To my family

Variation of Scores in Language Achievement Tests according to Gender, Item Format and Skill Areas

The Graduate School of Education
of
Bilkent University

by

Ayşe Engin

In Partial Fulfillment of the Requirements for the Degree of
Master of Art

in

The Program of
Teaching English as a Foreign Language
Bilkent University
Ankara

June 2012

BİLKENT UNIVERSITY

THE GRADUATE SCHOOL OF EDUCATION

MA THESIS EXAMINATION RESULT FORM

June 1, 2012

The examining committee appointed by the Graduate School of Education
for the thesis examination of the MA TEFL student
Ayşe Engin
has read the thesis of the student.
The committee has decided that the thesis of the student is satisfactory.

Thesis Title: Variation of Scores in Language Achievement Tests according to Gender, Item Format and Skill Areas

Thesis Advisor: Dr. Deniz Ortaçtepe
Bilkent University, MA TEFL Program

Committee Members: Asst. Prof. Dr. Julie Mathews-Aydınlı
Bilkent University, MA TEFL Program

Prof. Dr. Theodore Rodgers
University of Hawaii, Department of Psycholinguistics

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Teaching English as a Foreign Language.

_____

(Dr. Deniz Ortaçtepe)
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Teaching English as a Foreign Language.

_____

(Asst. Prof. Dr. Julie Mathews-Aydınlı)
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Teaching English as a Foreign Language.

_____

(Prof. Dr. Theodore Rodgers)
Examining Committee Member

Approval of the Graduate School of Education

_____

(Visiting Prof. Dr. Margaret Sands)
Director

ABSTRACT

VARIATION OF SCORES IN LANGUAGE ACHIEVEMENT TESTS ACCORDING TO GENDER, ITEM FORMAT AND SKILL AREAS

Ayşe Engin

M.A. Department of Teaching English as a Foreign Language
Supervisor: Dr. Deniz Ortaçtepe

June 2012

Students are assessed to collect information on their language ability or achievement. Some other factors as well as the proficiency level of a student may play a role in their language achievement scores. Gender, item format and skill areas are the factors that may cause variation in the scores, hence affecting the decisions made through these scores. However, there has not been a study that reveals if achievement scores of language learners vary depending on gender, item format or skill areas or the interaction among these factors. This study investigated how language learners scores in language achievement tests vary according to gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, and paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary); and whether the male and females' scores vary according to item format and skill areas. The research was conducted at T.C. Kadir Has University Preparatory School, Istanbul, Turkey. The second achievement test of the second module administered to 303 pre-intermediate level students from different majors was analyzed. The statistical analysis of data revealed that gender does not have a

significant effect on the total scores of the students in language achievement tests. On the other hand, students' total scores vary significantly depending on both the item format and skill areas in the test. In other words, it makes a difference which item format or skill area is used in a test because students' scores change according to the type of the item form and skill areas. Males' and females' mean scores also show differences depending on both item format and skill areas. According to the findings, females outperform males significantly in two item formats; 'find the correct form' and 'paragraph writing' questions, whereas males do not show any superiority in any item format. Also, in skill areas, females outperform males in three skill areas; 'writing,' 'grammar' and 'vocabulary' while males score higher only in one skill area; 'listening.' This study contributed to the existing literature by having studied gender differences. With results both confirming and contradicting the previous research, the present study has a unique place in the language testing literature by looking at the variation of scores according to three variables; gender, item format and skill areas, that have been studied together for the first time, and comparing males' and females' scores in terms of item format and skill areas again for the first time. The wide spectrum adopted while evaluating the differences in the results, and speculations made about these differences can benefit both future researchers in the field in terms of theoretical perspectives, and teachers and administrator in terms of practical perspectives.

Key Words: Language Assessment, Variation of Test Scores, Language Achievement Tests, Gender, Item Format, Skill Areas

ÖZET

DİL BAŞARI SINAVLARINDA PUANLARIN CİNSİYET, SORU TİPİ VE BECERİ ALANLARINA GÖRE FARKLILAŞMASI

Ayşe Engin

Yüksek Lisans, Yabancı Dil olarak İngilizce Öğretimi Bölümü
Tez Yöneticisi: Dr. Deniz Ortaçtepe

Haziran 2012

Öğrenciler dil becerileri ve başarıları ile ilgili bilgi edinmek amacıyla değerlendirilirler. Öğrencilerin dildeki yetkinlikleri dışında bazı faktörler sınav skorlarını etkileyebilir. Cinsiyet, soru tipi ve beceri alanları puanlarda farklılaşmaya neden olabilen faktörlerdir ve dolayısıyla puanlara dayanılarak alınan kararları etkileyebilirler. Ancak dil öğrencilerinin başarı puanlarının cinsiyet, soru tipi ve beceri alanlarına veya bu alanların birbirleriyle olan etkileşimlerine göre farklılık gösterip göstermediğini ortaya koyan bir çalışma daha önce yapılmamıştır. Bu çalışma dil öğrencilerinin başarı sınavlarındaki puanlarının cinsiyet, soru tipi (eşleştirme, boşluk doldurma, doğru formu bulma, çoktan seçmeli, açık uçlu, ve paragraph yazma) ve beceri alanlarına göre (okuma, yazma, dinleme, dilbilgisi ve kelime) nasıl farklılık gösterdiğini ve kız ve erkek öğrencilerin puanlarının soru tipi ve beceri alanlarına göre farkılılık gösterip göstermediğini incelemiştir. Araştırma T.C. Kadir Has Üniversitesi Hazırlık Okulu, İstanbul, Türkiye' de gerçeklerilmiştir. Farklı akademik bölümlerden 303 orta düzey öğrenciye verilen ikinci modülün ikinci başarı sınavı incelenmiştir. Verilerin istatiksel incelemesi cinsiyetin öğrencilerin dil

başarı sınavlarındaki toplam puanları üzerinde önemli bir etkisinin olmadığını ortaya koymuştur. Ancak öğrencilerin toplam puanları soru tipi ve beceri alanlarına gore önemli ölçüde farkılık göstermiştir. Diğer bir deyişle bir sınavda hangi soru tipi ve beceri alanının test edildiği öğrencilerin puanları soru tipi ve beceri alanına göre değişeceginden fark yaratır. Ayrıca erkek ve kız öğrencilerin puanları da soru tipi ve beceri alanına göre farklılık gösterir. Sonuçlara göre, kız öğrenciler 'doğru formu bulma' ve 'paragraf yazma' sorularında erkeklerden önemli bir biçimde daha başarılı olmuşlardır fakat erkek öğrenciler herhangi bir soru tipinde bir üstünlük gösterememişlerdir. Ayrıca beceri alanlarına göre, kız öğrenciler 'yazma,' 'dilbilgisi' ve 'kelime' alanlarında erkeklerden önemli bir oranda daha başarılı olmuşlardır, erkek öğrenciler ise 'dinleme'alanında kız öğrencilerden önemli bir oranda daha başarılı olmuşlardır. Bu çalışma var olan literature cinsiyet farkılıklarını çalışarak katkıda bulunmuştur. Önceki çalışmaları hem destekleyen hem de onlarla çelişen sonuçları ile bu çalışmanın, dil başarı punalarının cinsiyete, soru tipine ve beceri alanlarina göre farklılaşmasını ilk kez inceleyerek ve erkek ve kız öğrencilerin puanlarını soru tipi ve beceri alanlarına göre ilk kez karşıştırarak litratürde özgün bir yeri vardır. Sonuçlardaki farklılıkları değerlendirirken ve bu farklılıklar ile ilgili tahminlerde bulunurken benimsenen geniş bakış açısı gelecekteki araştırmacılara teorik anlamda, öğretmen ve yöneticilere ise pratik anlamda fayda sağlayacaktır.

Anahtar Kelimeler: Dil Değerlendirmesi, Sınav Puanlarının Farklılaşması, Dil Başarı Sınavları, Cinsiyet, Soru Tipi, Beceri Alanları

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor Dr. Deniz Ortaçtepe for her invaluable support, patience and feedback throughout the study. I thank her for her tolerance, positive and friendly attitude. Many special thanks to the jury members Asst. Prof. Julie Mathews Aydınlı and Prof. Dr. Theodore Rodgers.

I would also like to express my gratitude to Prof. Dr. Mustafa Aydın, the rector of T.C. Kadir Has University who made it possible for me to attend the program. Many thanks to Assoc. Prof. Serhat Güvenç, the coordinator of Foreign Languages, for his faith, support and guidance throughout the study. This thesis would not have been possible without his valuable contributions. I wish to thank my colleagues Patricia Marie Sümer, for her guidance, encouragement and mentorship. I am also indebted to my colleagues, Aylin Kayapalı and Gülşah Baysal Sadak who made it possible to conduct my study in the institution. I offer my regards to my colleagues in my home institution and my classmates for their support and motivation.

My greatest thanks to my family, my self-sacrificing mother Mirem Engin, my brother Abdullah Engin, my sister Mine Engin, for their constant support and encouragement. I also would like to thank my uncle Mehmet Paçacı and my aunt Gülten Paçacı, who have never left me alone, supported and believed in me. I wish to thank my precious, little nephew Eren Bartu Engin, having him in my life has always given me power and motivation to go on. Many thanks to my late father, Mehmet Engin. Even if he is not here today, I know he is somewhere proud of me.

Finally, I offer my regards to all of those who supported me in any respect during this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Introduction

Language learners are assessed with a variety of ways with the aim of collecting information on their language ability and/or achievement (Brindley, 2006). The time and type of assessment may change depending on several factors such as the aim of the assessment, the objectives of the course and/or student profile. Among different assessment types, achievement tests are defined as tests which gather information during, or at the end of, a course of study in order to examine if and in which aspects progress has been made in terms of teaching objectives (McNamara, 2000). Some other factors as well as students' proficiency level may play a role in achievement scores such as gender, item format or the skill being tested such as macro skills (e.g., reading, writing, listening, or grammar) or micro skills (e.g. organizing ideas or developing arguments) (Jordan, 1997). Variables such as proficiency level, students' testing strategies, and personal factors such as motivation or anxiety have been researched several times in language teaching (Dörnyei, 2001; MacIntyre, 1995) and testing literature. However, the effects of gender, item format and the skill areas tested seem to be an area which could be recognized more, and analyzed more deeply since they could affect the results of the tests, hence affecting the decisions made through these tests. These decisions include classifying test takers into appropriate proficiency levels, assigning grades, and accepting or rejecting test takers (Shohamy, 2001).

The aim of this study is twofold. First, it attempts to analyze whether language learners' scores in language achievement tests show any difference

according to gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary); second, to reveal whether male and females' scores show any difference according to item format and skill areas tested.

Background of the Study

Information about students' language ability or achievement is obtained through assessment. Depending on what kind of information the test developers want to obtain, different types of tests can be administered to the test takers. There are four main types of tests; proficiency, diagnostic, placement and achievement tests (Hughes, 2003).

Proficiency tests measure how much of a language someone has learned (Davies, 1999). They are designed regardless of any training and the content of a proficiency test is not based on the objectives or syllabus of any course (Hughes, 2003). Diagnostic tests provide information about the students' present situation; their strengths and weaknesses at the beginning of a course (Robinson, 1991), and its distance from target-level performance (Munby, 1978). Placement tests are administered to place the students at the right stage of an instructional program most appropriate to their abilities (Hughes, 2003). The last type of test is the achievement test which is closely associated with the process of instruction (McNamara, 2000, p.5). In most educational settings, achievement tests are designed with reference to specific objectives of a course or curriculum in order to learn how well students have achieved the instructional goals (Brown, 1996). The learning objectives of the syllabus constitute the abilities to be tested in achievement tests (Bachmann, 1990).

The scores received from these tests are used to make decisions about the course, students and instructional materials. Hence, any factors that may cause a variation in the scores should be taken into consideration because they will affect the decisions that may be made depending on the achievement scores (Bachman, 1990) such as classifying test takers into appropriate proficiency levels, assigning grades, and accepting or rejecting test takers (Shohamy, 2001).

Achievement scores of language learners in language tests could be affected by several factors as well as their proficiency level. Factors related to exams such as validity, reliability, practicality (Fulcher & Davidson, 2007; Harris & McCann, 1994; Hughes, 2003) and features of test takers such as attitude, motivation and aptitude (Dörnyei, 2001; Genesee, 1976; Obler, 1989) have been widely discussed, whereas gender as a variable has received little attention in the fields of second language learning and teaching (Catalan, 2003; Nyikos, 2008; Sunderland, 1994). According to Graham (1997), of all the factors that influence test outcomes, gender is the one to which the least attention has been paid. Socially-determined characteristics of males and females may relate to classroom interaction, learning styles and strategies or attitude towards language. Studies examining the effect of gender on learners' achievement have contradictory findings. In the UK, girls perform better than boys in the language part of the general certificate exam to secondary school (GCSE) (Arnot, David & Weiner, 1996), on the other hand, in some countries "girls perform so much better than boys that entrance requirements are lowered for boys applying to English-medium schools" (Byram, 2004, p. 230). Even though there is a common belief that girls perform better than boys at languages, in some mixed-sex schools, boys have

been found to perform better than girls (Cross, 1983). Neurological evidence, while still not clear, suggests that there are potentially relevant differences between male and female brain, yet these differences may be too small to account for gender differences in language achievement (Klann-Delius, 1981). Oxford (1996) argues that social factors such as parental attitude and gender-related cultural beliefs may influence students' success in language. Ryan and Demark (2002) also claim that differences caused by gender may be a reflection of instruction or socialization that varies according to the culture of the setting where teaching takes place.

Gender may also play an important role in students' achievement according to item format. Ryan and Demark (2002) address this issue through two related meta-analytic studies of published and present research. The analysis of students' achievement in language assessments suggests that females outperform males in language assessment if a constructed-response format (e.g., short answer, essay) is employed, but not when their language skills are measured with selected-response items (e.g. multiple choice, true/false, matching). This result reflects gender differences favoring females in writing performance scores. It also implies that, as a result of item format, there might be differences of achievement between males and females in the skill areas as well. Females' success in constructed-response format questions implies better achievement in writing skill compared to males. In Graham's (1997) study with German learners, students were asked about their opinions regarding different aspects of language. According to the results, male students felt less comfortable with reading than their female counterparts, but they felt more comfortable with oral work and general grammar. These differences in

attitude may result in differences of achievement according to the skill area tested; thus, affecting students' success.

While some studies show an advantage for women in language learning (Gu, 2002; Sunderland, 2000), some others report no significant relationship between gender and language learning (Ehrman & Oxford, 1995). Hence, there are some inconsistencies in the literature about the role or effect of gender on language learning, and there is not much information about the effect of gender in assessment results.

Statement of the Problem

Sunderland (1994) claims that even if the effects of gender differences are everywhere, it is ironic that gender appears rarely in writing and thinking on English language teaching: the fact that gender is often neglected as a variable in language learning by writers and language researchers has been pointed out by Nyikos as well (2008). Likewise, even if there are few studies about the effects of item format and skill areas on achievement of learners (Graham, 1997; Ryan & Demark, 2002), the relationship between gender, item format and skill areas has not been studied before. Careful analysis of these factors will be valuable for stakeholders while making decisions or evaluations based on test scores.

Although the studies conducted on gender and language learning mostly report a female dominance in terms of success in language learning, recent research points in a more complex direction, suggesting that males and females might differ in completing specific learning tasks and in different learning contexts (Gu, 1996).

Ignoring any potential differences between male and female scores, or possible relationships between item format, skill areas and gender may result in biased tests advantaging one gender or disadvantaging the other one unintentionally. Gender, in this study, is not only a biologically based term, but it also includes socially constructed roles (e.g. identity, reasoning skills, spatial skills) created by the ways sexes are raised from birth and socialized within a certain culture (Ellis, 1994). Hence, differences in language achievement, if any, caused by gender may reflect social factors which depend on the culture of the setting where teaching takes place. There has not been a study done in the Turkish educational and cultural context that looks at whether language learners' achievement scores vary depending on gender or whether gender interacts with other test features such as item format or target skill. This study attempts to address the following research questions:

1. How do Turkish EFL learners' scores on language achievement tests vary according to
    a. Gender?
    b. Item Format?
    c. Skill Areas?
2. To what extent do male and females' scores vary according to item format?
3. To what extent do male and females' scores vary according to skill areas?

Significance of the Study

Language tests are increasingly understood in terms of their political functions and social consequences (Brown & McNamara, 1998; Shohamy, 2001). Hence, inferences made about individuals based on language tests should be free of

bias and error. Good language testing should care for the rights and interests of particular social groups who may be at risk from biased language assessments (Davies, 1997). Any potential differential and unequal treatment of candidates in language tests based on gender is thus an ethical issue. To avoid such problems, a further insight into specific variables that might affect the achievement of learners and understanding the reasons causing variation of scores is crucial. This study may contribute to the existing literature by providing answers regarding how achievement scores vary according to gender, item format and skill areas, and whether gender has an effect on the students' scores received from different item formats and skill areas. Thus, the findings of this study might help resolve the inconclusiveness in the literature by either strengthening the idea of female dominance in language learning and contributing to growing concern over educational performance of boys (Tyre, 2005; Van Houtte, 2004) or by confirming the literature that emphasizes the variation resulting from individual differences other than gender (Ehrman & Oxford, 1995; Nyikos, 2008).

At the local level, by revealing the variation of scores in language achievement tests according to gender, item format and skill areas, it is expected that the results of the study may help test writers develop tests free of gender bias and predict potential challenges that may be encountered by either group; females or males. The interaction effect of these variables on language achievement scores, if any, will also reveal whether there are any curriculum materials or instructional practices that somehow favor or help either group to develop some skill areas more than the others or to succeed more in a particular question format. Hence, the results

will help unravel any dynamics resulting in differential opportunities for either gender. Sensitivity to the learning preferences or weaknesses of either gender will create a more supportive learning environment for all language learners and help teachers meet learners' needs more fairly.

Conclusion

This chapter aimed to introduce the study through a statement of the problem, research questions, and the significance of the study. Furthermore, the general frame of the literature review was outlined. The next chapter will review the relevant literature. In the third chapter, the methodology including the setting, participants, instruments, data collection methods and procedures will be described. The data collected will be analyzed and reported quantitatively in the fourth chapter. Finally, the fifth chapter will present the discussion of the findings, pedagogical implications, limitations of the study, and suggestions for further research.

CHAPTER II: LITERATURE REVIEW

Introduction

The scores of language learners in language achievement tests could be affected by several factors such as factors related to the exams themselves such as validity, reliability, practicality (Harris & McCann, 1994; Hughes, 2003; Fulcher & Davidson, 2007) and factors related to test takers such as attitude, motivation, aptitude (Dörnyei, 2001; Genesee, 1976; Obler, 1989) Gender, item format and skill areas are among the factors that can also cause variation in test scores; thus, deserve a closer look and analysis. The fact that gender is often neglected as a variable in language learning has been pointed out by Nyikos (2008). Likewise, even if there are a few studies about the effects of item format and skill areas on achievement of learners (Graham, 1997; Ryan & Demark, 2002) the interaction effect of these variables on language achievement scores has not been studied before. This study attempts to analyze whether language learners' scores in language achievement tests show any difference according to gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary), and to reveal whether male and females' scores show any difference according to item format and skill areas tested.

This chapter includes multiple sections. The first section summarizes the literature on the definition of tests in general, and then types of tests, proficiency tests, diagnostic tests, placement tests, and achievement tests which are followed by the sections, and the qualities of tests and different uses of tests. This is followed by

the second section on factors that may affect language test scores, and a particular factor which is gender. This part of the literature review also discusses gender as a construct, sources of gender differences, and the role of gender in English as a Foreign Language learning. The third section provides an insight into item formats, and item formats in language achievement tests; selected –response items, constructed-response items, and personal-response items. Finally, the last section focuses on skill areas in language achievement tests; reading, writing, listening, speaking, grammar, and vocabulary.

Tests

According to Carroll (1968), "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" (p. 46). In other words, a test is a measurement instrument used to draw out a particular sample of an individual's behavior. The inferences and uses made out of language test scores rely on the sample of language use obtained. Language tests can thus provide the means for focusing on the specific language abilities that are of interest (Bachman, 1990). Information about people's language ability is often useful and sometimes necessary. Universities need language test scores to evaluate students from overseas; they cannot accept these students without some information about their proficiency in English; thus, their ability to follow the courses delivered in English. The same is true for organizations hiring employees who are expected to have high language proficiency. Also, within teaching systems, dependable measures of language ability are crucial to be able to make rational educational decisions such as designing

appropriate course materials, setting educational objectives, and passing or failing the learners. Hence, tests serve as a common yardstick to be able make decisions about the test takers.

In a school environment, teachers must periodically evaluate student performance and prepare reports on student progress. Classroom tests play three important roles in the second language program; they are used to define course objectives, they stimulate student attention and progress, and they are also used to evaluate class achievement (Valette, 1977). Tests should provide an opportunity for students to show how well they can handle the target language. Through testing, teachers can determine which targets of the course are presenting difficulties for the learners, and which targets have been acquired. The type and content of tests should be in line with the course content and objectives; thus, tests have an important role in defining course objectives.

Types of Tests

The following section focuses on the types of tests, which are classified according to the type of information they provide. Such a classification may help stakeholders evaluate to what extent the tests they administer are appropriate, and gain insights about testing.

Proficiency tests. Proficiency tests are designed "to measure people's ability in a language, regardless of any training they may have had in that language" (Hughes, 2003, p. 11). Since they evaluate general knowledge or abilities, proficiency tests are not based on a specific syllabus, content or objectives of a

course. The aim is to determine whether the language ability of a test taker corresponds to specific language requirements in order to be considered proficient. Proficiency means having sufficient competency in the language for a specific purpose (Hughes, 2003). A test administered to determine whether a student's English is good enough to study at an American university is an example of such kinds of tests, such as TOEFL or IELTS Academic. Some proficiency tests may be designed taking into account the level and type of English needed to follow a particular course of study. Then, the test may have different forms depending on the subject knowledge needed by the test taker such as a test for arts or for sciences. In other words, such proficiency tests identify the actual ways the test takers will use English in most stages (Heaton, 1990; Hughes, 2003). The Interuniversity Foreign Language Examination (ÜDS) administered in Turkey, which has two forms (medicine and social sciences), is an example of this type of test. There are also some proficiency tests that do not have any occupation or course in mind. The idea of proficiency here is more general in these tests, such as Cambridge First Certificate in English examination (FCE) (Hughes, 2003).

Diagnostic tests. Diagnostic tests are used to detect those areas where learners are strong or weak. (Hughes, 2003). Identifying the strengths and weaknesses of students help teachers ascertain what learning needs to take place. Good diagnostic tests are useful for individualized instruction and self-instruction because they provide detailed analysis of a student's command of particular linguistic skills. One important feature of diagnostic tests is that "they are administered at the beginning or middle of a course, not at the end" (Brown, 1996, p.

15), it is also noteworthy to mention that diagnostic tests may be prepared according to the syllabuses of specific classes (Bailey, 1998). The preparation of a comprehensive diagnostic test of English is a hard work, and the size of such a test would make it impractical to apply regularly (Hughes, 2003). For this reason, few tests are designed for only diagnostic aims. Achievement or proficiency tests are often utilized for diagnostic purposes (Heaton, 1990).

Placement tests. Placement tests provide information that will help assign the students to the appropriate stage of the teaching program according to their abilities (Hughes, 2003). Typically, they are used to assign students to different levels, and thanks to placement tests, there are different groups of students consisting of similar language ability students at the beginning of a course (Brown, 2004). Placement tests can be purchased or produced in house. If they are purchased, the institution should be sure that the test will suit its particular teaching program. Placement tests are more successful when they are produced for particular situations because then, they can recognize the key features required at different levels of teaching in that institution. Effective placement tests are built on the features of the teaching context (e.g., the language level of the students, the methodology and the syllabus type (Bailey, 1998; Brown, 1996). Hence, a placement test which asks grammar questions is not appropriate for a course where a skill-based syllabus will be exploited. Brown (1996, p. 13) points out that if there is an inconsistency between the placement test and the syllabus, the danger is that the groupings of similar ability students will simply not occur indicating that placement test has not served its purposes. A good placement test should sort students into groups which are made up of students with rather

similar levels. As a result, teachers can give their full attention to the problems and learning points appropriate for that level of students (Brown, 1996).

Achievement tests. Achievement tests determine the success of individual students, groups of students, or the courses themselves in attaining objectives (Hughes, 2003). In other words, they are designed with a particular reference to objectives of a course or language program to measure learners' mastery of these objectives. Achievement tests have several functions in teaching programs.

As the definition also indicates, achievement tests are used to accumulate evidence for how much the learners have learned the content of a course and how successful they have been in achieving the objectives of that program (Brown, 1996); thus, also helping teachers evaluate the effectiveness of their teaching and methodology. As Spolsky (1995) points out, achievement tests help teachers continually check on their learners' progress to determine whether learning has been successful. By making use of the results, teachers may make decisions regarding appropriate changes in teaching procedures and learning activities (Bachman, 1990). Learners are provided with periodic feedback on their progress in language learning. Johnston (2003) states that learners need to have a sense of "how well they are doing: of their progress, of how their work measures up to expectations" (p. 77), and achievement tests can enable learners to monitor their weakness in the language as well as their overall strengths on a regular basis.

One function of achievement tests is to provide feedback on the effectiveness of teaching and the language program itself (Bachman, 1990; Bachman & Palmer, 1996; Bailey, 1998). They can be used to make adaptations in the language program,

as well as to evaluate those adaptations (Brown, 1996). Achievement tests may also help lead the curriculum developers and syllabus designers to make adaptations to increase the quality of language program offered (Brown, 1996). These adaptations improve the curriculum with appropriate changes so as to better suit the language needs of learners.

Qualities of Tests

Test results are used to make important decisions about the test takers such as classifying them into appropriate proficiency levels, assigning grades and accepting or rejecting test takers (Shohamy, 2001); thus, stakeholders must make sure that the tests administered possess good qualities. Bachman and Palmer (1996) suggest that quality of tests can be evaluated "on the basis of a model of test usefulness" (p. 17). The test usefulness model is concerned with six qualities: validity, reliability, authenticity, practicality, interactiveness and washback. Validity in general is defined "the degree to which a test measures what it claims to be measuring" (Brown, 1996, p. 231). If a test assesses what it should assess, it can be considered as valid. Reliability, on the other hand, refers to the consistency of test takers' scores (Bachman & Palmer, 1996). More specifically, a reliable test should provide similar results if it is given to two different groups with the same proficiency level or if it is given to the same group for the second time. Another good quality of a test; authenticity, is the degree to which test tasks are relevant to real life language use (Bachman, 1990). If a test and its tasks are closely related to the features of real-life language use, it is considered to be authentic. Practicality is also a quality of good tests which is defined as "the relationship between the resources that will be required

in the test design, development, and use of the test and the resources that will be available for these activities" (Bachman & Palmer, 1996, p. 36). If the tests in a course have practicality, it means they are easy and inexpensive to construct and administer. A further quality of good tests is interactiveness. Purpura (1995) considers interactiveness as the degree to which a test serves to engage a test taker's language ability, or the degree to which task elicits test performance which replicates a genuine interaction. In other words, interactiveness measures the extent and type of involvement of the test takers' individual characteristics in accomplishing a test task. The last quality of good tests is washback also known as backwash. The term 'washback' refers to the effects of testing on teaching and learning (Hughes, 2003).Washback is generally considered as being either positive (if a test promotes learning and teaching) or negative (if the test hinders learning and teaching).

Uses of Tests

The fundamental use of testing in an educational program is "to provide information for making decisions, that is, for evaluation" (Bachman, 1990, p. 54). This evaluation can be done regarding the students, the teachers, the course, or the institution itself. Information about educational outcomes is essential for effective formal education. A prime source for such kind of information is the test results. In order to be able to depend on test results while making decisions, accountability and feedback should be considered essential ingredients for the continued effectiveness of any educational program. Bachman and Savignon (1986) describe accountability as "being able to demonstrate the extent to which we have effectively and efficiently discharged responsibility" (p. 380). Feedback, on the other hand, simply refers to

information that is provided to teachers, students, and other interested persons about the results or effects of the educational program. Test results can be used to make decisions about the programs and courses to improve learning and teaching through appropriate changes. Without the opportunities to improve student performance and program effectiveness, there is no reason to test, since there are no decisions to be made, and therefore no information required. In educational programs the decisions made about students and teachers have some effects on their lives. The first decision that may be made about students is whether or not they should be accepted to a program. Learners are also assigned to levels, and in which class and with whom they will study is determined by the test results. Test results are also used to decide whether a person is eligible to be hired. For example, if teachers are not native speakers of the target language, institutions ask for information about their language proficiency by means of a proficiency test. It is therefore essential that the information upon which we base these decisions be as reliable and as valid as possible. Another use of test results is to provide information to evaluate a course, a language program or a teacher. Performance of students on achievement tests can indicate "the extent to which the expected objectives of the program are being attained, and thus pinpoint areas of deficiency" (Bachman, 1990, p. 62).

Factors Affecting Language Test Scores

In order to obtain reliable test scores, the abilities test developers want to measure should be differentiated from the other factors that might affect the test-takers' scores. Bachman (1990) groups those factors that affect test scores into three categories: "(1) test method facets, (2) attributes of test takers and (3) unpredictable

random factors" (p. 164). Test method facets are about the features of the exam, and they are systematic because they are the same in all test administrations. If there are matching questions in the test, it does not matter whether it is given in the morning or the evening. Attributes of individuals that are not related to language ability include learning styles, knowledge of the content, or group characteristics such as gender, ethnic background and race (Bachman, 1990). These attributes are related to who the students are, so they are systematic in a way because they will affect the scores regularly. Test scores are not only affected by systematic factors, there could be random, unsystematic factors that could affect the test results such as emotional state of the test taker, features of the environment like heating or noise, and the test administrators attitudes (Bachman, 1990). The results of the effects of these factors may vary because they are not equal every time the examinee takes a test. Different factors will affect different individuals in different ways. The following section will focus on three factors; gender, item format and skill areas which have not received enough attention in the testing literature despite the fact that they may affect test scores.

Gender

Gender as a broad term is often used to denote not only biologically based, dichotomous variable of *sex* (that is, male or female) but also the socially constructed roles (i.e. gender) created by "the ways sexes are raised and socialized within a certain culture" (Nyikos, 2008, p. 73). Hence, according to Nyikos (2008), it could be concluded that nature and nurture create the totality of what is classified as male and female. Individuals learn the characteristics and opportunities associated with

being male and female through socialization processes, in other words, these characteristics and opportunities can be considered context/ time-specific and changeable.

Sources of gender differences. There are biological and environmental hypotheses on performance differences between males and females. Three biological features are considered to be at work; genetic, hormonal and brain differences. (Halpern, 1992). It is difficult to distinguish these features because they are not separate, but rather interrelated. Genetic differences hypothesis accounts for performance differences by proposing the theory that males and females have different intellectual abilities because they inherit different genetic codes. With different genetic features, different performances are inevitable. Legato (2005) claims that women have more nerve cells in the left part of the brain where language is centered. There have also been plenty of studies seeking to determine the effect of hormones on the development of cognitive abilities. One theory links "early physical maturation with intellectual development in order to explain girls' assumed superiority in early language related skills" (Gipps & Murphy 1994, p. 58); on the other hand, another theory proposes that late maturers at puberty (typically boys) exhibit "more highly developed spatial skills than verbal skills, whereas for early maturers (typically females) the converse is true" (Gipps & Murphy 1994, p. 58). Another biological theory regarding gender differences is brain differences. Based on five reviews, Halpern (1992) proposes that males and females differ in brain organization for intellectual behaviors. The female brain is systematically more

organized than the male brain, which implies the female brain is less lateralized and language functions are represented in both hemispheres.

There are also interesting environmental theories regarding gender differences in linguistic performance. Wilder and Powell (1989) talk about the different ways boys and girls are encouraged to interact with the environment and the people around them. Gipps and Murphy (1994) propose that there are expectations that girls perform better in language domains than quantitative domains, and children's judgments closely reflect those of their teachers and parents. Different approaches to different sub-groups (i.e. males & females) may encourage different skill development; in particular, boys are encouraged to develop independent, self-confident behaviors which are required more for future achievement in mathematics and science (Gipps & Murphy, 1994). On the other hand, Nyikos (2008) argues that adults have a subconscious perception of females' language superiority, and talk more to baby girls than boys, respond more to girls' early attempts to talk and have longer, more complex conversations with daughters. One environmental hypothesis suggests that students perform better when there is a close correspondence between their self-image and gender stereotyping of the task (Nyikos, 2008). Wilder and Powell (1989) mention the item content is also a source of differential performance because content reflects different life experiences of males and females. Perceptions of students regarding the value of certain contents and subject areas also connected their performance. Boys are inclined to see mathematics and science as more valuable for their future; however, both boys and girls think they are not as important

for girls' future. Such perceptions affect motivation; thus, influence their engagement with certain subjects (Gipps & Murphy, 1994).

Gender in EFL context. Sunderland (2000) points out that a wide range of language phenomena, such as language tests, language performance, styles and strategies have been shown to be gendered because females and males tend to behave differently. Therefore, it is inevitable to expect a gender effect on language learning.

There are studies with confirming results that gender causes differential performance, as well as other studies, which found no differences between males and females in foreign language skills. Feyten (1991) did not find any differences in general language learning skills of male and female foreign language learners. Likewise, Bacon (1993) and Markham (1988) looked at listening comprehension abilities of foreign language learners and could not identify any gender based differences. Nyikos (1990) also looked at gender effect on foreign language learning. She found no difference in males' and females' rote memorization skills. However, some studies did reveal significant differences between males and females regarding the factors related to the language itself. According to the results of Catalan's (2003) study, females use a higher number of vocabulary learning strategies than males. She also looked at the difference between male and female students in terms of the range of vocabulary learning strategies, and the results indicate very small differences between the genders regarding the ten most and least frequently used vocabulary strategies; however, there are differences in other strategies in the middle in terms of frequency of use. For example, analysis of the part of speech for a new word is reported as a more preferred strategy by females, and more males report that they

analyze affixes and roots. Another study on strategies was conducted by Politzer (1983) who studied language learning behavior and social behavior and found out that social strategies are more used by females. In Politzer's (1983) study, females expressed more interest in interpersonal relationships ; for example, cooperativeness and less interest in competitiveness and aggression. Gu's (2002) revealed that female participants outperformed male participants on both vocabulary size and general English proficiency. Oxford and Nyikos (1989) found that formal rule-based strategies, general study strategies, and conversational input- elicitational strategies were used more often by females than males. Another study conducted by Boyle (1987) found that female Chinese learners were stronger in overall language ability, on the other hand, male learners of English in China were found to be stronger in terms of vocabulary recognition in a listening task. Farhady (1982) reported a study that revealed that females were better at recognizing the constituents of more or less prestigious dialects; thus, females were able to differentiate among dialects better than males. Bensoussan and Zeidner (1989) found that males reacted less negatively and experienced less anxiety than women toward oral language tests.

As indicated above there are studies with contrasting results regarding gender differences. While several studies indicate a female superiority in language achievement, there are also many studies which came up with no significant differences between males and females. Analysis of gender differences in different contexts may put forward different results.

Another factor that may affect test scores is item format. Questions asked in different formats may result in different achievement scores. The following section will focus on test items and item formats.

Test Items

Brown and Hudson (2002) describe test item as "a unit of measurement with a prompt and a prescriptive form for responding, which is intended to yield a response from an examinee from which a performance in some language construct may be inferred in order to make decisions" (p. 57). In other words, test items are used to obtain samples of behaviors from which decisions and inferences can be made about the test taker. A language test item should be quantifiable either objectively or subjectively in order to serve as a unit of measurement. Since the test item involves a prompt, the portion of a test item to which examinees must respond, and a prescriptive form for responding, the examinee responds in a way prescribed by the item, and s/he is directed to write an essay, perform a task, select an answer or respond the task in some other way (Brown & Hudson, 2002). The performance of the examinee is evaluated in order to make inferences in terms of performance in some language construct. Language construct may refer to a language skill, success in an instructional objective, pragmatic competence or any other language performance.

Brown and Hudson (2002) propose general rules to help write good tests; four rules of Grice (1975); Grice's Maxims for Cooperative Principle of Discourse, can also be used to cover test writing. These four maxims are, maxim of quantity; being as informative as required not more or less, maxim of quality; being truthful,

maxim of relation; being relevant, and finally maxim of manner; being orderly and avoiding obscurity. According to these rules, test writers are advised to write relevant, unambiguous items by providing information not more than required. They are also advised to be orderly in test preparation. Test writer may prefer using different test item formats depending on the objectives of the course or the language construct being tested. The following subsections will provide an overview of the different test item formats.

Selected response items. Popham (1978) refers to *selected-response items* as those which involve simply selecting the correct answer from several alternatives. The test taker does not need to produce any language; thus, these items are more preferred to test receptive skills; reading and listening. Administering and scoring selected-response items are relatively easy, and the scoring is objective. However, writing selected-response items takes a lot of work on the part of the test writer. Since there is no language production from the students, guessing should be limited as a factor in the test takers' scores, and correct answers should be randomly dispersed in order to avoid a pattern. Within selected-response items, the most common item formats are binary choice, matching and multiple choice questions (Brown & Hudson, 2002). Binary choice questions require the examinees to choose from one of two choices, for instance, between true and false. While this format provides simple and direct indices of whether a particular point has been comprehended, there is a high chance of guessing and test writers may be inclined to write deceptive items to make the items work well. In matching questions, examinees match words/phrases in one list with the ones in another list. While the guessing

factor is low, matching questions are limited to measure whether the test taker can associate one set of facts with another. Another selected-response item format is multiple choice questions. They are good for testing a variety of learning points, yet it is challenging to write quality distracters, and there is still a guessing factor.

Constructed response items. Popham (1978) refers to *constructed-response items* as those which involve the production of a language sample in response to the input material. Such language production may be highly structured; tests that elicit single sentence or phrasal responses such as the Ilyin Oral Interview (Ilyin, 1972). In some tests, on the other hand, the response is fairly unstructured such as the ILR Oral Interview (Lowe, 1982). Research supports the hypothesis that constructed response types are generally more difficult than selected response types (Shohamy, 1984). Constructed-response items eliminate most of the guessing, but pose challenges for the raters. Constructed-response items require subjective scoring, and their scoring is time consuming. Since there is language production, this format is appropriate for productive skills; speaking and writing. The advantage of constructed-response item format is that it allows for testing the interaction of receptive and productive skills; like interaction of listening and speaking in an oral interview (Brown & Hudson, 2002). Three types of constructed-response item format are common in language teaching; fill-in, short answer and performance items. In fill-in format, the examinee is provided with a context, but a part of the context is removed and the examinee fills in the gap. These items are easy to construct and administer, but are limited to the length of the blank which is a short phrase or a word, and there could be more than one possible answer. In short answer item format, the test taker responds to the prompt with one or more phrases, or sentences. While they are easy to create, and

take a short time to administer, each examinee can come up with a unique answer which makes the scoring more challenging and subjective. The last type of constructed-response item format is performance items which require examinees to perform a task using the spoken or written language. Most common performance question in writing is a paragraph or essay writing question. While such questions may stimulate authentic language, they are difficult to create and relatively more time to score.

Personal response items. Personal response items ask for students to produce language, but they permit the responses and even the ways the tasks are completed to be quite different for each student, in other words, this is personal assessment because students communicate what they want (Brown & Hudson, 2002). Personal response items are directly related to and integrated into the curriculum. They are also suitable for evaluating learning process. On the other hand, they require subjective scoring because the grader evaluates the personal work and there is no one correct answer, and they are also hard to create and structure. Conferences, portfolios and self-assessments are considered to be personal response items. Conferences require the student to visit the teacher's office and discuss a particular piece of work. While these help students understand the learning process and develop better self images, it is extremely time consuming for the teacher, and it is hard to use the conference meetings for grading purposes. A popular way of personal response items is portfolios. Portfolios are "collections of work designed for a specific objective that is, to provide a record of accomplishments" (NLII, 2004). Portfolios develop student self-reflection, critical thinking, and responsibility for

learning, but pose decision and interpretation problems. In self-assessment, students rate themselves through performance, comprehension or observation self-assessments (Yamashita, 1996). These involve students in the assessment process and encourage autonomy, but they are prone to subjective errors (Brown & Hudson, 2002).

Item formats are among the factors that may affect test scores because each format may appeal to students with different personalities and language learning styles. A student who is good at writing and expressing feelings in a constructed-response item format may find it hard to answer selected-response questions. Furthermore, a student who is used to formal testing format and being given strict guidelines may feel uncomfortable when assessed on freer, personal response items such as portfolios. Other than gender and item formats, students may show differing success rates depending on the skill being tested. While some students find certain skills easier, they may struggle with others. The following section will focus on the nature of language skills and the ways of testing these skills.

Skill Areas

Language learners may not succeed equally in all language skill areas (reading, writing, listening, speaking, grammar and vocabulary). While a student is good at reading, s/he may not be as successful in grammar because different skill areas require different strategies. Hence, the nature of the skill areas can also be considered a factor that may affect learners' success, in this case, their test scores.

Reading. Reading is a complex activity that involves both perception and thought. It involves recognizing the words which refers to the process of perceiving how written symbols are parallel to spoken language, and comprehension; process of understanding utterances (Pang, Muaka, Bernhardt, & Kamil, 2003). The goal in reading is direct comprehension without recourse to the native language (Valette, 1977). To this end, readers need to employ their existing knowledge of the topic, vocabulary, grammatical knowledge, and other strategies to help them comprehend the written text (Bernhardt et al., 2003). Hughes (2003) classifies these other strategies are into two; macro skills of reading such as scanning, skimming, identifying an argument, identifying examples, and micro skills of reading such as understanding relations, guessing meaning, identifying referents. Reading is a language skill that is also essential to the development of other skills. Learners can learn new vocabulary, grammar topics, and sentence structures by reading in English. Reading texts are also models for students' writings.

Testing reading is a challenging task in that receptive skills may not present themselves directly in overt behavior. The important job of the test writer is to set tasks which will not only cause the candidate to exercise reading, but will also result in behavior that manifests successful use of reading skills (Hughes, 2003). Reading skills are also referred as operations. Depending on their purpose, readers employ different operations which can be classified under two main headings; there are expeditious operations which require speed such as skimming the text for main ideas, or scanning the text to find specific information; the second type involves careful reading operations which require more in-depth analysis and comprehension of the

text for the purposes such as identifying reference, making inferences or outlining logical organization of texts (Hughes, 2003). What kind of operations the test writer wants to test determines the item formats and the nature of the exam texts. Choice of text can be specified with a number of parameters such as type, form, graphic features, topic, style, length, readability, range of vocabulary and structures and so on (Hughes, 2003). After choosing the text, the test writer should decide what a competent reader should and can derive from the text, and write tasks which can be carried out in a number of ways; reading aloud, written response, multiple-choice, picture-cued items, matching tasks, editing tasks, gap-filling tasks, cloze-tasks and so on (Brown, 2004). Asking for colleagues' recommendations and moderation of the test should be the final step while developing the test.

Writing. Writing is a method of expressing language in a written form. Of all the language skills, writing is considered the most sophisticated (Vallette, 1977). It requires real proficiency on the part of the writers, and involves the development and presentation of thoughts in a structured way. The genre, the addressee, the topics determine the way writers produce texts. Martin (1984) describes genre as "a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture" (p.25). He gives examples of genres from different skills of language such as poems, narratives, expositions, lectures, seminars. Learners are expected to use a language in line with the genre such as an informal language in a letter to a friend, and a formal language in an academic essay. The type of writing teachers teach depend on the objectives of the course, students' age, interests and levels. As in reading there are some subskills in writing too so as to produce effective texts such

as writing grammatical sentences, using correct words with correct forms, paraphrasing, developing an argument in a coherent way, supporting the main idea with details. Process and product approaches have dominated much of the teaching of writing in the last 20 years with genre approaches gaining adherents in the last ten years (Badger & White, 2000). A process approach involves writing multiple drafts and editing, in a product approach students imitate a model, and in a genre approach they are asked to follow predetermined genre conventions.

There are many kinds of writing tests because of a wide variety of writing tasks learners need to engage in (Madsen, 1983). Since writing is a productive skill, assuming that the best way to test writing is to make the language learner write is a reasonable assumption. However, to state the testing problem in a writing task is not an easy job. Hughes (2003) recommends some steps to be followed to develop a good writing test, these steps are specifying all possible content, including representative samples of the specified content, setting as many tasks as feasible, testing only writing ability and nothing else, restricting candidates, and setting tasks which can be reliably scored (p. 83). According to Brown (2004), there are four categories of written performance to be tested depending on the range of written production. Imitative writing requires learners to attain fundamental skills; writing letters, words, punctuation and very brief sentences. In this stage form is more important than context and meaning. This category can be tested with tasks such as copying words, listening to cloze selection tasks, form completion, converting numbers and abbreviations to words (Brown, 2004). The next stage comprises intensive tasks which require learners to produce appropriate vocabulary in a context

in length of a sentence. To achieve these tasks learners are asked to transform sentences, describe pictures, order words and complete sentences. Intensive writing is followed by responsive writing in which learners are expected to perform at a discourse level, and to connect sentences into paragraph. The last stage is extensive writing; "successful management of all the processes and strategies of writing for all purposes, up to a length of an essay" (Brown, 2004, p. 220). Intensive and extensive writing can be tested with tasks such as writing reports, narrating, responding to a text, writing opinions, interpreting graphs and so on. After setting the writing task, the next step is to score the writings. There are two basic approaches to scoring; analytic and holistic scoring. Holistic scores involves assigning a single total score to a piece of writing on the basis of an overall impression of it, and analytic scoring involves assigning a separate score for different aspects of writing and adding those scores up (Hughes, 2003).

Listening. Listening is the process in which spoken language changes into meaning in the mind. Valette (1977) proposes that listening requires proficiency in three areas: "discrimination of sounds, understanding of specific elements, and overall comprehension" (p.140). Language learners need to be familiarized with sound system of the target language and should be trained to make the necessary sound distinctions to understand the message. Just like the other skill areas, there are some macro skills in listening necessary for comprehension such as obtaining the gist, listening for specific information, following directions (Hughes, 2003). Listening requires learner engagement; thus, the type of text learners are exposed to is crucial for their engagement. Basic principles of teaching listening (Harmer, 2007)

are summarized as: “the tape recorder is just as important as the tape, preparation is vital, once will not be enough, students should be encouraged to respond to the content of a listening, different listening stages demand different listening tasks, and good teachers exploit listening texts to the full” (pp. 99-100).

Listening skill can be incorporated into two broad categories of tests, one that utilizes listening to evaluate something else such as vocabulary or speaking, and one that uses listening to assess proficiency in the listening skill itself (Madsen, 1983). Depending on their purpose, listeners employ different operations. They can execute macro skills (i.e. global operations) and depend on overall grasp of the text for purposes such as obtaining the gist, following an argument, recognizing attitudes. As an alternative, they can also execute micro skills and attend to smaller bits and chunks of language for purposes such as discriminating among sounds, recognizing reduced forms, distinguishing word boundaries (Brown, 2004, Madsen, 1983; Richards, 1983). Texts to be used in listening tests should be specified in terms of type; monologue, dialogue, conversation, announcement, etc. form; description, narration, argumentation, etc. and length; expressed in seconds or minutes; speed of speech; expressed as words per minute (Hughes, 2003). After specifying the operations and selecting the text, questions are prepared. Possible techniques to ask listening questions are multiple choice, gap filling, short answer, information transfer, note-taking, and transcription (Hughes, 2003). The moderation of the test items is essential, which could be done by piloting the test with colleagues, and analyzing the items and reactions to the items (Hughes, 2003).

Speaking. Speaking is the productive skill in the oral mode, and it is more than just pronouncing words. It is more about making oneself understood. Improving your English speaking skills will help you communicate more easily and effectively. There are seven major principles of teaching speaking (Brown, 2001); "using techniques that cover the spectrum of learner needs, providing intrinsically motivating techniques, encouraging the use of authentic language in meaningful contexts, providing appropriate feedback and correction, capitalizing on the natural link between speaking and listening, giving students opportunities to initiate oral communication, and encouraging the development of speaking strategies" (pp. 275-276). Teaching speaking provide students with opportunities for developing oral fluency through interpersonal interactions, and the ability to speak coherently and intelligibly on a topic is an essential goal for ESL students.

Speaking has an integrative nature; vocabulary, grammar, pronunciation play a role in the overall speaking skill of a language learner; thus, while assessing speaking a number of factors should be taken into account such as command of linguistic features (grammar,vocabulary, pronunciation), the interlocutors' familiarity with the accent of the speaker and also the topic itself , the speakers' anxiety, and approval of the topic by the audience (Dalkılıç, 2001; Khamkhien, 2010; Kitao & Kitao,1996). Test developers need to follow a number of criteria. First, they should include a representative sample of the specified content when setting tasks, choose appropriate techniques (e.g. interview, interaction, response to recordings), plan and structure the testing carefully, and ensure valid and reliable scoring (Hughes, 2003). Depending on the purpose of the task in the test, test takers employ different

operations; they can either execute informational skills for purposes such as providing personal information, presenting arguments, expressing opinions, or execute interactional skills for purposes such as expressing agreement or disagreement, eliciting information, changing the topic etc. (Hughes, 2003). Similar to writing, speaking skill can also be graded according to analytic or holistic scales.

Grammar. Grammar is defined as the structural foundation of our ability to express ourselves (Crystal, 2004). Crystal (2004) claims that the meaning and effectiveness of the way people use language can be monitored more closely when if there is an awareness of how grammar works. According to him, grammar can help improve precision, identify ambiguity, and make use of the richness of expression available in English, and it can help everyone not just teachers or learners of English. Learners want to know why and how a foreign language works, and answers are found in its grammar. Grammar is a conscious learning and it is an essential part of awareness raising.

Testing grammar covers a wide range of topics from simple inflections to complex syntax (Madsen, 1983). Lack of grammatical ability sets limits to learners' language achievement; thus, testing grammar is essential to detect language abilities of learners. The type of test to be given determines the specifications of the test; thus, affecting the questions and the content. For example, for achievement tests, either objectives or the syllabus list grammatical structures to be taught, so test developers follow either of them to write specifications and relevant questions (Hughes, 2003). There are four widely used techniques to test grammar; gap filling, paraphrasing, completion and multiple choice, and whatever techniques are chosen, it is important

for the text of the item to be written grammatically correct and in natural language (Hughes, 2003).

Vocabulary. Wilkins sums up the importance of vocabulary by proposing that nothing can be conveyed without vocabulary (as cited in Thornbury, 2002). Learning vocabulary is a challenge for learners because it is a never ending task with multiple facets. Richards (1976) and Nation (2001) (as cited in McCarten, 2007) list the different things learners need to know about a word. These include: the meaning(s) of the word, its spoken and written forms, what word parts it has (e.g., any prefix, suffix, and "root" form), its grammatical behavior (e.g., its word class, typical grammatical, patterns it occurs in), its collocations, its register, what associations it has (e.g., words that are similar or opposite in meaning), what connotations it has, and its frequency. Vocabulary is the most critical component of learning English because no other skill can develop without it. There is plenty of research on how learners learn vocabulary best and how teachers might best teach. Some key principles can be taught to help students learn vocabulary more effectively such as giving vocabulary an important place in the syllabus and the classroom so that students can see its significance and see that learning a language does not only include learning grammar (O'Dell, 1997). A variety of material, repeating and recycling in class, providing opportunities to organize vocabulary, and making vocabulary learning persona can be offered to learners.

Testing vocabulary is a bit different from testing other skills with regard to its specifications. If vocabulary is explicitly taught, all the items could be included in the specifications. Grouping the words according to whether production or

recognition is necessary is also a part of writing specifications, the subsequent step is to group them again according to importance (Hughes, 2003). To test recognition ability, recognizing synonyms, definitions, and appropriate word for context; to test production ability, gap filling, picture naming, rewriting techniques could be used (Hughes, 2003).

Conclusion

In this chapter, factors that may affect the language test scores; gender, item format and skill areas have been reviewed and relevant literature has been summarized, the sections of this part provide information about types of tests; proficiency tests, diagnostic tests, placement tests, and achievement tests, qualities of tests and uses of tests. Next, factors that may influence language test scores, gender; sources of gender differences, and gender in EFL context were discussed, item formats in language achievement tests; selected –response items, constructed-response items, and personal-response items, and skill areas in language achievement tests; reading, writing, listening, speaking, grammar, and vocabulary were reported in the light of the relevant literature.

The next chapter will provide information about the methodology of the study focusing on setting, data collection, samples, and data analysis.

CHAPTER III: METHODOLOGY

Introduction

This study attempts to analyze whether language learners' scores in language achievement tests show any difference according to gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary), and to reveal whether male and females' scores show any difference according to item format and skill areas tested. In this respect, the present study addresses the following research questions:

1. How do Turkish EFL learners' scores on language achievement tests vary according to
    a. Gender?
    b. Item Format?
    c. Skills Areas?
2. To what extent do male and females' scores vary according to item format?
3. To what extent do male and females' scores vary according to skill areas?

The present chapter will cover the setting, sample, data collection, and data analysis.

Setting and Samples

The research was conducted at T.C Kadir Has University Preparatory School, a private university situated in Istanbul, Turkey. As for the choice of the institution,

eligibility and convenience were of primary concern. The school is in charge of providing compulsory English language education for students who have passed the university exam before they start their bachelor's education in their departments. The program lasts for one year and consists of five proficiency levels: zero beginners, false beginners, elementary, pre-intermediate and intermediate. Students are allocated into groups based on the scores they receive on the placement test administered at the beginning of the year. The students take the proficiency exam at the end of the year and start their majors, the ones who fail take the proficiency exam in September again, and if they fail, they repeat the preparatory year. The school is applying a modular system. There are five modules of seven weeks. Beginners and elementary level students attend the classes for five modules, and pre-intermediate and intermediate level students attend the classes for four modules. If a student fails, he repeats the same level, so the same module. There are two short quizzes and two achievement tests during a module. The students take an end of module exam at the end of each module to move on to the next level. Beginner level students have main course and writing lessons, elementary level students have main course, reading, writing, and listening lessons, pre-intermediate and intermediate level students have main course, integrated skills and extra writing lessons. Each lesson is taught by a different instructor, and the same lesson is sometimes shared by more than one instructor such as beginner and elementary level main course lessons.

The sample for this study comprised pre-intermediate level students in the second module. It was purposeful sampling: pre-intermediate was chosen because it is the level with the highest number of students, and it was assumed that the higher

number of participants would provide more reliable results. There are 303 students in total with 163 males and 140 females. The students are from different majors with different educational, social and economic backgrounds. They are young adults around 18 years old. They have 26 hours of English instruction a week. Their main course book is Language Leader Pre-Intermediate, they also have an academic skills book; Academic Encounters, and a separate writing book; Introduction to Academic Writing. Each book is taught by a different instructor.

Data Collection

Data Source: The Achievement Test

The second achievement test of the second module (see Appendix 1) administered to pre-intermediate level students was analyzed. The test was written by one of the members of the testing office, and it was administered for the first time. The test designer stated that the question types and the content of the test were determined according to the textbook, and only the question types in the textbook were included in the test. The test included 5 sections; reading, writing, listening, grammar and vocabulary. The question formats are as follows:

Table 1

Skill Areas and Distribution of Item Formats (see Appendix 1)

Skill Area	**Selected-Response Questions**	**Constructed-Response Questions**
Reading	matching finding the definition	open-ended
Writing	none	write a paragraph
Listening	multiple choice fill in the blanks	none
Grammar	fill in the blanks write the correct form	none
Vocabulary	matching	none

The tests to be analyzed were taken from the testing office of KHU with the permission of the director. There were fourteen pre-intermediate classes in total. All parts of the tests belonging to students in class pre-intermediate 1 were photocopied as a source, and grading sheets belonging to all pre-intermediate classes were taken. The researcher did not need to photocopy all the tests because detailed grading sheets were filled out by the teachers who marked the exam. In the grading sheets, all questions were allocated a slot and the teachers filled in the slots with grades depending on the answers of the students. All the exams were graded by two teachers and discrepancies were checked after both graders marked the tests. The final grades were given by both teachers after discussing the discrepancies.

Data Analysis

The data analysis was done in several steps. The first step was to enter the data into SPSS; a software for running statistical tests for the social sciences and categorize the variables as in female/male, item format and skills. The items (i.e. questions) were classified according to their format; constructed response questions such as essay questions, short answer questions, or selected-response items such as matching questions, multiple choice, and the skill areas tested were identified.

In order to answer the research question 1a and to determine the variation, if any, in the scores according to gender, the researcher divided the scores into two; males' and females' scores. First, the researcher looked for the effect of gender on total achievement scores to see whether there was any difference between males' and females' scores or dominance of either gender in terms of success by using an independent samples T-test. In order to answer the research question 1b, the researcher entered the test scores according to the item formats included in the test; matching, fill in the blanks, find the correct form, multiple choice, open ended, and paragraph writing questions. Scores in these different item formats were compared by using one-way ANOVA test to identify whether the students performed better at a particular item format. Likewise, in order to answer the research question 1c, the researcher entered the test scores according to the skill areas tested in the exam; reading, writing, listening, grammar and vocabulary. Scores in these different skill areas were compared by using one-way ANOVA test to identify whether the students performed better at a particular skill area.

In order to answer the research questions 2 and 3, the researcher compared male and females' scores in terms of item format and skill areas and looked at whether either gender was better at a particular question format or a particular language skill. To identify whether male and females' scores show any difference according to item format and skill areas tested, the researcher entered the data into SPSS and ran two-way ANOVA tests.

Conclusion

In this chapter, the methodology used to carry out the study was described in terms of its setting and samples, data collection, and data analysis. In the next chapter, the details of the data analysis as well as the results revealed will be discussed in detail.

CHAPTER IV: DATA ANALYSIS

Introduction

This study investigates the variation of scores in language achievement tests. It aims to analyze how language learners scores in language achievement tests vary according to gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, and paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary); and whether the male and females score vary according to item format and skill areas. In this respect, the study addresses the following research questions:

1. How do Turkish EFL learners' scores on language achievement tests vary according to
   a. Gender?
   b. Item Format?
   c. Skills Areas?
2. To what extent do male and females' scores vary according to item format?
3. To what extent do male and females' scores vary according to skill areas?

Data Analysis Procedures

The research was conducted at KHU, Istanbul, Turkey. The sample of this study comprised 303 pre-intermediate level students in the second module of the preparatory school, 143 being females and 160 being males. The instrument of the research was the second achievement test of second module (see Appendix 1) which included five sections; reading, writing, listening, grammar and vocabulary.

The data analysis was carried out in several steps. The first step was to enter the data into SPSS and categorize the variables as according to female/male, item format and skills. First, the gender of each student was identified, and the total scores gotten by each student were entered into SPSS. Then, the total scores gotten from each skill were entered separately for each student. The test items (i.e. questions) were classified according to their format. There were both selected-response items; matching questions, fill in the blanks, find the correct form, multiple choice, and constructed-response questions; open-ended and paragraph questions. The total scores received from each item format were entered for each student.

In order to answer the research question 1a, the researcher divided the scores into two; females and males scores. First, the researcher used independent samples T-test in order to determine the variation, if any, in the scores according to gender, or dominance of either gender in terms of success. In order to answer the research question 1b, first the scores in different item formats were normalized by converting them out of 100. Since the total score of each item format differed. One-way ANOVA test was run to find out the means of different item format scores and to compare the different formats with one another. In order to answer the research question 1c, the researcher entered the test scores according to skill areas tested. The total score in different skill areas ranged from 25 to 10; hence, the researcher normalized the scores by converting them out of 100. The data were analyzed by conducting one-way ANOVA test, and the mean scores in different skill areas were compared to see whether the difference in the means of skill areas was statistically significant.

In order to answer the research questions 2 and 3, the researcher compared male and females' scores in terms of item format and skill areas and looked at whether there was a statistically significant difference between scores of males and females at a particular question format or a particular language skill. To identify the interaction effect between gender and item format, and gender and skill areas, the researcher ran two-way ANOVA test.

Variation of Turkish EFL Learners Scores in Language Achievement Tests

In this part the findings in regards to RQ 1 will be presented by examining whether the students' scores change depending on gender (RQ1a), item format (RQ1b), and skill areas (RQ1c).

Gender

The analysis of total scores and gender sought to answer research question 1a.

Table 2

Variation of Students' Scores according to Gender

scores			T-test		
	$\bar{x}$	*SD*	*df*	*t*	*p*
Males	60.07	16.17	301	-1.483	.139
Females	62.81	15.82			

According to the descriptive statistics, the difference between the means of males' and females' total scores was small ($\bar{x}$ male = 60.07, $\bar{x}$ female = 62.81). An

independent samples T-test was conducted in order to identify whether the difference between males' and females' mean scores is statistically significant. Even though the females performed a bit higher than males, the difference was not statistically significant. Further analysis was then conducted to identify whether the other independent variables; item format and skill areas have an effect on students' scores.

Item Format

The analysis of total scores and item format sought to answer research question 1b.

Table 3

Variation of Students' Scores according to Item Formats

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	467398.647	5	93479.729	167.250	.000*
Within Groups	1012769.611	1812	558.924		
Total	1480168.257	1817			

Note. df= degree of freedom; F=found variation of the group averages; Sig=significance; *p< .01

A one-way between subjects ANOVA was conducted to identify whether there was a statistically significant difference among the students' total scores according to the item formats. As shown in Table 3, there is a statistically significant difference between different item formats in terms of students' scores at the p< .01 level (F (5, 1812) =167.250, p = .000). That is, it makes a difference which item format is used in a test because students' scores change according to the type of the item format. Table 4 below shows the differences between students' mean scores in selected response questions and constructed response questions.

Table 4

Variation of Scores according to Selected Response and Constructed Response Questions

scores			T-test		
	$\bar{x}$	*SD*	*df*	*t*	*p*
Selected Response	69.42	14.40	604	7.558	.000*
Constructed Response	57.69	22.84			

Note. *p< .01

A further analysis was conducted to identify whether there was a statistically significant difference among the mean scores of students in selected response questions and constructed response questions. Selected response questions included 'matching,' 'fill in the blanks,' 'find the correct form' and 'multiple choice' questions, whereas the constructed response questions included 'open ended' and 'paragraph writing' questions. While the mean of selected response questions is 69.42, it is 57.69 in constructed response questions. The difference between the means is statistically significant at the p< .01 level. The students are apparently better at selected response questions. Table 5 below shows where the differences between students' scores in terms of individual item formats lie.

Table 5

Variation of Students Scores across Different Item Formats

IF(I)	N	Mean	Std. Dev.	IF (J)	Mean Diff. (I-J)	Sig.
1	303	90.8416	21.73824	2	33.30363	.000*
				3	44.90759	.000*
				4	7.47525	.001*
				5	30.58746	.000*
				6	35.70627	.000*
2	303	57.5380	19.42700	3	11.60396	.000*
				4	-25.82838	.000*
				5	-2.71617	.718
				6	2.40264	.811
3	303	45.9340	27.37976	4	-37.43234	.000*
				5	-14.32013	.000*
				6	-9.20132	.000*
4	303	83.3663	13.04175	5	23.11221	.000*
				6	28.23102	.000*
5	303	60.2541	33.87899	6	.11881	.083
6	303	55.1353	20.88200			

Note. IF=Item Format; 1=Matching; 2=Fill in the blanks; 3=Find the correct form; 4=Multiple Choice; 5=Open-Ended; 6=Paragraph ; * p< .01

As shown in Table 5, the students scored best in 'matching' questions (1) with a mean of 90.84. According to the results of Tukey HSD post-hoc analysis, the difference in the means of 'matching' questions with all the other item formats is statistically significant at the p< .01 level. The second best mean score was obtained in 'multiple choice' questions (4), x□ = 83.36. The mean of this item format is statistically different from all the other item formats at the p<.01 level. There was a big difference, 23 points, between the second best scored item format and the third. The mean of the third best score the students received was 60.25 in 'open ended'

questions (5). This finding reveals an interesting fact that while selected response questions seem to be easier for students to answer because they offer options and do not require production, the students scored better in 'open ended' questions, a constructed response item, than 'fill in the blanks' and 'find the correct form' formats which are selected response items. This finding may have implications regarding the scoring of 'open ended' questions. The answers to these questions might have been leniently scored and given easy points. Another implication might be a difference in the difficulty of selected response questions and constructed response questions in that selected response questions may have been more tricky and challenging than constructed resulting in lower grades in 'fill in the blanks' and 'find the correct form' formats. The mean differences between 'open ended' questions and 'matching,' 'find the correct form,' and 'multiple choice' are statistically significant at the $p < .01$ level; however, there is no statistically significant difference between 'open ended' and 'fill in the blanks,' and between 'open ended' questions and 'paragraph' questions. The mean of the fourth best score was in 'fill in the blanks' (2) questions , $x\square = 57.53$. The difference in the means of 'fill in the blank' between 'matching,' 'find the correct form' and 'multiple choice' questions was significant at the $p < .01$ level. The mean of the fifth best score the students obtained was in 'paragraph' questions (6), $x\square = 55.13$. Similar to 'open ended' questions, the mean differences between 'paragraph' questions and 'matching,' 'find the correct form,' and 'multiple choice' are statistically significant at the $p < .01$ level. The lowest mean belonged to 'find the correct form' questions (3), $x\square = 45.93$. This was the item format the students had most difficulty with and scored worst in. The difference between the means of 'find the correct form'

questions and all the other item formats is statistically significant at the p< .01 level. To sum up, item format has an effect on students' scores. While differences between different item formats are mostly statistically significant, few exceptions have been observed as presented in table 5. The next section will discuss whether the same observation can be made regarding the skill areas.

Skill Areas

The comparison of different skill areas sought to answer research question 1c.

Table 6

Variation of Students' Scores according to Skill Areas

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27634.960	4	6908.740	15.202	.000*
Within Groups	686225.431	1510	454.454		
Total	713860.390	1514			

Note. df= degree of freedom; F=found variation of the group averages; Sig=significance; *p< .01

A one-way between subjects ANOVA was conducted to identify whether there was a statistically significant difference among students' achievement scores according to the skill areas. As shown in Table 6, there is a statistically significant difference between different skill areas in terms of students' scores at the p< .01 level (F (4, 1510) =15.202, p = .000). That is, it makes a difference which skill area is assessed in a test because students' scores change according to the type of the skill areas. However, it should be noted that not all skill areas revealed a statistically significant difference. Table 7 below shows where the differences between students' scores in terms of different skill areas lie.

Table 7

Variation of Students Scores across Different Skill Areas

SA(I)	N	Mean	Std. Dev.	SA (J)	Mean Difference (I-J)	Sig.
1	303	65.5446	22.06051	2	10.40924	.000**
				3	3.18482	.352
				4	5.64356	.010**
				5	-1.65017	.876
2	303	55.1353	20.88200	3	-7.22442	.000**
				4	-4.76568	.047*
				5	-12.05941	.000**
3	303	62.3597	16.32531	4	2.45875	.615
				5	-4.83498	.042*
4	303	59.9010	20.09576	5	-7.29373	.000**
5	303	67.1947	20.06128			

Note. SA=Skill Areas; 1=Reading; 2=Writing; 3=Listening; 4=Grammar; 5=Vocabulary; * p< .05, ** p< .01,

As shown in Table 7, the students scored best in 'vocabulary' (5), x□ = 67.19. According to the results of Tukey HSD post-hoc analysis, the difference between the means of 'vocabulary' and 'writing' and 'grammar' is statistically significant at the p< .01 level, and with 'listening' at the p< .05 level. However, there is no statistically significant difference between the means of 'vocabulary' and 'reading.' The mean of the second best score was obtained in 'reading' (1), x□ = 65.54. The difference between the means of this skill area and 'writing' is statistically significant at the p< .01 level, and the significance level of the difference between the means of 'reading' and 'listening' is p< .05. On the other hand, there is

no statistically significant difference between the means of 'reading' and 'listening,' and 'reading' and 'grammar.' The mean of the third best score the students received was 62.35 in 'listening' (3). The mean difference between 'listening' and 'writing' is statistically significant at the p< .01 level, it is significant at the p< .05 with 'vocabulary.' However, there is no statistically significant difference between 'listening' and 'reading,' and 'listening' and 'grammar.' The mean of the fourth best score was received 'grammar' (4), x□ =59.90. The difference in the means of 'grammar' and 'vocabulary' is statistically significant at the p< .01 level, and the difference in the means of 'grammar' and 'reading,' and 'grammar' and 'writing' is statistically significant at the p< .05. On the other hand, there is no statistically significant difference between the means of 'grammar' and 'listening.' The lowest mean belonged to 'writing' (2), x□ = 55.13. This was the skill area the students had most difficulty with and scored worst in. Only 'writing' is statistically significant different from all the other skill areas. The difference between the means of 'writing' and 'reading,' 'listening' and 'vocabulary' is statistically significant at the p< .01 level, and it is statistically significant at the p< .05 level with grammar. The mean scores of 'grammar' and 'writing' are below the passing grade of KHU Preparatory School; 60, while the other skills mean scores are above. The fact that the students got higher in 'listening' is an interesting finding because while the students complain about the difficulty of 'listening,' and claim that they feel more comfortable with 'grammar,' they scored better in the 'listening' section of the test. To sum up, skill areas have an effect on students' scores. While differences between different skill areas are mostly statistically significant, a few exceptions have been observed as

presented in table 7. The next section will discuss to what extent the scores of male and females vary according to item format.

The Extent to which Male and Females' Scores Vary according to Item Format

The comparison of scores in different item formats in terms of gender sought to answer research question 2.

Table 8

Variation of Students' Scores according to Gender and Item Format

Dependent Variable: Score

Source	SS	df	MS	F	Sig.	Partial Eta Squared
Gender	1765.906	1	1765.906	3.189	.074	.002
Item Format	460666.757	5	92133.351	166.392	.000*	.315
Gender * Item Format	10997.074	5	2199.415	3.972	.001*	.011
Error	1000006.631	1806	553.714			
Total	9282594.000	1818				

a. R Squared = .324 (Adjusted R Squared = .320)

A two-way ANOVA test was conducted to identify whether gender has an effect on the achievement scores in different item formats. As presented in Table 8, gender does not have a main effect on students' scores in different item formats. This result suggests that being male or female does not influence students' performance in achievement tests.

On the other hand, the main effect of item format on students' scores is statistically significant such that scores in certain item formats are significantly higher than the others (F (5, 1806) = 166.392, p = .000, $\eta^{2 =}$.315). The significant main effect of item formats on the scores is expected given the one-way ANOVA results discussed in research question 1b.

More importantly, there is a highly significant interaction effect between gender and item format (F (5, 1806) = 3.972, p = .001).However, the eta squared statistic ($\eta^{2 =}$.011) indicated a small effect size which may be related to lack of main effect of gender on the scores. Because of the small effect size of interaction between gender and item format, item formats ($\eta^{2 =}$ .315) alone explain more of the variation while interaction does less. Figure 2 below shows the interaction effect between gender and skill areas.

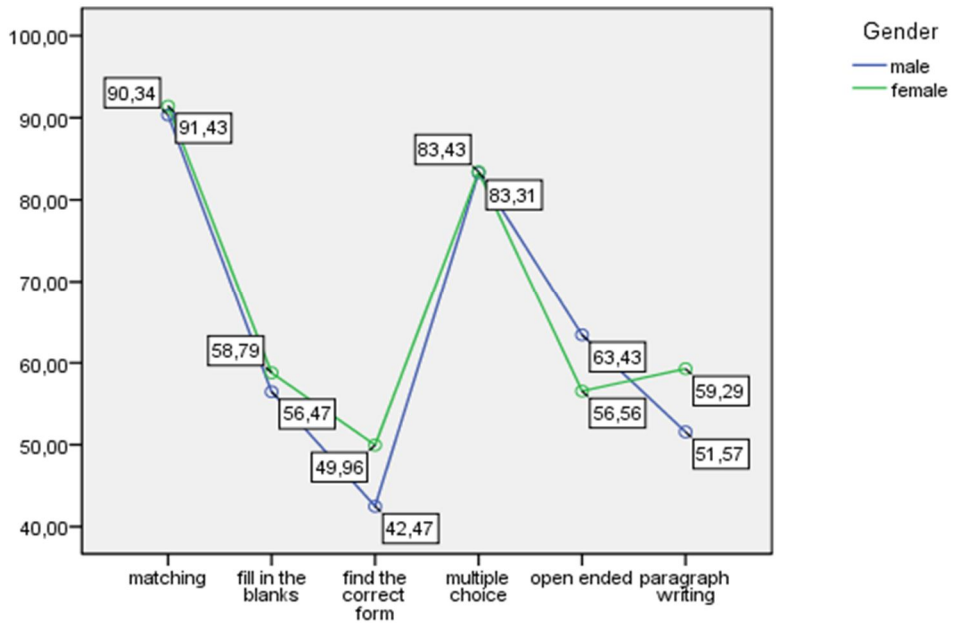


Figure 1 Gender and Item Formats

As shown in Figure 1, the interaction effect of gender and item format lies in 'open ended' questions where, unlike any other item format, males outperform females. Table 9 below shows the difference between males' and females' scores according to item format.

Table 9

Comparison of Mean Scores of Males and Females in Item Formats

		Scores		T-test		
		x□	SD	t	df	Sig. (2-tailed)
1.Matching Females	Males	90.34	21.75	-.435	301	.664
	Females	91.43	21.78			
2.Fill in the blanks	Males	56.46	19.65	-1.036	301	.301
	Females	58.78	19.14			
3.Find the correct form	Males	42.47	27.40	-2.393	301	. 017*
	Females	49.96	26.89			
4.Multiple Choice	Males	83.31	12.62	-.077	301	.939
	Females	83.42	13.55			
5.Open-ended	Males	63.42	34.12	1.767	301	.078
	Females	56.55	33.32			
6.Paragraph	Males	51.57	19.90	-3.257	301	.001*
	Females	59.28	21.29			

Note. *p< .01

As presented Table 9, the biggest difference between the means of males' and females' scores is in 'paragraph questions' (6). The mean of males' scores in 'paragraph questions' is 51.57, and it is 59.28 in females, and the difference is 7.71 which is a statistically significant result at the $p < .01$ level. The second highest difference between the means of males and females scores is in 'find the correct form' questions (3). The difference is 7.49; a statistically significant result at the $p < .05$ level, with a mean of 49.96 in females and 42.47 in males. There is again a female superiority in this item format. This item format is the format where both genders scored worst and received a failing grade. 'Find the correct form' and 'paragraph writing' questions are the only item formats with a statistical difference between males' and females' scores. The third highest difference between the means of males and females scores is in 'open ended' questions (5). The mean difference is 6.87, close to the difference paragraph questions and 'find the correct form' questions. Males' mean score is 63.42, whereas females' mean score is 56.55. This item format shows a different pattern because it is the only item format where males scored higher than females. Another interesting result is that while males mean score is higher than the passing grade at KHU Preparatory School; 60, females mean is below, so a failing grade. The fourth biggest difference between the means of males and females scores is in 'fill in the blanks' questions (2). There is 2.32 points difference between males mean scores; $x\square = 56.46$, and females mean score; $x\square = 58.78$. The fifth biggest difference is in 'matching' questions (1), the item format where both genders scored best. There is a small difference, 1.09, between the males; $x\square = 90.33$, and females $\bar{x} = 91.42$. The smallest difference between the means of males and females is in 'multiple choice' questions (4). Females mean score is $x\square =$

83.42, and males mean score is x□ = 83.31; thus, the difference between the means of males and females scores is only 0.11.

Overall, these results indicate that there is no main effect of gender on students' scores according to different item formats; however, item format does have a main effect on scores influencing students' success in different item formats. There is also an interaction effect between gender and different item formats, when these two variables interact, they do affect the scores students receive from different item formats.

The Extent to which Male and Females' Scores Vary according to Skill Areas

The comparison of scores in different skill areas in terms of gender sought to answer research question 3.

Table 10

Comparison of Gender and Skill Areas

Dependent Variable: score

Source	SS	df	M	F	Sig.	Partial Eta Squared
Gender	3710.506	1	3710.506	8.284	.004*	.005
Skill	26741.258	4	6685.315	14.926	.000*	.038
Gender * Skill	8418.024	4	2104.506	4.699	.001*	.012
Error	674096.900	1505	447.905			
Total	6542605.500	1515				

Note. *p< .01, a. R Squared = .056 (Adjusted R Squared = .050)

A two-way ANOVA test was conducted to identify whether gender has an effect on the achievement scores in different skill areas. As shown in Table 10, there is a significant main effect of gender on the scores received in different skill areas (F (1, 1505) = 8.284, p =.004, η^2 = .005). That is, being male or female influences students' performance in different skill areas.

As for skill areas, the main effect of skill areas on students' scores is also statistically significant such that scores in certain skill areas are significantly higher than the others (F (4, 1505) = 14.926, p = .000, η^2 = .038). The significant main effect of skill areas on the scores is expected given the one-way ANOVA results discussed in research question 1c.

More importantly, there is a significant interaction effect between gender and skill areas (F (4, 1505) = 4.699, p = .001). However, the eta squared statistic (η^2 = .012) indicated a small effect size. Although gender is significant, its effect size is really small, thus, the effect size for the interaction is also small. Figure 2 below shows the interaction effect between gender and skill areas.

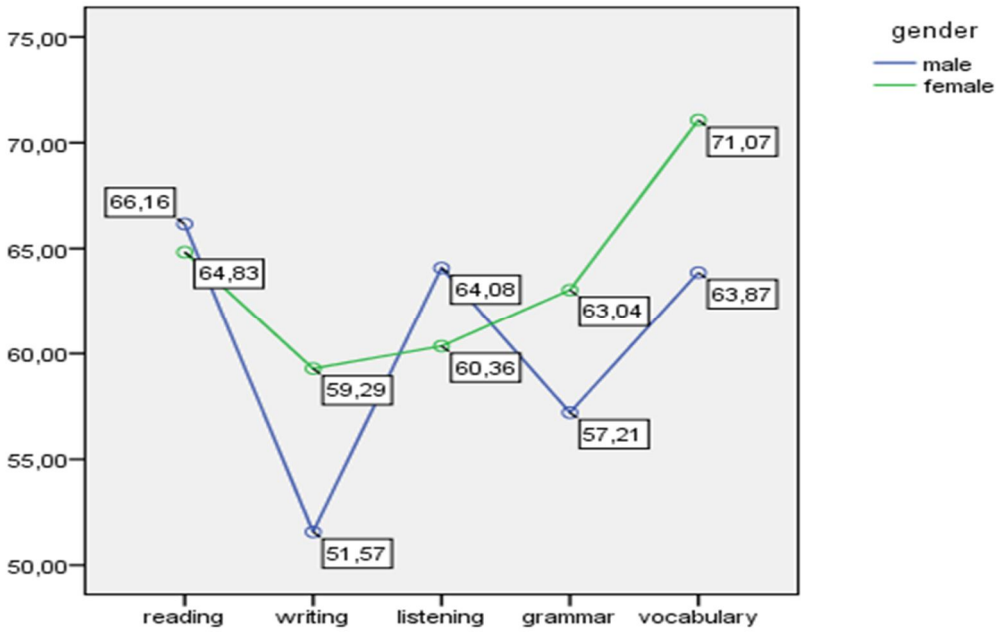


Figure 2 Gender and Skill Areas

As shown in Figure 2, the interaction effect of gender and skill areas lies in 'listening" questions where unlike any other item format males outperform females significantly. Table 11 below shows the difference between males' and females' scores according to skill areas.

Table 11

Comparison of Mean Scores of Males and Females in Skill Areas

	Scores		T-test		
	x̄	*SD*	*t*	*df*	*Sig. (2-tailed)*
1.Reading Males	66.16	22.29	.523	301	.601
Females Females	64.83	21.84			
2.Writing Males	51.57	19.90	-3.257	301	.001*
Females	59.28	21.29			
3.Listening Males	64.07	16.77	1.989	301	.048*
Females	60.35	15.60			
4.Grammar Males	57.20	20.20	-2.539	301	.012*
Females	63.03	19.57			
5.VocabularyMales	63.86	26.83	-2.419	301	.016*
Females	71.07	24.06			

Note. *p< .01

As shown in Table 11, the biggest difference between the means of males and females is in 'writing'; females' mean score is higher than males. While males' mean score is 51.57, females' mean is 59.28. The difference between the means of males and females is 7.71 which is a statistically significant difference at the p< .01 level. However, even if there is a big difference between the means of males and females compared to other skill areas, both groups have a mean below the passing grade. The second highest difference between the means of males and females is in

'vocabulary.' There is again a female superiority in this skill area. The difference is 7.21, with females mean score of 71.07, and males mean score of 63.86. The difference in this skill area is statistically significant at the $p < .05$ level. The third highest difference between the means of males and females is in 'grammar.' The difference between the mean scores of males and females is 5.83. Males' mean score is 57.20, whereas females' is 63.03. The difference in this skill area is again statistically significant at the $p < .05$ level. While females' mean score is above the passing grade, males' is below; thus, it is a failing grade. The fourth biggest difference between the means of males and females is in 'listening.' Different from 'writing,' 'vocabulary' and 'grammar,' in 'listening' males ($x\square = 64.07$) performed better than females ($x\square = 60.35$), and the difference in this skill area is statistically significant at the $p < .05$ level. Reading is the second skill where, even if small, only 1.33 points, males performed better ($x\square = 66.16$) than females ($x\square = 64.83$). This is the smallest difference between the means of males' and females' scores, and also it is the only difference which is not statistically significant.

Overall, these results indicate that there is a main effect of gender on students' scores according to different skill areas, and also, skills areas do have a main effect on scores influencing students' success in different skill areas. There is also an interaction effect between gender and different skill areas, when these two variables interact, they do affect the scores students receive from different skill areas.

Conclusion

In this chapter, the data obtained from the students' scores of a sample achievement test from KHU Preparatory School were analyzed and presented. In the first part, the total scores of 303 pre-intermediate level students in the sample achievement test, and results of the analysis of these scores in terms of their variation according to gender, item format and skill areas are shown. In the second part, males' and females' scores in different item formats, and results of the analysis of these scores in terms of their variation according to item format are provided. In the third part, males' and females' scores in different skill areas and results of the analysis of these scores in terms of their variation according to skill areas are presented.

According to the results, gender does not have a significant main effect on total scores of the students; however, there is a significant main effect of both item format and skill areas on the students' scores influencing their success in the test. There is also an interaction effect between gender and item format and gender and skill areas, when these two variables interact, they do affect the scores students receive from different skill areas. However, because of the small effect size of interaction between gender and item format, item format alone explain more of the variation while interaction does less. It is the same for skill areas as well, the effect size of interaction between gender and skill areas is small; thus, skill areas alone explain more of the variation while interaction does less.

The next chapter will continue with a discussion of the findings, pedagogical implications, limitations of the study, and implications for further studies.

CHAPTER V: CONCLUSION

Introduction

This study attempted to analyze whether language learners' scores in language achievement tests show any difference according to gender, item format (matching, fill in the blanks, find the correct form, multiple choice, open ended, paragraph writing) and skill areas (reading, writing, listening, grammar, and vocabulary), and to reveal whether male and females' scores show any difference according to item format and skill areas tested. In this respect, the study addressed the following research questions:

1. How do Turkish EFL learners' scores on language achievement tests vary according to
    a. Gender?
    b. Item Format?
    c. Skills Areas?
2. To what extent do male and females' scores vary according to item format?
3. To what extent do male and females' scores vary according to skill areas?

The sample of this study comprised 303 pre-intermediate level students at KHU, 163 being males, and 140 being females. The second achievement test of the second module belonging to these students was analyzed. The test included five sections; 'reading,' 'writing,' 'listening,' 'grammar' and 'vocabulary.' There were six different item formats in these sections; 'matching,' 'fill in the blanks,' 'find the correct form,' 'multiple choice,' 'open ended' and 'paragraph writing' questions.

Students' total scores were analyzed according to gender, item format and skill areas, and the means of males' and females' scores in terms of different item formats and skill areas were also analyzed and compared.

In this chapter, the research findings will be discussed and evaluated in light of the research questions and the relevant literature. Within the scope of the chapter, pedagogical implications, limitations of the study, and suggestions for the further research will also be presented.

Findings and Discussion

Variation of Scores according to Gender, Item Format and Skill Areas

Gender. The second achievement test of the second module in pre-intermediate level classes at KHU Preparatory School was analyzed. The total test scores of the students were first grouped according to the gender of the students. Then, the scores of males and females were compared. There was a small difference between the means of males ($x\square = 60.07$) and females' ($x\square = 62.81$) scores. This result can be interpreted as an indication of language learning skills since the students who are good at language learning are expected to score high in a language test. The literature is torn in between no difference between males' and females' language achievement and females' superiority. For example, Feyten (1991) investigated and compared the general language learning skills of male and female language learners, and she found no differences in general language learning skills of male and female foreign language learners. Similarly, Ehrman and Oxford (1995) did not find a difference in performance between males and females in language

achievement “by any measure” (p. 81). However, according to the results of Boyle’s (1987) study, the female language learners were stronger in overall language ability, and Gu (2002) also reported that females outperformed males significantly on a general proficiency test. The findings of the present study, in terms of students’ total scores, seem to be more in line with studies such as Sunderland’s study (2000) which show a slight advantage for females rather than big.

To sum up, the findings of this study revealed a small difference between males and females in terms of total scores. These results confirm the literature that emphasizes the variation resulting from individual differences rather than gender (e.g., Ehrman & Oxford, 1995; Nyikos, 2008).

Item format. There were two basic question types in the test analyzed in the present study; selected response questions and constructed response questions. Selected response questions included ‘matching,’ ‘fill in the blanks,’ ‘find the correct form’ and ‘multiple choice’ questions, whereas the constructed response questions included ‘open ended’ and ‘paragraph writing’ questions. While the mean of selected response questions is 69.42, it is 57.69 in constructed response questions. The students are apparently better at selected response questions. This is an expected result because the skills students are required to possess while answering constructed response questions are much more varied than selected response questions. More specifically, it is enough to recognize or identify the correct option in selected response questions; however, in order to answer a constructed response question, the test takers are required to actually produce the language; therefore, a variety of more complex skills are at play.

Research also revealed that in the field of educational testing that selected response questions, especially 'multiple choice,' and constructed response questions provide essentially the same information by measuring the same constructs; hence, 'multiple choice' questions can be used as a substitute for constructed response questions (Lukhele, Thissen, & Wainer, 1994). This claim may have been made depending on those studies indicating a high level of agreement between the scores on 'multiple choice' and constructed response questions (e.g., Godschalk, Swineford, & Coffman, 1966). However, as the findings of the present study indicate, the results of 'multiple choice' and constructed response questions are quite different. To be more specific, while the mean score of 'multiple choice' questions is 83.36, it is 55.13 in 'open ended' questions and 60.25 in 'paragraph writing' questions. The big difference between the scores received in 'multiple choice' and 'constructed response' questions in the test show no agreement of any sort between these two formats. Hence, the findings are more in line with the research which concluded that selected response questions, especially 'multiple choice,' and constructed response questions probably examine different levels of cognition (e.g., Bridgeman & Rock, 1993; Walstad & Becker, 1994, Kuechler & Simkin, 2004).

Other than the difference between selected and constructed response questions, there are also significant differences among individual item formats. The students received the best score in 'matching' questions, and the second best score in 'multiple choice' questions. While these results were expected, that the students would be more successful with selected response questions, there is an interesting finding about the third best score received which is in 'open ended' questions. The

mean score of 'open ended' questions is higher than 'fill in the blanks' which is the fourth best score and 'find the correct form,' which is the lowest mean score. The reason why 'open ended' questions have a higher mean might be that these questions were asked in the 'reading' section of the text (see Appendix 1), and included scanning questions as well. Scanning questions do not require higher order skills, or as various or complex skills as open ended comprehension, inference or paragraph writing questions. Thus, the students might have found the 'open ended' questions easy, especially the scanning questions. Another reason might be related to the skill area tested, which will be discussed in detail later in this chapter. Compared to the other skill areas tested, the students' second best mean score is in 'reading.' Thus, the students are more successful at 'reading,' which might be the reason why they scored higher in the questions asked in the 'reading' section even though the questions are constructed response type.

'Fill in the blanks' questions follow 'open ended' questions as the fourth best score. The reason why 'fill in the blanks' comes after 'open ended' and other selected response item formats might be that this item format requires the ability to understand context and vocabulary, and then depending on this understanding, the ability to identify the correct words or type of words that belong in the deleted passages of a text. Thus, the cognitive load on the students seems to be higher with this type of questions.

'Paragraph writing' question has the fifth best score in the analysis. It is an expected result because the question has a higher cognitive demand. Apart from being a constructed response question and requiring the complex skills any

constructed response question type requires, ‘paragraph writing’ questions also pose another challenge for the students; writing anxiety (Cheng, 2004; Cheng, Horwitz & Schallert). Besides the linguistic requirements, writing anxiety or apprehension, which is defined as the tendency to avoid writing situations or to react in an anxious manner because of the anticipation of negative consequences (Daly & Miller, 1975), might also have an effect on ‘paragraph writing’ question’s mean being lower than other item formats. Hence, the students’ writing anxiety, if they had, might be the cause of poor grades in paragraph writing.

Another interesting finding of the analysis of item formats is that ‘find the correct form’ questions has the lowest mean among all the other item formats, even lower than ‘paragraph writing’ questions. One reason why ‘find the correct form’ questions have a lower mean than ‘paragraph writing’ questions might be related to scoring. ‘Paragraph writing’ questions are open to subjectivity during scoring. The teachers grading the papers might have been lenient towards the students’ answers and might have assigned some easy points. On the other hand, while grading ‘find the correct form’ questions there is only one answer specified in the key; thus, the scoring is objective and standard across all the papers. Also, ‘find the correct form’ type is only used in the ‘grammar’ section of the test. There are two parts with this item format, one part asking the students to ‘find the correct form’ of the verbs in the parenthesis and complete the sentences with either the Past Simple or the Past Continuous tense, and the other part asking the same task with the Present Simple, the Present Continuous, and the Present Perfect tense. There is a short paragraph with blanks in these two parts (see Appendix 1). The low mean of ‘find the correct form’

questions might be related to the exam itself or the instruction. The paragraphs seem to be lacking enough content for students to refer to and the blanks are too close, especially in the past tense section. The students might not have found the correct answers because they had difficulty in understanding the context and the vocabulary. Another reason might be instruction. The mean of 'find the correct form' questions is 10 points higher than the mean of 'paragraph writing' questions even though 'paragraph writing' question type is assumed to be more challenging. This fact might imply that there was not enough or effective instruction in the classroom regarding the topics asked with this item format. The students might have failed because they had not learnt the grammar topics asked in the exam at all.

To sum up, there are both expected and unexpected findings regarding the scores in different item formats. These differences might be caused by several different reasons as aforementioned. There has been no research comparing test takers' achievement in different item formats in a second language test; thus, the findings in the present study might be confirmed or contradicted by further research in the future.

Skill areas. The test analyzed is composed of five skill areas; 'reading,' 'writing,' 'listening,' 'grammar' and 'vocabulary.' Previous research mostly focused on one individual skill area and factors that might cause variation in success in that skill (e.g., Catalan, 2003; Gu, 2002; Knudson, 1995); however, in this study, it was aimed to compare the different skill areas in terms of students' success. According to the results, the highest mean score belongs to the 'vocabulary' section of the test. One reason might be that the words asked in the exam were directly taken from the

course books of the students. The students usually do not have an idea about the stimulus text in 'reading' or 'listening' and the specific topic in 'grammar'; thus, there is the surprise factor in these skill areas. However, in the 'vocabulary' section of the test, they have more ideas about what can be measured in the test; at least they have a list of words to refer to; hence, the surprise factor in the 'vocabulary' section of the test is minimized. Also, 'vocabulary' seems to be the skill area where study skills benefit the students more. When the students study hard or memorize the words, they have a good chance to answer the test questions correctly. Another reason might be the help of contextual clues. There are many clues in the context that help students identify the correct word. They may refer to part of speech, preceding or following words, or connotation of the word required. The context the 'vocabulary' questions asked is also similar to the ones in their books; thus, they have a good chance to remember the information they learnt about the context. This is another factor that might help them score better in the 'vocabulary' section since the context will help them activate their schemata (i.e., prior information, knowledge or experience of the topic of the text) so that the students can make sense of the material (Thornbury, 2006).

The second best score in the test belongs to 'reading', and 'listening' skills; the third best score, follows it with a small difference. The reason why students scored better in these skill areas than in 'grammar' or 'writing' might be as a result of the test itself, instruction and/or individual factors. The topic of the stimulus test might be relevant to the students' previous learning; hence, they might be familiar with the vocabulary, the questions might be well designed and prepared with no

ambiguity which might contribute to student' success. Also, most of the questions in the 'reading' section except for 'open ended' questions are selected response questions which are appropriate for objective scoring. Another reason might be that students might have been trained well for reading and listening strategies. Strategic reading and listening is very important and emphasized in KHU's academic English classes; therefore, the students might have applied these strategies in the exam received higher scores.

The fourth best mean score the students obtained in the test is in the 'grammar' section. Even if students mostly consider grammar an easier skill because they find it more concrete and they have rules to refer to, the students in the present study received low scores in 'grammar.' This finding might be related to the item formats in the 'grammar' section, as mentioned before, or problems in the instruction of the topics asked.

Finally, the lowest mean among the skill areas belongs to the 'writing' section. Writing has always been a challenging skill area for language learners. Previous studies indicate that productive skills cause more anxiety; thus, may result in poor achievement (e.g., Hilleson, 1996; Zhang, 2001). The fact that writing has the lowest mean score which is below the passing grade might be caused by writing anxiety or simply insufficiently developed writing skills (MacIntrye, 1995).

To sum up, there are both expected and unexpected findings regarding the scores in different skill areas. These differences might be caused by several different reasons as aforementioned.

The Extent to which Males' and Females' Scores Vary according to Item Format

The mean scores of males and females in different item formats in the test were analyzed and compared. Males' and females' scores varied significantly in two item formats, 'find the correct form' and 'paragraph writing' questions. In both item formats, females scored significantly higher than males. This higher achievement of females in some item formats ('find the correct form' and 'paragraph writing') both contradicts and confirms the findings of Ryan and Demark's study (2002). In their study, the researchers found out that in language measures, males score higher than females using selected response questions, while females have higher achievement in constructed response items. The findings of the present study confirms Ryan and Demark's study (2002) in the sense that females outperform males in constructed response questions; however, according to the present study, in 'find the correct form' format, which is a selected response item, females again outperform males. The reason why females' achievement is higher than males in some item formats ('find the correct form' and 'paragraph writing') of the test can be attributed to several factors such as different study skills, differences in cognitive abilities, language competence, affective factors such as anxiety, confidence, motivation, and the skill area tested with these item formats. In their review article Oxford, Nyikos and Ehrman argued that (1988) women remember more details. This might explain why females scored higher in 'find the correct form' questions; remembering the correct verb form is essential to answer the 'find the correct form' questions in the test because all the questions in this format were about verb tenses. If females are better at remembering things, then this ability gives them a great advantage over

males in question types such as ‘find the correct form’ because the memory factor is really influential. Also, according to the findings of Graham’s (1997) study, females are more likely to learn grammatical items “by heart” (p.81), this finding confirms the assumption that females are better at remembering things. Oxford and Nyikos(1989) argue that women have a greater tendency for social approval, and this tendency motivates them to strive for higher grades than men. Confirming this argument, Van Houtte (2004) claims that females’ culture is more study-oriented and supportive of academic achievement. If so, females might have studied harder than males to memorize the verb forms, and learnt the previously taught grammar topics better, and thus, received higher grades in ‘find the correct form’ questions.

Most studies focusing on gender differences investigated the differences in skill areas and strategies used (e.g., Bacon, 1993; Boyle, 1987; Catalan, 2003; Gu, 2002) rather than item formats. The findings at this study regarding the difference in the item formats in this study might also be explained in reference to the skill areas these item formats were used. In her study, Graham (1997) found out that male learners of German were more comfortable about grammar than females, in the same study, female learners of French expressed greater worries for grammar. Hence, even though Graham (1997) has participants who were learning two different languages, the findings are consistent in that the male participants in her study felt more comfortable with grammar. However, in the present study, the fact that the males performed worse than the females in ‘grammar’ where ‘find the correct form’ questions are asked, might imply that males do not feel comfortable with ‘grammar.’

This attitude, which may stem from lack of competence, confidence or other affective factors such as motivation may have affected their test results as well.

As far as the 'paragraph writing' question, which cannot be evaluated separately from writing skill, is concerned, the fact that females outperformed males was an expected result (e.g., Graham, 1997; Knudson, 1995). According to the results of the study conducted by Knudson (1995) females are better writers. Graham (1997) also found out that in writing tasks, females used a careful, organized approach which included planning, monitoring and evaluating more frequently than males. This kind of approach might explain the higher scores of females in the 'paragraph writing' question of the test.

To sum up, females significantly outperformed males in two item formats; 'find the correct form' and 'paragraph writing' questions. The reason why females scored higher, as aforementioned, might be related to study skills, differences in cognitive abilities, language competence, or affective factors such as anxiety, confidence, or motivation. The next section will discuss to which extent male and females' scores vary according to skill areas.

The Extent to which Males' and Females' Scores Vary according to Skill Areas

The scores of males and females in different skill areas in the test were analyzed and compared. Males' and females' scores varied significantly in four skill areas; namely 'writing,' 'listening,' 'grammar' and 'vocabulary.' In 'writing,' 'grammar' and 'vocabulary,' females outperformed males; however, only in 'listening' males scored higher than females.

The female superiority in most skill areas in the test confirm the results of many other studies conducted (e.g., Burstall, 1975; Boyle, 1987; Nyikos, 1990). The reason why females' achievement is higher than males in most part of the general language skill areas ('writing,' 'listening,' 'grammar' and 'vocabulary') of the test can be attributed to several factors such as different study skills, different strategies for language learning, differences in cognitive abilities, biological factors, language aptitude, affective factors such as attitudes, motivation and cultural norms.

According to the literature, females are better at general study strategies, refer to rule-related strategies (Oxford & Nyikos, 1989), and use conscious learning strategies more than males do (Oxford, 1993). Yang (2001) also notes that there are gender differences regarding the development of cognitive abilities. These differences may result in different cognitive strategy preferences by males and females. These different cognitive strategies may favor one group in language learning, apparently females in this study. Ehrman and Oxford (1995) list the language learning strategies more often used by females as metacognitive strategies (higher order executive skills; planning, evaluating, organizing), affective (emotional and motivational) and social. Thus, females in the present study may have surpassed their male counterparts because they have more advanced and effective language learning strategies. There could also be some biological factors that provide females with an advantage over males. According to Springer and Deutsch (1989), behavioral and clinical data indicate that women are less lateralized for language functions, and they are superior to men in language skills. There are also some researchers who related the performance differences between males and females to aptitude. Powell

(1979) claims that females are superior to males in all aspects of linguistic process; therefore, show a greater aptitude for language. Rua (2006) also states that "although both males and females have the same linguistic potential as human beings (aptitude in general sense), females' linguistic skills somehow seem more prone to be stimulated in order to reach higher levels of linguistic competence" (p. 103).

Another reason why females outperformed males in three major skill areas might be related to attitudes. Ellis (1994) recognizes attitude as the obvious explanation for females' greater success in L2 learning. Spolsky (1990) looks at the relationship between attitude and success in language learning from a different perspective. He suggests that "attitudes do not have direct influence on learning, but they lead to motivation, which does" (p.49). In this respect, the studies conducted by Pritchard (1987), Powell and Littlewood (1983), and Powell and Batters (1985) show that unlike males, females are more favorably inclined to the language itself, the speakers and cultures of other languages. Burstall (1975) argues that females manifest more integrative reasons for studying a language (e.g., interest in getting to know the speakers and culture of the language), while males' motives are more instrumental (e.g., seeing language as an instrument). Krashen (1988) regards integrative motivation as a stronger predictor of achievement than instrumental motivation. Socialization also seems to determine, or at least, influence motivation, and also cognitive development (Slavin, 1988). Social forces such as parental attitude and gender related cultural beliefs determine how males and females perceive the process of language learning, and the value they attach to other languages. Hence, in a culture, like Turkish, where language is seen as a women's topic, it is natural to

find out female superiority in language learning. This cultural norm of associating language with females, seeing it as ‘a girl thing’ might have affected the results in the test.

To sum up, females’ higher achievement in general language skill areas of the test can be attributed to several factors such as different study skills, different strategies for language learning, differences in cognitive abilities, biological factors, language aptitude, affective factors such as attitudes, and cultural norms.

Apart from these reasons, specific comparison of scores in the language skill areas is essential to understand the sources of differences between males and females. As for writing, the females outperformed males significantly. This finding confirms Knudson’s (1995) findings which revealed that females are better writers. On the other hand, there are also studies with contradictory findings. Pajares and Valiente (1996) found no differences between males’ and females’ writing performances in their study; however, they reported that females had higher writing self-efficacy. Self-efficacy is defined as “people's beliefs about their capabilities to produce designated levels of performance that exercise influence over events that affect their lives; self-efficacy beliefs determine how people feel, think, motivate themselves and behave” (Bandura, 1994, p. 2). Meier, McCarthy and Schmeck (1984) reported that self-efficacy predicted the writing performance of the undergraduate students. Shell (1989) also investigated writing self-efficacy of the undergraduate students and found a strong correlation between students’ confidence in their writing skills and their grades in holistic scales. Thus, the differences between males’ and females’ achievement in the present study might be related to

students' self-efficacy in writing in the sense that the females might have had higher self-efficacy than the males which advantaged them in writing scores. There are numerous studies which reported that females express stronger self-beliefs in language arts than do males (e.g., Eccles, Wigfield, Flanagan, Miller, Reuman, & Yee, 1989; Pajares, Miller, & Johnson, 1999; Pajares & Valiante, 1997; Wigfield et al., 1991).

Since females express stronger confidence in their writing capabilities than do males (Pajares & Valiante, 1997), this might affect their ability to employ various self-regulatory strategies such as self-observation, self-evaluation, and self-correction while writing (Zimmerman & Martinez-Pons, 1990) positively which may result in higher achievement. Another reason that led females to perform better in writing might be related to writing apprehension. Daly and Miller (1975) reported strong correlation between apprehension and perceived likelihood of success in writing; they also found out that males were more apprehensive about writing than females. Similarly, the male students in the present study might have experienced more apprehension towards writing, and this might have affected their achievement in writing negatively.

Apart from the affective factors, the approaches and strategies employed by males and females during writing might also have an effect on their scores. According to Graham (1997), females use a careful, organized approach to writing which includes planning, monitoring and evaluating in writing tasks more frequently than males, and in the same study males expressed a dislike for planning in writing. Hingley (1983) also notes that females are encouraged to be more conscientious than

males, and this might have given them an advantage in written work and formal language use. This more planned and organized approach of females to writing might have favored them in getting higher scores.

'Grammar' is the second skill area in which females scored significantly higher than males. This was an expected result because females have been reported to use formal rule-related practice more frequently than males (Oxford & Nyikos, 1989). Rule-related practice obviously provides an advantage in the grammar skill which prescribes specific rules to be followed. The females in the study, as indicated in the literature, might have executed more appropriate strategies such as rule practice while studying for grammar, and this might have favored them in achievement in the grammar skill. Other than using the right strategies, females have also been reported to have more desires for higher grades than males (Oxford, Nyikos & Ehrman, 1988). The combination of desire, thus motivation, and more frequent strategy use might have naturally led the female students in the present study to greater success in learning grammar and greater achievement in the test. The aforementioned affective issues such as self-efficacy, confidence, and motivation might have also played a role in the difference between the 'grammar' scores of males and females.

The other skill females scored higher in than males is 'vocabulary.' The finding regarding higher achievement of females in 'vocabulary' is confirmed by Gu's (2002) study in which females reported more use of almost all the strategies that are associated with success in EFL learning (Gu & Johnston, 1996) such as guessing, using contextual clues, taking notes and employing oral repetitions.

Catalan (2003) also found out that females use a higher number of vocabulary strategies than males. In Gu's (2002) study, it was also reported that females spend significantly more extra-curricular time on learning English than their male counterparts. When combined, employing more effective strategies and investing more in language learning by spending more time for it might explain female superiority in 'vocabulary.' In Oxford, Lavine, Hollaway, Felkins and Saleh's (1996) study, females have been found to try out new techniques for vocabulary learning. Different preferences in the number and type of vocabulary learning strategies might have caused a difference in the students' achievement in vocabulary. Moreover, just as 'writing' and 'grammar,' affective factors might have affected the results of 'vocabulary' section of the test as well.

The results revealed a different pattern in the 'listening' skill; the males significantly outperformed the females. This finding conflicted with the studies of Markham (1988) and Bacon (1993) who found no differences in the listening comprehension of males and females. The finding of this study regarding males' superiority in 'listening' also contradicts the findings of the study conducted by Farhady (1982) who found out that females have higher comprehension in listening than males. Eisenstein (1982) argued that females can discriminate dialects and prestige of dialects better than males. Different from these studies, Boyle (1987) found out that males are better in recognizing words in 'listening' texts, which may bring higher comprehension; however, he also reported that females are better in general 'listening' comprehension (as cited in Kunnan, 1998). Larsen, Freeman and Long (1991) also argued that females are better in listening. The findings of the

present study seem to contradict the aforementioned studies in the sense that, instead of females, males performed better in 'listening.' The reason why the present study contradicts the other studies might lie in affective factors such as anxiety, confidence, or motivation. Bacon's (1992) study on the listening strategies which indicates that males are more confident in their ability to tackle an aural passage might support this assumption. The difference might be also related to the listening strategies employed by males and females. In her study, Bacon (1992) also reported that females used both metacognitive and cognitive strategies; the strategies that manipulate information such summarizing or reorganizing. Females adjusted metacognitive strategies according to the difficulty of the passage, but they used cognitive strategies in the same fashion without adjusting, even when they listened to more difficult passages. Males in the same study (Bacon, 1992) dealt with more difficult texts more aggressively with reference to bottom-up strategies more frequently, and reference to their mother tongue. In Graham's (1997) study, males were more likely to use problem-identification, a more direct even confrontational strategy. Hence, strategy preferences of males and females in the exam might have resulted in different achievement scores favoring males. If the superiority of males stems from their use of strategies, it means that the strategies they employ, work.

To sum up, this study yielded both expected and unexpected results in terms of variation of males' and females' scores in skill areas. While females were significantly better at 'writing,' 'grammar' and 'vocabulary,' males outperformed females at 'listening.' These results might have been caused by a number of different reasons or interaction of these reasons. While some findings of this study confirm the

literature of gender studies, some contradict. The findings might be supported or contradicted by further research comparing genders in their achievement in terms of item formats and skill areas.

Pedagogical Implications

According to the findings, females outperform males significantly in two item formats; 'find the correct form' and 'paragraph writing' questions, whereas males did not show any superiority in any of the item formats. Also, in skill areas, females outperformed males in three skill areas; 'writing,' 'grammar' and 'listening,' while males scored higher only in one skill area; 'listening.' These differences might have been caused by several factors such as differences in strategy use, and affective factors that can be remedied in the classroom by instructors' guidance and support; as well as other factors related to the exam itself which can be overcome by the test developers at schools.

First of all, understanding the strategies employed by male and female students will help instructors guide their students. Even though it does not directly stem from the results of this study, strategy training is recommended and considered a vital step to minimize the differences in achievement of males and females by the researcher. By informing the students and training them to self-control, teachers can help students monitor their comprehension and learning processes better. Also, being aware of the variety of the strategies used by the students, teachers might be more sensitive to different learning styles and intelligences. If there are any strategies that work better for a particular skill, students' might be trained to employ those strategies, or at least taught and given an option to select those strategies depending

on which skill they are required to use. By this way, teachers can take the steps to maximize the chances of success.

Another finding is related to the aforementioned affective factors that might accelerate or inhibit learning. Howell-Richardson and Parkinson (1988) pointed out that students may lack motivation if they do not see that "there is something in it" for them (p. 79). Hence, it is very important teachers and students to have a good relationship and open communication. Teachers should inform students about the benefits and ways of learning English. They should try to appeal to students with different learning styles and strategies. To minimize apprehension, teachers should consider having individual meetings with students, and should create a non-threatening, practice-like environment where grades are de-emphasized.

Another implication of this study is that factors related to the exam itself such as the question type, or scoring might cause variation in the scores. The test developers must make sure that the instructions, the question types, the context where the questions are placed in (e.g., 'fill in the blanks') are unambiguous and comprehensible. Variety of question types in any skill area is also important not to favor one group who is good at a particular question format. The topics of the stimulus texts in the exam should be also neutral; not appealing to either males or females in an obvious way. Given that females are better at memorizing, the questions requiring memorization should be minimized not to give an unfair advantage to the females. To eliminate the surprise factor, which affects the scores in some skill areas negatively, the topics and themes to be focused on in the test could be selected from those taught in the class. Scoring procedures are very important as

well. Scoring must be made as objective as possible maybe by having multiple scorers grading the same papers, having discussion meetings, and revising answer keys with the teachers' feedback.

Limitations of the Study

There are several limitations of this study which suggest that the results should be interpreted cautiously. As mentioned before, any differences in the scores depending on gender may reflect instruction or socialization processes that vary according to the culture of the setting where teaching takes place; thus, the same test may provide different results if it is administered in a different cultural context.

Another limitation of the study which stems from the test analyzed is that it does not include a speaking section. If there were a speaking section in the test, it would provide valuable results regarding differences, if any, depending on the gender of the students, and also its place compared to other skill areas in terms of students' success.

A further limitation is the number of different item formats. While there are four different selected response questions; 'matching,' 'fill in the blanks,' 'find the correct form' and 'multiple choice' questions, there are two types of constructed response questions; 'open ended' and 'paragraph writing.' Hence, there is a difference in the number of item format types. If the item format types and the number of the questions in these formats were equal, a better comparison of the results could have been done.

Suggestions for Further Research

Based on the findings and the limitations of this study, suggestions can be made for further research. This study can be replicated with tests administered at different proficiency levels. Students at different proficiency levels might tend to succeed differently in different item formats or skills than the students in this study.

Also, analysis of tests administered to students in a different culture or students who have come from different cultures and receive education in the same class might provide different results. Third, because the test investigated in this study lacks a speaking part, analysis of a test with a speaking section or an individualized speaking test could provide valuable results.

Conclusion

This study revealed that gender does not have a significant effect on the total scores of the students in language achievement tests. On the other hand, students' total scores vary significantly depending on both the item format and skill areas in the test. In other words, students' success differs in different item formats or skill areas. Males' and females' mean scores also show differences depending on both item format and skill areas. According to the findings, females outperform males significantly in two item formats; 'find the correct form' and 'paragraph writing' questions, whereas males do not show any superiority in any item format. Also, in skill areas, females outperform males in three skill areas; 'writing,' 'grammar' and 'listening' while males score higher only in one skill area; 'listening.'

This study compared the performances of two different groups, males and females, success in different item formats and skills. It also investigated the variation of scores in different item formats and skill areas according to gender. Thus, the present study contributes to the existing literature by having studied gender differences. With results both confirming and contradicting the previous research, this study has a unique place in the language testing literature by looking at the variation of scores according to three variables; gender, item format and skill areas, that have been studied together for the first time, and comparing males' and females' scores in terms of item format and skill areas again for the first time. The wide spectrum adopted while evaluating the differences in the results, and speculations made about these differences can benefit both future researchers in the field in terms of theoretical perspectives, and teachers and administrators in terms of practical perspectives so that learners are measured more appropriately and correctly based on their true abilities.

REFERENCES

Alexander, L. G. (1990). Why teach grammar? In J.E. Alatis (Ed.) *Linguistics, language teaching, and language acquisition: The interdependence of theory, practice, and research* (pp. 377-382). Georgetown University Press.

Arnot, M. & David, M. & Weiner, G. (1996). *Educational reforms and gender equality in schools. Equal Opportunities Commission Research Discussion Series*, No:17, Manchester: EOC.

Bachman, L. & Savignon, J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal, 70* (4), 380- 390.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Bachman, L.F., & Palmer, A.S. (1996). *Designing and developing useful language tests.* New York: Oxford University Press.

Bacon, S.M. (1992). The relationship between gender, comprehension, processing strategies and cognitive and affective response in foreign language listening. *The Modern Language Journal, 76* (2), 160-176

Bacon, S. M. (1993). The relationship between gender, comprehension, processing strategies and cognitive and affective response in foreign language listening. *Modern Language Journal, 76,* 160-178.

Badger, R & White, G. (2000). A process genre approach to teaching writing. *ELT Journal*, *54* (2), 153-160.

Bailey, K. M. (1998). *Learning about language assessment.* U.S.A.: Heinle & Heinle.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior,* vol. 4, pp. 71-81). New York: Academic Press. (Reprinted in H. Friedman [Ed.], Encyclopedia of mental health. San Diego: Academic Press, 1998).

Bensoussan, M. & Zeidner, M. (1989). Anxiety and achievement in a multicultural situation: the oral testing of advanced English reading comprehension. *Assessment and Evaluation in Higher Education*, *14* (1), 40-54.

Boyle, J. P. (1987). Sex differences in listening vocabulary. *Language Learning, 37,* 3273-3284.

Bridgeman, B. & Rock, D. (1993). Relationship among multiple choice and open ended analytical questions. *Journal of Educational Measurement, 30* (4), 313-329.

Brindley, G. (2006). Assessment. In R. Carter, & D. Nunan (Eds.) *Teaching English to speakers of other languages* (pp. 137-143*)*.Cambridge: Cambridge University Press.

Brown, A & McNamara, T. F. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics, 18*, 304–319.

Brown, D., H (2001). *Teaching by Principles: An Interactive Approach to Language Pedagogy.* San Francisco: Longman Inc.

Brown, H.D. (2004). *Language assessment: Principles and classroom practice*. Longman.

Brown, J. D. (1996). *Testing in language programs.* New Jersey: Prentice Hall Regents.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. New York: Cambridge University Press.

Byram, M. (2004). *Routledge encyclopedia of language teaching and learning.*

Routledge. Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.) *Language testing symposium: A psycholinguistic perspective* (46-69) London: Oxford University Press.

Cheng, Y. (2004). A measure of second language writing anxiety: Scale development and preliminary evaluation. *Journal of Second Language Writing, 13* (4), 13-335.

Cheng, Y., Horwitz, E., & Schallert, D. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning, 49*, 3.

Cross, D. (1983). Sex differences in achievements. *System*, 11(2), 159-62.

Crystal, D. (2004). In word and deed. *TES Teacher,* 26.

Dalkılıç, N. (2001). The role of foreign language classroom anxiety in English speaking courses. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8,* 70-82.

Daly, J.A. & Miller, M.D. (1975). Apprehension of writing as a predictor of message intensity. *Journal of Psychology, 89,* 175–7.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing 14* (3), 328 – 339.

Davies, A. (1999). *Dictionary of testing.* Cambridge University Press.

Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge University Press.

Eccles, J. S., Wigfield, A., Flanagan,C., Miller ,C., Reuman, D., & Yee, D. (1989). Self -concepts, domain values, and self-esteem: Relations and changes at early adolescence. *J. Person. 57,* 283-310.

Ehrman, M. E. & Oxford, L. R. (1995). Cognition plus: Correlates of language learning success. *Modern Language Journal, 79* (1), 67-89.

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

Eisenstein, M. (1982). A study of social variation in adult second language acquisition. *Language Learning, 32,* 367-391.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quarterly*, *16*, 43-59.

Feyten, C. M. (1991). The power of listening ability: An overlooked dimension in language acquisition. *Modern Language Journal*, 75, 173-180.

Fulcher, G. &Davidson, F. (2007). *Language testing and assessment*. Routledge.

Genesee, F. (1976). The role of intelligence in second language learning. *Language Learning*, *26* (2), 267-280.

Gipps, C. & Murphy, P. (1994). *A fair test? Assessment, achievement and equity.* Buckingham Open University Press.

Godschalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability.* New York: College Entrance Examination Board.

Graham, S. (1997). *Effective language learning: positive strategies for advanced level language learning.* Multilingual Matters.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.) *Syntax and Semantics,* Volume 3 (pp. 41-58). New York: Academic Press.

Gu, Y. & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. *Language Learning, 46,* 643-679.

Gu, Y. (2002). Gender, academic major, and vocabulary learning. *A journal of Language Teaching and Research*, 33 (1), 35-54.

Harris, M. & McCann, P. (1994). *Assessment.* Oxford: Heinemann.

Halpern, D. F. (1992). *Sex differences in cognitive abilities.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Harmer, J. (2007). *How to teach English*. Essex: Pearson Longman.

Heaton, J.B. (1990). *Classroom testing.* New York: Longman.

Hilleson, M. (1996). I want to talk with them, but I don't want them to hear. In K. M. Bailey and D. Nunan (Eds.), *Voices from the Language Classroom* (pp. 248-277). Cambridge: Cambridge University Press.

Hingley, P. (1983). Modern languages. In J. Whyld (Ed.) *Sexism in the secondary curriculum* (pp. 99-110) London: Harper and Row.

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly: 40* (1), 109-131.

Howell-Richardson, C. & Parkinson, B. (1988). Learner diaries: possibilities and pitfalls. In P. Grunwell (ed.) *Applied Linguistics in Society*. Papers from the Annual Meeting of the British Association for Applied Linguistics. London: CILT/BAAL.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Jımenez Catalan, R. M. (2003). Sex differences in L2 vocabulary learning strategies. *International Journal of Applied Linguistics. 13* (1), 54-77.

Johnston, P.H. (2003). Assessment conversations. *The Reading Teacher, 57*, 90–92.

Jordan, R.R (1997). *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Ilyin, D. (1972). Ilyin oral interview. Rowley, Mass.: Newbury House. Ingram, D. E. 1984. Australian Second Language Proficiency Ratings.

Kitao, K.S., & Kitao, K. (1996). Testing speaking. (ERIC Document Reproduction Service No: ED 398261.

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.

Klann-Delius, G. (1981). Sex and language acquisition: Is there any influence? *Journal of Pragmatics*, *5,* 1-25.

Knudson, R. E. (1995). Writing experiences, attitudes, and achievement of first to sixth graders. *Journal of Educational Research, 89*, 90-97.

Krashen, S. (1988). *Second language acquisition and second language learning.* Englewood Cliffs, NJ: Prentice Hall.

Kuechler, W. L., & Simkin, M. G. (2004). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education, 14* (4), pp. 389-400.

Kunnan, A. J. (1998). Approaches to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 1-16). Mahwah, N.J.: LEA.

Larsen-Freeman, D., & Long. M. H. (1991). *An introduction to second language acquisition research.* New York: Longman.

Legato, M. J. (2005). *Why men never remember and women never forget.* New York: Rodale.

Lowe, P. Jr. (1982). ILR Handbook on Oral Interview Testing. Washington, DC: DLVLS Oral Interview Project.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple choice, constructed response, and examinee selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234–250.

MacIntyre, P. (1995). How does anxiety affect second language learning? A reply to Sparks and Ganschow. *Modern Language Journal*, *79* (1), 90-9.

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

Markham, P. L. (1988). Gender differences and the perceived expertness of the speaker as factors in ESL listening recall. *TESOL Quarterly, 22*, 297-406.

Martin, J. R. (1984). Language, register and genre. In F. Christie (ed.) *Language Studies: Children's Writing: Reader*. (21-29). Geelong: Deakin University Press.

McCarten, J. (2007). *Teaching vocabulary: Lessons from the corpus, lessons for the classroom.* Cambridge: Cambridge University Press.

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

Meier, S., McCarthy, P. R. & Schmeck, R.R. (1984). Validity of self-efficacy as a predictor of writing performance. *Cognitive Therapy and Research, 8*, 107-120.

Munby, J. (1978*). Communicative syllabus design.* Cambridge University Press.

Murphy, J. M. (1991). Oral communication in TESOL: Integrating speaking, listening,and pronunciation. *Tesol Quarterly*, *25* (1), 51-74.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nyikos, M. (1990). Gender-related differences in adult language learning: Socialization and memory factors. *Modern Language Journal, 74* (3), 273-287.

Nyikos, M. (2008). Gender in language learning. In C. Griffiths (Ed.), *Lessons from good language learners: Insights for teachers and learners* (pp.73-82). Cambridge University Press.

Obler, L. K. (1989). Exceptional second language learners. In: S. Gass, C. Madden, D. Preston & L. Selinker (eds.), *Variation in second language acquisition: Vol. II. Psycholinguistic Issues* (141-159). Clevedon: Multilingual Matters.

O'Dell, F. (1997). Incorporating vocabulary into the syllabus. In N. Schmitt & M. McCarthy (Eds.) *Vocabulary: Description, Acquisition and Pedagogy* (pp. 258-278). Cambridge: Cambridge University Press.

National Learning Infrastructure Initiative (NLII, 2004). Retrieved from http://www.elearnspace.org/Articles/eportfolios.htm

Oxford, R., Nyikos, M. & Ehrman, M. (1988) Vive la difference? Reflections on sex differences in use of language learning strategies. *Foreign Language Annals, 21,* 321-329.

Oxford, R., & Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *Modern Language Journal, 73* (3), 291-300.

Oxford, R. (1993). Research on second language learning strategies. *Annual Review of Applied Linguistics, 13,* 175-187.

Oxford, R. L. (1996). *Language learning strategies around the world: cross-cultural perspectives*. Natl Foreign Lg Resource Ctr.

Oxford, R. L., Lavine, R. Z., Hollaway, M. E., Felkins, G., & Saleh, A. (1996). Telling their stories: Language learners use diaries and recollective studies. In R. L. Oxford (Ed.), *Language learning strategies around the world: Cross-cultural perspectives* (pp. 19- 34). Manoa: University of Hawaii Press.

Pajares, F. & Valiente, G. (1996). *Predictive utility and causal influence of the writing self-efficacy beliefs of elementary students.* Paper presented at the annual meeting of the American Educational Research Association. New York.

Pajares, F.,& Valiante, G. (1997). The predictive and mediational role of the writing self-efficacy beliefs of upper elementary students. *Journal of Educational Research, 90,* 353–360.

Pajares, F., Miller, M. D., & Johnson, M. J. (1999). Gender differences in writing self-beliefs of elementary school students. *Educational Psychology, 91,* 50–61.

Pang, E. S., Muaka, A., Bernhardt, E. B., & Kamil, M. L. (2003). *Teaching reading.* Brussels, Belgium: International Academy of Education.

Politzer, R.L. (1983). An exploratory study of self-reported language learning behaviors and their relation to achievement. *Studies in Second Language Acquisition, 6,* 54-68.

Popham, W. J. (1978). *Criterion-referenced measurement.* NJ: Prentice-Hall.

Powell, R.C. (1979). Sex differences in language learning: a review of the evidence. *Audiovisual Language Journal, 17* (1), 19-24.

Powell, R.C. & Batters, J.D. (1985). Pupils perceptions of foreign language learning at 12+: some gender differences. *Educational Studies, 11* (1), 12-23.

Powell, R.C. & Littlewood (1983). Why choose French? Boys' and girls' attitudes at the option stage. *The British Journal of Foreign Language Teaching, 21* (1) 36-9.

Pritchard, R. M. O. (1987). Boys' and girls' attitudes towards French and German. *Educational Research, 29* (1), 65-72.

Purpura, J. E. (1995). *Fundamental considerations in the design of CB language tests*. Paper presented at the 1st INGED Conference, Middle East Technical University, Ankara, Turkey.

Richards, J. C (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*, 219-239.

Robinson, P. C. (1991). *ESP today: A practitioner's guide*. Hemel Hempstead: Phoenix ELT.

Ryan, J. M. & DeMark, S. (2002). Variation in achievement scores according to gender, item format and content area tested. In G. Tindal & T. M. Haladyna (Eds.) *Large Scale Assessment Programs for all students* (67-88). Routledge.

Rua P. L. (2006). The sex variable in foreign language learning: an integrative approach. *Porta Linguarum, 6,* 99-114.

Burstall, C. (1975). Factors affecting foreign-language learning: a consideration of some relevant research findings. *Language Teaching and Linguistics Abstracts, 8,* 105-125.

Ruth E. Knudson. (1995). Writing experiences, attitudes, and achievement of first to sixth graders. *The Journal of Educational Research, 89* (2), 90-9.

Shell, D.F. (1989). Self-efficacy and outcome expectancy mechanisms in reading and writing achievement. *Journal of Educational Psychology, 81*, 91-100.

Shohamy, E. (1983). Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew. In J. W. Oller (Ed.) *Issues in language testing research* (229-36). Newbury House.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Reading Research Quarterly, 17*, 229-255.

Shohamy, E. (2001*). The power of tests: A critical perspective on the uses of language tests*. London: Pearson.

Slavin, R.E. (1988). Cooperative learning and student achievement. In R.E. Slavin (Ed.) *School and Classroom Organization*. Hillsdale, NJ: Erlbaum.

Springer, S. & Deutsch, G. (1989). *Left brain, right brain*. New York: Freeman.

Sunderland, J. (1994). *Exploring gender: Questions and implications for English language Education.* Hemel Hempstead: Prentice Hall.

Sunderland, J. (2000). Issues of language and gender in second and foreign language education. *Language Teaching, 33*, 203-223.

Spolsky, B. (1990). *Conditions for second language learning.* Oxford: Oxford University Press.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Thornbury, S. (2002). *How to teach vocabulary*. Malaysia: Longman.

Thornbury, S. (2006). An A to Z of ELT. Oxford: Macmillan.

Tyre, P. (2005). Boys' Brains, Girls' Brains. *Newsweek,* CXLVI (12), 58.

Valette, R. M. (1977). *Modern Language Testing*. New York: Harcourt Brace Jovanovich,Inc.

Van Houtte, M. (2004). Why boys achieve less at school than girls: the difference between boys' and girls' academic culture. *Educational Studies, 30* (2), 159-173.

Walstad, W. B., & Becker, W. E. (1994). Achievement differences on multiple-choice and essay tests in Economics. *American Economic Review, 84,* 193-196.

Wilder, G. Z. & Powell, K. (1989*). Sex differences in test performance: A survey of the literature.* New York: College Board Publications.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics*. Honolulu, HI: University of Hawai'I Press.

Zhang, L. J. (2001). Exploring variability in language anxiety: Two groups of PRC students learning ESL in Singapore. *RELC Journal, 32* (1), 73-91.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. Journal *of Educational Psychology, 2* (1), 51-59.

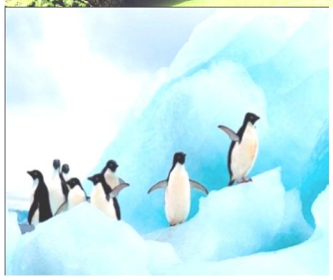
Appendix 1

The Achievement Test Administered to Pre-Intermediate Level Preparatory School Students at T.C. Kadir Has University

READING

THE BEST DOCUMENTARY SERIES OF BBC4



One of the most popular documentary channels is BBC4 of the United Kingdom. Since 2002, the channel has broadcast only documentaries and alternative news programmes. A wide variety of topics are covered in these programmes. The channel usually airs old and well-known documentaries. However, it also produces new shows every year. This season, BBC4 has four new documentaries on the most interesting and shocking sides of nature and the world around us.

Paragraph A: _____

Planet Earth is the most successful documentary television series that has ever been filmed. It was also the most expensive documentary of BBC done on nature. It gives the audience a chance to see the different characteristics of the Earth's nature. *Planet Earth* has eleven episodes and each episode is concerned with a different habitat on the Earth. The first episode features the frozen parts of our planet. It shows the audience rare animals that live in the polar regions. Another episode takes place in the jungles, that is, thick tropical forests with many large plants. There are many types of animals in the jungles and the episode includes lots of information about them.

Paragraph B: _____

Another important documentary on nature is *The Blue Planet*. It is known as the first documentary that gives a great amount of information about the oceans. The whole world is surrounded by water and the programme has eight sections that take place in a different type of water system. Furthermore, it gives valuable information about the coasts (places where the land meets the sea). The documentary also shows what happens when the Earth orbits the Sun, and how the Earth's circling around the Sun affects the oceans and the waters.

Paragraph C: _____

Nature's Great Events is a documentary series which is mostly about wildlife, or the animals and plants growing in their natural places. The programme looks at how the seasonal changes affect the nature. It opens with summer season and continues with the other seasons. In this documentary, we see, for example, what happens to animals when the temperature (hotness or coldness) rises and ice melts. The series includes the life of different fish, the oceanic life of South Africa and the eating habits of sea lions of Canada.

Paragraph D: _____

Life is also another documentary about nature. It especially focuses on how animals survive, or continue to live, in dangerous conditions. The show has ten episodes. The first episode is an introduction to the living habits of various animals and the second episode is about plants. The other episodes are all about animals of the hottest and the coldest places on the Earth.

A. Match the headings with the paragraphs and write the paragraph letters (A, B, C or D). (2 pts each)

1. Water Systems of the Earth — Paragraph _____
2. As Seasons Pass — Paragraph _____
3. Different Corners of the Earth — Paragraph _____
4. Survival Techniques — Paragraph _____

_____ / **8**

B. Answer the questions. (2 pts each)

1. What kind of programmes does BBC4 have? (Write 2 types)

   _____

2. Which documentary gives information about the North and South Poles?

   _____

3. In which documentary can a person learn more about the Earth's orbit and its effects?

   _____

4. Which documentary shows us what the sea lions eat?

   _____

5. Which documentary shows the way animals live in danger?

   _____

6. What is the third episode of *Life* about?

   _____

_____ / **12**

C. Look at the definitions below and find the words <u>in the text</u> using the *clues that signal definitions.*
Write the words in the blank. (1 pt each)

1. tropical forests with huge plants *(Par A)* _____
2. the land next to or close to the sea *(Par B)* _____
3. animals and plants that grow in natural conditions *(Pa* _____
4. a measure of how hot or how cold a place or thing is (_____
   C)
5. to continue to live after almost dying, or very difficult _____
   situations *(Par D)*

_____ / **5**

WRITING

Write a paragraph describing <u>one</u> of your <u>high school</u> teachers. (85 words minimum)

You can use the back of this sheet for brainstorming and/or drafting.

LISTENING

From a quiz show on TV:

Part 1

Host: Welcome to this week's quiz show! Our theme this week is "natural disasters" and our contestant James is here with us. Hello James, how're you doing?

James: Great, thank you.

Host: So if you're ready James, let's start. Here's your first question about world's natural disasters: Where in the world do most earthquakes occur? The Pacific Ocean, Anatolia, or Africa?

James: Hmm, Anatolia?

Host: No, actually not. Around 80 percent of the planet's earthquakes happen in countries on the coast of the Pacific Ocean. The area is also called the "Ring of Fire"… Ok, next question: How often does lightning strike Earth's surface? About 10, 100, or 1000 times per second?

James: I'll say... 100 per second?

Host: Well done! That is the correct answer! There's always lightning in some part of the world. And lightning is also extremely hot. In fact it writes here that a flash of lightning can heat the air around it to temperatures five times hotter than the sun's surface… Alright then, your next question: Flows of lava can burn everything in its path, including whole towns. Which is a town destroyed by lava flows? Hinode in Japan, Kalapana in Hawaii or Kalamos in Greece?

James: Ok, I know this one. I remember reading about a volcano that erupted there in 1990. It must be Kalapana in Hawaii.

Host: Kalapana ***is*** the correct answer. And how hot can lava be? Lava can reach a temperature of 1,250 degrees Celsius, and it can sometimes destroy a whole town. …

Alright, we'll meet James again in the second part of the show. Now a short commercial break…

Part 2

Host: We're back at the studio, ladies and gentlemen, and James is still with us. Ready to continue, James?

James: Absolutely!

Host: Let's start then. On March 11th 2011, an earthquake occurred on the Eastern coast of Japan. What was the magnitude? 7.5, 9.2, or 8.9?

James: 7.5?

Host: Unfortunately, it was a lot bigger than that. It measured 8.9 on the Richter scale. It was a disastrous earthquake that destroyed many things. But the worst damage was to nuclear power stations. It caused accidents at four major stations… And your last question, James: Which natural disaster ***cannot*** be a result of tectonic movements? A volcanic eruption, a tsunami or a hurricane?

James: A hurricane, of course!

Host: Correct! Hurricanes are ***not*** caused by tectonic movements. Actually, they are giant storms. They bring a lot of rain. And they can pack wind speeds of over 257 kilometers an hour. … You've done great, James.

James: Thank you!

Host: Hope to see you in the finals.

LISTENING

Read questions 1-5 below in 1 minute. Then listen to the <u>first part</u> of the quiz show. Choose the best option according to the information you hear. (2 pts each)

1. Most earthquakes happen in countries _____

a) on the Pacific coast.

b) in Africa.

c) around Anatolia.

2. Lightning on Earth happens _____

a) 100 times per second.

b) 1000 times per second.

c) 10 times per second.

3. Lightning _____

a) never heats the air.

b) can be 5 times brighter than the sun.

c) is very hot.

4. Flows of lava burned a town in _____

a) Japan.

b) Hawaii.

c) Greece.

5. Lava can be as hot as _____

a) 1050 ^{o}C.

b) 1250 ^{o}C.

c) 1520 ^{o}C.

_____ **/ 10**

B. Read the sentences below in 1 minute. Listen to the second part of the quiz show to answers questions 6-10. Write ONE WORD in each blank, OR if the information is numerical, write it in NUMBERS.
(2 pts each)

Eastern Japan Earthquake

The 2011 earthquake on the Eastern coast of Japan measured **6** _____ on the Richter scale. As a result, there were accidents at **7** _____ big nuclear power stations.

Hurricanes

Hurricanes are not about the movement of the Earth because they are **8** _____ **9** _____. They can reach speeds of over **10** _____ km per hour.

_____ **/ 10**

USE OF ENGLISH

A. Complete these sentences with a/an, the or no article (write Ø). (1 pt each)

Virginia Woolf was [1] _____ important English writer. Her family was full of artists. Her mother and sister died when she was only thirteen so she had

psychological problems. Virginia met her husband, Leonard Woolf, in London. He was [2] _____ successful publisher. In fact, he was one of [3] _____ most successful publishers in London. Later, he started to publish Virginia's novels. In the 19th century, [4] _____ women writers weren't very popular, but [5] _____ people liked reading Virginia's novels. *Mrs. Dalloway* and *To the Lighthouse* were two of her most important books.

_____ / **5**

B. Complete the dialogue with words from the box. (0.5 pts each)

so do I	**neither do I**	**why don't we**	**I don't**
should x 2	**what about**	**shouldn't**	**let's x 2**

Two students are talking about their project:

Susan: So, we have to finish our project by the end of this month, and we haven't chosen our topic yet!
Matt: We have only three weeks. What [1] _____ we do first, then?
Susan: [2] _____ look at the list of topics that the teacher gave us?
Matt: Sure, [3] _____ see. I think climate change is an excellent topic.
Susan: Do you? [4] _____. I think it's too popular - everybody is writing about it. I want to do something more original.
Matt: [5] _____. The list is very long. There is another topic, future earthquakes, for example. What do you think about it?
Susan: I think it will take a lot of time, and it is a very difficult subject to study. I don't think we [6] _____ choose that one.
Matt: [7] _____. We [8] _____choose a difficult topic, because we don't have much time.
Susan: You're right, I suppose. [9] _____ animals in the North Pole?
Matt: That is a great topic! We can write about their lives and the dangers of living in very cold conditions.
Susan: I really like this topic. [10] _____ start working!

_____ / **5**

C. Read the passage below. Complete it with the <u>past simple</u> or <u>past continuous</u> of the verbs in parentheses. (1 pt each)

Sir Isaac Newton and the "apple story":

One day, while Newton and one of his friends [1] _____ (have) tea in the garden, an apple [2] _____ (fall) on Newton's head from the tree above them. The apple [3] _____ (still / roll) on the ground when Newton
[4] _____ (start) thinking about gravity. Why [5] _____ (the apple / fall) towards the centre of the earth?

_____ / 5

D. Complete the news report with the <u>present simple</u>, <u>present continuous</u> or <u>present perfect</u> of the verbs in parentheses. (1 pt each)

From a news report about the earthquake in Van:

One month [1] _____ (pass) since 23 October 2011. Currently, thousands of earthquake survivors [2] _____ (suffer) from the cold and snow because they have to live in tents. So far, Turkish Red Crescent and other organisations
[3] _____ (send) thousands of tents and tons of food supplies to the area. However, some local people say their children [4] _____ (not eat) much proper food for two weeks. We [5] _____ (believe) this city is still in need. Therefore, as the national and international news channels here, we will continue to be the voice of Van.

_____ / 5

VOCABULARY

A. Fill in the gaps with a suitable word from the box. There is one extra word. (1 pt each)

dedicated	**drought**	**malnutrition**	**insurance**	**diseases**	**injuries**

Global warming has affected Kenya extremely. Now the summers are very dry in most cities. The lack of rain causes heavy [1] _____. There isn't enough water, so local people have a lot of [2] _____, such as cholera, diphtheria and hyperthermia. Lack of water also causes lack of food. People can't find enough food, so they suffer from [3] _____. This year lots of [4] _____doctors and nurses are trying to help the people of Kenya - they are working day and night. People are very poor, and they don't have [5] _____, so the doctors are doing surgery free of charge. There are also various charity organizations which are actively working in the country, such as United Nations.

_____ / **5**

B. Fill in the gaps with a suitable word/phrase from the box. There is one extra word.
(1 pt each)

depends on	**surface**	**made up of**	**continents**	**floods**
solar				

Planet Earth is [1] _____ three layers. The outer layer is called the crust. A large part of the crust is below the oceans. In fact, 71% of Earth's [2] _____ is covered with water. That is why it is also called the Blue Planet. There are [3] _____ surrounded by oceans on the Blue Planet. The biggest one is Asia. Ours is the only planet in the [4] _____ system

that has life. Life exists on Earth because there is water on it. Therefore, we can say that life on Earth [5] _____ water.

_____ / **5**