_____
5. Language Use	_____
6. Vocabulary	_____
7. Grammar	_____
8. Translation	_____

9) Please tick the skills you assess in a typical achievement and write down their weight.

<u>Skills</u>	<u>Weight</u>
Listening _____	_____
Watching _____	_____
Reading _____	_____
Language Use _____	_____
Vocabulary _____	_____
Writing _____	_____
Grammar _____	_____
Translation _____	_____
speaking _____	_____

10) Please tick the skills you assess in a typical placement test and write down their weight.

<u>Skills</u>	<u>Weight</u>
Listening _____	_____
Watching _____	_____
Reading _____	_____
Language Use _____	_____
Vocabulary _____	_____
Writing _____	_____
Grammar _____	_____
Translation _____	_____
speaking _____	_____

11) Please tick the skills you assess in a typical proficiency test and write down their weight.

<u>Skills</u>	<u>Weight</u>
Listening _____	_____
Watching _____	_____
Reading _____	_____
Language Use _____	_____
Vocabulary _____	_____
Writing _____	_____
Grammar _____	_____
Translation _____	_____
speaking _____	_____

**PART II) Please tick the subskills of the skills you assess in various achievement tests and the item types you use depending on the frequency.**

	<b>Always</b>	<b>Often</b>	<b>Some times</b>	<b>Rarely</b>	<b>Never</b>
<b>I.Listening/Watching</b>					
1. Skimming					
2. Scanning					
3. Guessing the title					
4. Understanding the main idea					
5. Referencing					
6. Dictation/note taking (word/phrase level)					
7. Note taking (guided)					
8. Guessing the meaning of unknown words/phrases					
9. Information transfer					
10. Inferencing					
11. Speaker's attitude or opinion					
12. Identifying facts or opinions					
13. Recognizing discourse markers, patterns					
14. Other specify.....					
<b>I.A. Item Types Used</b>					
1. Multiple Choice at word/phrase level					
2. Multiple Choice at sentence level					
3. Multiple Choice in a cloze test					
4. True/False					
5. Open-ended items					

	Always	Often	Some times	Rarely	Never
6. Matching at word/phrase level					
7. Matching at sentence level					
8. Matching at paragraph level					
9. Filling out a form/a table					
10. Sequencing sentences to make a summary					
11. Sequencing the paragraphs					
12. Sequencing the sentences to put in a correct order					
13. Sentence completion					
14. Choosing the irrelevant statement in a paragraph					
15. Placing the appropriate sentence in a paragraph					
16. Writing a response to a given situation					
17. Finding the paraphrased statement					
18. Filling in the blanks at word/phrase level					
19. Filling in the blanks at sentence level					
20. Completing the dialogue with MC					
21. Summarizing the text					
<b>II. Reading</b>					
1. Skimming					
2. Scanning					
3. Guessing the title					
4. Understanding the main idea					
5. Referencing					
6. Outlining (paragraph level)					
7. Outlining (text level)					
8. Guessing the meaning of unknown words/phrases					
9. Information transfer					
10. Inferencing					
11. Speaker's attitude or opinion					
12. Identifying facts or opinions					
13. Recognizing discourse markers, patterns					
14. Other specify.....					



<b>II.A. Item Types Used</b>	<b>Always</b>	<b>Often</b>	<b>Some times</b>	<b>Rarely</b>	<b>Never</b>
1. Multiple Choice at word/phrase level					
2. Multiple Choice at sentence level					
3. Multiple Choice in a cloze test					
4. True/False					
5. Open-ended items					
6. Matching at word/phrase level					
7. Matching at sentence level					
8. Matching at paragraph level					
9. Filling out a form/a table					
10. Sequencing sentences to make a summary					
11. Sequencing the paragraphs					
12. Sequencing the sentences to put in a correct order					
13. Sentence completion					
14. Choosing the irrelevant statement in a paragraph					
15. Placing the appropriate sentence in a paragraph					
16. Writing a response to a given situation					
17. Finding the paraphrased statement					
18. Filling in the blanks at word/phrase level					
19. Filling in the blanks at sentence level					
20. Completing the dialogue with MC					
21. Summarizing the text					
<b>III. Writing</b>					
<b>A. Paragraph Level</b>					
1. Description					
2. Opinion					
3. Comparison/contrast					
4. Cause and effect					
<b>III.A. Item Types Used</b>					
1. Controlled writing task					
2. Guided writing task					
3. Free writing task					
4. Writing an e-mail					
5. Writing a letter of advice					
6. Writing an application letter					
7. Making an outline of the given text					
8. Writing notes about daily issues					

	Always	Often	Some times	Rarely	Never
9. Re-writing some sentences (paraphrase)					
10. Finding and Correct the misspelled words					
11. Putting the cohesive words in appropriate place					
12. Writing a topic sentence/ conclusion sentence in a text					
13. Other specify.....					
<b>B. Essay Level</b>					
1.Descriptive					
2.Classification					
3.Focus on the cause					
4.Focus on the effect					
5.Comparison/Contrast					
6. Problem solution					
7. Opinion					
8. Argumentative					
9. Narrative					
<b>III.B. Item Types Used</b>					
1.Controlled writing task					
2.Guided writing task					
3.Free writing task					
<b>IV. Speaking</b>					
1. Descriptive					
2. Narrative					
3. Having a dialogue on a topic					
4. Comparison/Contrast					
5. Problem solution					
6. Cause and effect					
7. Persuasive Talk					
8. Argumentative					
9. Talking about a process (how to make a cake, etc.)					
10. Giving a presentation on a topic					

	Always	Often	Some times	Rarely	Never
<b>IV. A. Item Types</b>					
1. Picture Talk					
2. Having a dialogue on a topic					
3. Choosing a topic and talking about it					
4. Summarizing a film/a story/a novel					
<b>IV. B Application</b>					
1. Testing one student at a time					
2. Testing two students at a time					
3. Testing three students at a time					
4. Testing a group of 4 or 5 students at a time					
5. Recording all the students					
<b>V. Translation</b>					
1. From Turkish to English					
2. From English to Turkish					
<b>VI. Item Types:</b>					
1. Multiple Choice sentence level					
2. Paragraph translation					
3. Essay translation					
<b>Language Use</b>					
<b>A. Structural use</b>					
<b>B. Functional use</b>					
<b>VI. A. Item Types</b>					
1. Multiple Choice					
2. Writing appropriate form of verbs in a cloze test					
3. Filling in the blanks					
4. Completing the dialogue					
5. Sentence completion					
6. Making statements					
7. Making sentences					
8. Odd one out					
9. Matching					
10. Open-ended items					
11. Re-writing the sentence (transformation)					
12. Find the mistake and correcting it					
13. Other specify.....					
<b>VII. Vocabulary</b>					
1. Deducing the meaning of new words from context					
2. Finding synonym/antonym of a give word in context					
3. Using the appropriate form of the					

words					
	<b>Always</b>	<b>Often</b>	<b>Some times</b>	<b>Rarely</b>	<b>Never</b>
4.Finding the function of the word in a context such as noun, adj. or verb, etc.					
<b>VII. A. Item Types</b>					
1.Cloze test with multiple choice					
2.Cloze test by choosing the appropriate word from a word box.					
3. Filling in the blanks with appropriate words					
4. Matching words with pictures					
5. Matching words with their simple definitions					
6. Matching words with their synonyms or antonyms					
7. Multiple Choice					
8. Writing definition of a word in English					
9. Writing Turkish definition or meaning of the given word					
10. Using the words in a sentence					
11. Other specify.....					

**Thank you for your contribution!**

-

## CURRICULUM VITAE

### PERSONAL DETAILS

**Name-Surname** : Gülce Dursun

**E-mail adres** : [gulcedursun@hotmail.com](mailto:gulcedursun@hotmail.com)

**Date of birth** : 14th November 1989

**Date of birthplace** : Muğla

### EDUCATION

#### 2011 Oxford University Webinar

- Designing Good Tests: Principles into Practice
- Activating Speaking with Young Learners

#### 2011 Shaping the Way We Teach English Fifth Webinar

- The Musical Classroom: Teaching English with Tunes
- Towards a Pedagogy of Peace
- Large Classes: Tips & Techniques for Teachers
- Phrasal Verbs: Why We Have Them and How to Teach Them
- 50 years of English Teaching Forum: Teachers Collaborating Worldwide

#### 2011 Shaping the Way We Teach English Fourth Webinar

- Exploiting Literature in Project Based Learning
- The Question of Grammar in the English Classroom
- More Critical Thinking: Leading your Class to English Language Awareness through Questioning
- Understanding & Applying the Concept of Multiple Intelligences
- Language Games for all levels)

**2011-2007:** B.A at Pamukkale University, Faculty of Education , Department of English Language Teaching

**2007-2003:** Muğla Turgutreis Anadolu High School

**2003-1995:** Muğla Atatürk Primary School

**PROFESSIONAL EXPERIENCES**

**2013-** English Teacher in Gaziantep Şehitkamil Mahmut Fehime Güleç  
Secondary School

**2011- 2013** English Teacher in Gülemdil İngilizce Kursu

**2010 - 2011** English Teacher in Denizli Doğan Demircioğlu Emsan  
Secondary School (pre service teaching)

**2009-2010:** English Teacher in Denizli Youth Centre