REPUBLIC OF TURKEY
ÇANAKKALE ONSEKİZ MART UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

ETHICAL CONCERNS IN LANGUAGE TESTING AT UNIVERSITY LEVEL

MA THESIS

Supervisor
Prof. Dr. Dinçay KÖKSAL

Submitted by
Anıl CEYLAN

Çanakkale-2007

Sosyal Bilimler Enstitüsü Müdürlüğü’ne

Anıl CEYLAN’ a ait

‘*Ethical concerns in language testing at university level*’ adlı çalışma, jurimiz tarafından Yabancı Diller Eğitimi Anabilim Dalı İngilizce Öğretmenliği Programında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Başkan Prof. Dr. Dinçay KÖKSAL
Akademik Ünvanı, Adı Soyadı (Danışman)

Üye Yrd. Doç Dr. Ece ZEHİR TOPKAYA
Akademik Ünvanı, Adı Soyadı

Üye Yrd. Doç. Dr. Aysun YAVUZ
Akademik Ünvanı, Adı Soyadı

Üye Yrd. Doç. Dr. Cevdet YILMAZ
Akademik Ünvanı, Adı Soyadı

Üye Yrd. Doç. Dr. Çavuş ŞAHİN
Akademik Ünvanı, Adı Soyadı

Çanakkale 2007

ÖZET

Bu çalışma İngiliz Dili Eğitimi Bölümünde görev yapan öğretim elemanlarının dil sınavlarındaki etik kavramı hakkındaki görüş ve düşüncelerini öğrenmek amacıyla yapılmıştır. Bu çalışma aynı zamanda, Türk Eğitim Sisteminde dil sınavlarının uygulanmasında bir 'etik yasası' veya bir 'uygulama yasası' geliştirmek için öğretim elemanlarının görüş ve düşüncelerinden faydalanmayı amaçlamaktadır.

Bu çalışma Çanakkale Onsekiz Mart Üniversitesi'nde İngiliz Dili Eğitimi Bölümünde görev yapan öğretim elemanlarının katılımıyla yapılmıştır. Veriler nitel veri toplama metotlarından olan anket yöntemiyle elde edilmiştir. Anket, Çanakkale Onsekiz Mart Üniversitesinde görev yapan 28 öğretim elemanına uygulanmıştır. Veriler bilgisayarda SPSS (Statistical Package for the Social Sciences) adlı programla analiz edilmiştir.

Anket kullanılarak elde edilen bulgular öğretim elemanlarının çoğunun Türkiyede dil sınavlarının uygulanması hakkında bir 'etik yasası' nın mevcut olup olmadığı hakkında bilgi sahibi olmadığı göstermiştir. Bununla birlikte, öğretim elemanları dil sınavlarının uygulanmasında bir 'etik yasası' nın gerekli olduğuna inanmaktadır. Bunun yanısıra, öğretim elemanları geliştirilecek 'etik yasası'nın Türk toplumunun değer sistemlerinin göz önüne alınarak yapılması gerektiğine inanmaktadırlar.

ABSTRACT

The present study investigated the perceptions of ELT Department instructors on ethics concept in language testing. The study further aimed at collecting information in order to develop a ‘code of ethics’ or ‘code of practice’ in language testing for Turkish Education System.

The study was carried out with the instructors of ELT Department at Çanakkale Onsekiz Mart University. The data were collected via qualitative research methodology; the questionnaire was administered to 28 instructors at Çanakkale Onsekiz Mart University. The data were analysed and interpreted with the help of computer program SPSS (Statistical Package for the Social Sciences).

The findings obtained through the questionnaire revealed that most of the instructors do not know whether there is a ‘code of ethics’ for language testing in Turkey or not. However, they believe that a ‘code of ethics’ for language testing is necessary. Furthermore, they indicated that ‘a code of ethics’ should be developed considering the value systems of the Turkish culture.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

CHAPTER TWO
TESTING IN LANGUAGE TEACHING

CHAPTER THREE
ETHICS IN LANGUAGE TESTING

CHAPTER FOUR
METHODOLOGY

CHAPTER FIVE
FINDINGS

CHAPTER SIX
DISCUSSIONS, CONCLUSIONS AND IMPLICATIONS

ABBREVIATIONS

CTS	Classical True Score
G-theory	Generalizability Theory
IRT	Item Response Theory
RQ	Research Question
SPSS	Statistical Package for Social Sciences

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First of all I would like to express my gratitude to my supervisor Prof. Dr. Dinçay Köksal for his constant guidance, support, and encouragement throughout the different stages of my thesis. I would like to thank Assistant Prof. Dr. İsmail Hakkı Erten and Assistant Prof. Dr. Ece Zehir Topkaya and Assistant Prof. Dr. Aysun Yavuz for their encouragement. I am also grateful to them for the knowledge and skills I acquired during their master courses.

I am grateful to Assistant Prof. Dr. Çavuş Şahin for helping me to analyze all the data obtained from the questionnaires. I appreciate him for his patience and sharing his knowledge with me.

I would also like to thank Associate Prof. Dr. Salih Uşun for his suggestions in the preparation and the analyzing stages of the questionnaire.

I would also like to thank all instructors who participated in the study.

To

“My family”

CHAPTER ONE
INTRODUCTION

1.0 INTRODUCTION

This chapter starts with the brief description of the background to the study and continues with the purpose of the study and the research questions. The significance of the study, its assumptions and limitations are also stated. Finally, the organisation of the thesis is outlined.

1.1 BACKGROUND OF THE STUDY

Tests are used for many different types of educational purposes such as selection, placement, diognosis and evaluation. “There are two major uses of language tests: First, they are used as sources of information for making decisions within the context of educational programs and second as indicators of abilities or attributes that are of interest in research on language, language acquisition, and language teaching” (Bachman 1990: 54).

However, Schemeiser (1995: 1) points out that “People who are involved in test preperation and application processes have ethical responsibilities to the test takers and to the public such as integrity, honesty, confidentiality, objectivity”. Unfortunately, unethical test practices are still common literally in every country in the world.

“The ethics and fairness concerns are not new in language testing because these issues have been taken into consideration in the framework of reliability and validity” (Alderson 1997; Shohamy 1997b in Kunnan 1999: 4).

However, in the last two decades language testing researchers have begun to focus on issues such as test standards, test bias, equity and ethics for testing professionals besides reliability and validity issues.

Furthermore, no study was carried out regarding the 'ethics in language testing' in Turkey.

1.2 PURPOSE OF THE STUDY AND THE RESEARCH QUESTIONS

The concept of ethics covers a vast area ranging from the medicine to law and from religion to education. Therefore, it can be concluded that the effects of 'ethics' can be seen literally in every field in our life and its consequences are of vital importance as well. Ethics in language testing field is one of the areas that should be dealt in education. Since, it affects all the stakeholders more or less, that is, the overall language testing field.

Therefore, this study aims to determine the instructors' conception of ethics, and what is considered ethical and unethical in language testing. Furthermore, it also investigates the roles of stakeholders in ethical test use and the importance of a 'code of ethics' in language testing. Therefore, this study will help to develop a 'code of ethics in language testing'. Moreover, instructors' views regarding the ethical issues will be identified. Finally, this study will explore how language tests can be carried out ethically in the classroom and standards for language test use at university level.

In conclusion, the study aims to answer the following research questions:

RQ 1: What is the test developers' and test users' level of making ethical choices in using English language tests?

RQ 2: What is the stakeholders' level of making ethical choices in using English language tests?

RQ 3: What is the stakeholders' level of responsibility in ethical use of English language tests?

RQ 4: 'Confidentiality' and 'access to information' are considered among the fundemental rights of test takers. What is the level of contribution of these rights to ethical use of English language tests?

RQ 5: What is the level of acceptance of language test givers to using language tests for non-intended purposes?

RQ 6: What is the level of awareness of language test givers to availability of a 'code of ethics' for language test use?

RQ 7: How much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests?

RQ 8: What is the level of acceptance of language test givers to unethical behaviors in language testing?

RQ 9: What do the test givers think about the content of a 'code of ethics' for language test use?

RQ 10: What do the test givers consider the unethical behaviors of a test giver?

RQ 11: What do the test givers consider the unethical behaviors of a test taker?

RQ 12: What do the test givers consider the qualities of a test giver?

RQ 13: What is the test givers level of considering value systems of the society in ethical use of English language tests?

RQ 14: Is there a significant diference between the views of the instructors and the members of the faculty holding PhD Degree with regard to;

a: What is the test developers' and test users' level of making ethical choices in using English language tests?

b: What is the stakeholders' level of making ethical choices in using English language tests?

c: What is the stakeholders' level of responsibility in ethical use of English language tests?

d: 'Confidentiality' and 'access to information' are considered among the fundemental rights of test takers. What is the level of contribution of these rights to ethical use of English language tests?

e: What is the level of acceptance of language test givers to using language tests for non-intended purposes?

f: How much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests?

g: How much do the test givers believe in the reasons for creating such a code for ethical use of language tests?

h: What is the level of acceptance of language test givers to unethical behaviors in language testing?

i: What do the test givers think about the content of a ‘code of ethics’ for language test use?

j: What do the test givers consider the unethical behaviors of a test giver?

k: What do the test givers consider the unethical behaviors of a test taker?

l: What do the test givers consider the qualities of a test giver?

m: What is the test givers level of considering value systems of the society in ethical use of English language tests?

1.3 SIGNIFICANCE OF THE STUDY

'Ethics in Language Testing' issue has not got the attention it deserved although it is one of the most important elements in language testing process. The first and only big event was the 19th Language Testing Research Colloquium in 1997 which discussed 'Fairness in Language Testing'. Furthermore, the issue 14, 3, 1997 of Language Testing Journal was wholly dedicated to 'ethics in language testing' (Fulcher 1999).

It is clear that only recently the significance of 'ethics in language testing' has been understood. For example, some countries such as the USA has developed a code named 'Code of Fair Testing Practices in Education'. However, 'ethics in language testing' issue has been ignored in Turkey as well as in many other countries.

Therefore, the first contribution of this study will be to help test developers and test users to carry out appropriate and ethical tests in the field of foreign language teaching and learning in Turkey. Secondly, this study will help to determine the standards of appropriate and ethical test use for foreign language education in Turkey and develop a code of ethics for language testing in education.

1.4 LIMITATIONS OF THE STUDY

The study is limited to the ELT instructors of Çanakkale Onsekiz Mart University and will be carried out during 2006-2007 academic year. For this reason, it may not be possible to generalize the results of this study for all ELT instructors in Turkey.

In this study, ‘questionnaire’ is used as data collecting instrument. Some of the questions were developed by the researcher and some of them were adapted from a different questionnaire carried out to collect data from Polish instructors to form a ‘code of ethics’. Therefore, the results of the study are limited to these instruments.

1.5 ASSUMPTIONS OF THE STUDY

Ethics forms an essential part of not only the society itself but also the education. Moreover, all stakeholders involve the education process. However, the instructors are considered as the primarily responsible people in ethical use of tests in education. Therefore, all the participants are assumed to contribute to the study willingly.

There will be no significant difference between the views of instructors and assistant professors, since it is assumed that holding a PhD degree will not make any difference in their views regarding ethical test use.

It is assumed that all the participants will answer the questionnaire honestly which will be used to collect data.

1.6 ORGANIZATION OF THE THESIS

This thesis is composed of six chapters. The background of the study is presented in Chapter One. Then, the purpose of the study and the research questions are stated. The significance, assumptions and limitations of the study are also included in Chapter One. Finally, Chapter One ends with the description of the organization of the thesis.

Chapter Two reviews the literature on testing, measurement and evaluation. The types and uses of tests and basic qualities of a test (reliability and validity) are discussed in detail in this chapter.

Chapter Three discusses the ‘ethics’ concept in detail. Approaches to language testing ethics, inappropriate uses of language tests, roles of stakeholders in language test use, models developed for ethical test use and language test standards are taken into consideration in this chapter.

Chapter Four reports the methodology of the study. The rationale for the questionnaire study is stated in this chapter. The participants, the setting and the procedures to the pilot and main study are described in Chapter Four.

Chapter Five states the findings of the study in detail to provide answers to the research questions.

Chapter Six discusses the findings of questionnaire and aims at drawing conclusions through analyses of the findings of the study. Furthermore, implications and suggestions for further research are stated in this chapter.

1.7 CHAPTER SUMMARY

This chapter discussed the basic literature regarding language testing and ethics concept. The purpose of the study was stated. The significance, the assumptions and limitations of the study were also discussed. Finally, the organization of the thesis was presented.

CHAPTER TWO
TESTING IN LANGUAGE TEACHING

2.0 INTRODUCTION

This chapter aims to describe the basic concepts and procedures in language testing. First, measurement, test, evaluation and assessment concepts will be defined. Second, uses of language tests in educational programs will be described. Third, the types of language tests will be explained. Finally, the basic qualities of a test will be discussed in this chapter.

2.1 BASIC TERMS: MEASUREMENT, TEST, EVALUATION, ASSESSMENT

The terms 'measurement', 'test', 'evaluation' and 'assessment' are the common terms which are used in the language testing field. These terms are used interchangeably not only in different published materials but also in lectures and discussions. For this reason, it will be very useful to identify the differences between them and explain them thoroughly in order to develop and use language tests efficiently and appropriately.

2.1.1 MEASUREMENT

The term 'measurement' has been defined by a number of language testing specialists. For instance, Genesee et al. (1996: 145) define measurement as "the assignment of numbers to qualities or characteristics according to some standard and rational method and the interpretation of those scores with respect to some frame of reference". Furthermore, they point out that there are two types of measurement tools. While the first one is used to measure differences in kind, the other is designed to measure differences in degree.

McNamara (2000) argues that there are two basic steps in measurement process: Quantification and procedures to provide validity of the test. Similarly, Bachman (1990: 19) claims that 'measurement' has three main qualities. These are "quantification, characteristics, rules and procedures."

First, quantification refers to labelling and ranking the characteristics of people with numbers (Bachman 1990). Therefore, quantification serves to analyze and interpret the characteristics of the test takers (Bachman 1990; McNamara 2000).

Second, the test takers' characteristics are measured in the testing process. However, it should be taken into account that a test taker has both physical characteristics and mental characteristics. While physical characteristics are measured directly, mental characteristics can be measured indirectly (Bachman 1990).

Third, there should be procedures and rules to provide the reliability and the validity of the test and to obtain meaningful results about the characteristics of the test takers (Bachman 1990; McNamara 2000).

Finally, all these qualities facilitate the measurement process. Because, test taker characteristics are varied and many different ways of measurement methods should be implemented to obtain valid results.

2.1.2 TEST

Test is undoubtedly a method which is designed to gather information about a person's characteristics. These characteristics may refer to a person's special ability, knowledge as well as a certain type of behavior (Bachman 1990; Brown 2001). Genesee et al. (1996: 141) come up with a more detailed description of a 'test':

1. A method for collecting information.
2. Has subject matter and content
3. A task or set of tasks that elicits observable behavior from testee.
4. Yield scores presenting attributes or characteristics of individuals.

Therefore, many types of tests have been designed which serve to collect information about the characteristics and measure the knowledge of people. Because, while one test aims to measure one ability of a person, it may fail to measure the other abilities of himself or herself.

2.1.3 EVALUATION

The concepts 'measurement' and 'evaluation' should not be confused. For this reason, the difference between these terms should be clearly defined. The main purpose of measurement is to obtain information regarding the characteristics of a test taker. However, evaluation serves to interpret the data provided through measurement in order to make decisions about the test takers. Furthermore, making correct decisions is of vital importance in the testing process. It depends on two factors. First, the decision maker should have the qualities to make the correct decision about the test takers. Second, the information obtained from the test takers should have quality (Bachman 1990).

2.1.4 ASSESSMENT

Assessment concept is widely used as the other terms stated above. Airasian (1991 in Bell 1994: 3) defines assessment as “the process of collecting, synthesizing and interpreting information for use in decision making”.

Vansickle (2003: 4-5) states that “assessment is typically the larger umbrella under which judgments, actions or decisions are made based on the tests and measurements used in a given situation. Therefore, assessment includes testing and measurement, and in many contexts is used in place of either or both terms”.

There are two types of assessment. Large-scale assessment and classroom assessment. Large-scale assessments are used in schools located in the different districts of a country. However, classroom assessment is carried out in a classroom and administered by the teacher. Furthermore, there are different types of assessment methods. Traditional assessments and performance assessments. The former one makes the student select the right answer. For example, multiple-choice, true-false and matching tests. These are also called traditional paper-and-pencil tests. On the other hand, students should produce an appropriate answer in performance assessments. Essay exams, presentations and projects can be counted among performance assessments (Bell 1994).

The term ‘test’ is used the most in the language testing field. For this reason, it will be preferred in the following sections. Moreover, specific test types have been developed to obtain and measure different qualities of a test taker. Uses of tests in educational settings, approaches to language testing and classification of tests will be dealt in detail in the following parts.

2.2 USES OF LANGUAGE TESTS IN EDUCATIONAL PROGRAMS

Tests serve for many different purposes. However, language tests are primarily used in educational settings. For this reason, the term ‘educational program’ should be defined. Bachman (1990: 54) points out that “when one or more than one person participate in teaching and learning, this situation is called an educational program.”

The basic purpose of a test is to evaluate the student’s performance in order to compare and select students. External examinations are designed for selection purposes. Moreover, classroom tests provide the necessary information for the teacher to become more effective. For instance, the teacher may change his/her teaching styles to enable the students to learn more. Another function of the test is determining the students’ weaknesses. When the teacher has identified the areas that students have difficulty, he cannot only make modifications in the syllabus, but also improve methods and materials he is employing in the classroom. Finally, a well-designed classrom test enables students to show their performance on specific tasks in the language. The feedback helps the students to identify their weaknesses and learn from them (Heaton 1988).

Tests also serve for evaluation and decision making regarding individuals (micro-evaluation) and educational programs (macro-evaluation). This can be carried out through using test results. The decision makers are test takers, teachers, administrators, district school boards, employers and state boards of education. The most important of these decision makers are test takers, teachers and programs. Decisions about students and teachers refer to decisions about individuals. The selection and placement of students for appropriate programs, diagnosing the students’ strengths and weaknesses and getting feedback about the students’ progress and grading them according to their performance constitute the decisions regarding students. It is necessary to make decisions about teachers in educational programs as well. Since it is important for teachers to improve themselves continuously. For instance, teachers’ language proficiency can be determined through proficiency tests so that they can identify their lacks and become more successful teachers. Language tests can also be used in making decisions about programs. The parts of a program can be evaluated and improved by making use of tests when developing new programs (Bachman et al. 1996).

Davies (1990) claims that language tests serve to gather information and there are five ways to obtain information. First, they are used in research and experiments to test hypotheses regarding language learning. Second, they are used to draw conclusions about the syllabus and the teaching process. Third, they serve for assessing learners. Fourth, they provide information to select students. Finally, they are used to evaluate courses, materials and methods in language teaching and learning.

Finally, language tests play an important role in research as well. Language proficiency, language acquisition and language teaching can be counted among reserach areas in language testing(Bachman 1990).

2.3 APPROACHES TO LANGUAGE TESTING.

Four main approaches to language testing were developed so far: The essay translation approach, the structuralist approach, the integrative approach and the communicative approach. A well-designed test involves the qualities of some of these approaches. The essay translation approach refers to the pre-scientific stage of language testing. The subjective judgment of the teacher is of vital importance and special skills are not necessary. The structuralist approach is based on the view that language learning mainly focuses on the learning of habits in a systematic way. Identifying and measuring the learner's understanding of the different parts of the target language such as vocabulary and grammar are of vital importance to this approach. The integrative approach is about the testing of language in context. These tests are designed to assess the learner's ability to use two or more skills simultaneously. Finally, the communicative approach refers to the use of language in communication. Therefore, it aims to apply tasks to the real life situations as much as possible (Heaton 1988).

There are many different types of tests. Test users and developers have difficulty in deciding the type of test to use in different situations. Because, selecting the appropriate type of test is of vital importance in order to collect both reliable and valid results according to the situation. For this reason, tests were classified according to their purposes.

2.4 CLASSIFICATION OF LANGUAGE TESTS

The new Standards for Educational and Psychological Testing (AERA, APA, & NCME 1999 in Vansickle 2003: 4) defines test as “all evaluative devices such as inventories [and] scales.” Therefore, test are primarily used to make judgements. However, tests can be classified in different categories changing from the information type the tests provide to the analyses of the collected information.

There are four main categories of language tests according to the information they provide: Achievement tests, proficiency tests, aptitude tests and diognostic tests. Achievement tests are designed to measure what the students have learnt in a language course. School examinations and public tests are common examples to these tests and they are based on course syllabus. Proficiency tests, on the other hand, are not related to any syllabus. These tests are developed to measure a student’s language proficiency according to a certain task. Aptitude tests are constructed to measure a student’s strengths and weaknesses in a foreign language before he has started to learn. Language learning aptitude is related to several factors such as intelligence, age, motivation, memory and etc. These characteristics differ from one individual to another. Finally, diagnostic tests are intended to determine the weaknesses and strengths of a test taker’s various language abilities (Heaton 1988).

Language tests are also classified according to scoring procedure and testing method. Subjective and objective tests are examples to language tests according to scoring procedure. In subjective tests, the test developers decide subjectively how a test can be constructed and similarly test takers answer the test questions in the same manner. However, in objective tests the test taker's response is analyzed according to the scoring criteria which were determined in advance. Multiple-choice test technique is the most widely used example to objective tests. Test constructors still continue to develop new methods for language tests. Therefore, there are a lot of test methods. Performance test is the most common testing method. Essay and oral interview are examples to these tests. Moreover, multiple-choice, completion, dictation and close tests can also be counted among the testing methods (Bachman 1990).

However, tests cannot be limited to these categories stated above. The other catogories that should be taken into consideration are: Direct-indirect testing, discrete point-integrative testing, objective-subjective testing, computer adaptive testing communicative language testing and norm-referenced/criterion referenced testing. Direct testing aims to measure a certain skill of a test taker. For example, a speaking test is carried out to determine the pronunciation skill of a learner. However, the authenticity concept is very important for such exams. Indirect testing, on the other hand, tries to measure the abilities of the testee that the teacher is interested in. Discrete point tests aim to measure one quality at a time, item by item. In contrast, in integrative testing the testee has to combine a number of language qualities in a task. e.g. completing a close passage, taking notes while listening a lecture. Computer adaptive testing enables to collect information on people's language abilities efficiently. Finally, communicative language tests are intended to measure how the individuals are able to use language in real life situations. (Hughes 2002).

Language test results can also be interpreted by employing norm-referenced and criterion referenced tests (Bachman 1990). Vansickle (2003: 8) explains these two types of tests as follows:

> "Norm-referenced tests report scores or profiles based on reference to a standard group. In these tests, a normative sample of individuals is used to determine the distributional characteristics of the responses for that group (e.g., mean and standard deviation). The test is scaled so that various scores can be reported to test takers based on the typical response patterns of the standardization group. The score or scores a test taker receives reflect the person's performance compared to the normative sample.
> In criterion-referenced tests, an individual's responses are compared to some predetermined standard (i.e., criterion). The standard may be a cut-off score expressed as a raw score, a percentage, a standard score, or some other value. If the test taker reaches or exceeds the specified standard or criterion, he or she is classified as having learned the material, achieved a specific level of mastery, or falling into some group or category."

Test developers have constructed many types of tests to measure the particular qualities of test takers. However, a test should have some specific qualities in order to provide accurate and meaningful results.

2.5 QUALITIES OF A LANGUAGE TEST

Bachman et al. (1996) state that a test should be used for the intended purposes. Therefore, test developers and test users should attach importance to use of test for the intended purpose in test design and development process. Furthermore, reliability and validity concepts are considered as the basic qualities of measurement (Heaton 1990; Bachman 1990; Davies 1990; Alderson et al. 1995). These two fundemental concepts of language testing will be explained in this section.

2.5.1 RELIABILITY

Reliability is one of the essential characteristics of a good test. A test cannot be considered valid unless it is reliable. These two qualities (reliability and validity) are of vital importance not only interpreting the test scores appropriately but also constructing and using tests in education (Bachman 1990; Heaton 1988; Alderson et al. 1995).

Consistency is the primary concern of the reliability concept (Bachman 1990; Davies 1990; Alderson et al. 1995). That is the consistency of test scores and results. In other words, since reliability is a quality of test scores there should be no measurement errors in a reliable test.

Alderson et al. (1995: 87) argue that "The aim in testing is to produce tests which measure systematic rather than unsystematic changes, and the higher the proportion of systematic variation in the test score, the more reliable the test is. A perfectly reliable test would measure only systematic changes". Therefore, the causes of unsystematic variation should be reduced to minimum in order to develop reliable tests.

However, it is impossible to achieve a perfectly reliable test. Since there are a lot of factors that affect the reliability of test scores and decisions based on these scores. For this reason, it is the test constructors' responsibility to make the tests as reliable as possible.

2.5.1.1 FACTORS AFFECTING LANGUAGE TEST SCORES

Measurement specialists claim that “the examination of reliability depends on our ability to distinguish the effects of the abilities we want to measure from the effects of other factors on test scores” (Stanley 1971: 362 in Bachman 1990: 163). Therefore, the abilities to be measured and the factors that are likely to affect test scores should be defined in order to provide the reliability of the test scores.

Heaton (1988: 162-163) enlists the factors affecting the reliability of the test as follows:

> "1. The extent of the sample of material selected for testing: whereas validity is concerned chiefly with the content of the sample, reliability is concerned with the size. The larger the sample, the greater the probability that the test as a whole is reliable.
> 2. The administration of the test: is the same test administered to different groups under different conditions or at different times?
> 3. Test instructions: are the various tasks expected from the testees made clear to all candidates in the rubrics?
> 4. Personal factors such as motivation and illness.
> 5. Scoring the test: objective tests overcome the problem of marker reliability, but subjective tests are sometimes faced with it: hence the importance of the work carried out in the fields of the multiple-marking of compositions and in the use of rating scales."

However, the factors affecting language test scores are varied and cannot be limited to those mentioned above. Bachman (1990: 164) argues that "Communicative language ability, test method facets, attributes of the test taker that are not considered part of the language abilities to be measured are the factors affecting language test scores".

First of these factors is the 'communicative language ability' which refers to student's achievement in different areas of communication. For example, listening comprehension, speaking and listening, reading and writing. This led to the development of communicative language testing in language testing field (Heaton 1988). Second of these factors is the test method facet. According to Bachman (1990: 118) "test method facet has five main elements. These are, the testing environment, the test rubric, the nature of the input the test taker receives, the nature of the expected response to that input, and the relationship between input and response". Third of these factors is the characteristics of individuals such as cognitive style, sex, race and ethnic background (Bachman 1990).

In order to observe and interpret the test taker's actual performance on a given test, theories were developed to measure the reliability as well as validity of a test.

2.5.1.2 WAYS OF MEASURING THE RELIABILITY OF A TEST

There are various ways to measure the reliability of a test. Three fundemental theories were developed in order to estimate the reliability of a test: The classical true score measurement theory, generalizability theory and item response theory.

2.5.1.2.1 CLASSICAL TRUE SCORE MEASUREMENT THEORY

Bachman (1990) points out that it is really difficult to measure the language abilities of an individual since they are not concrete. For this reason, a test should be designed that measures the specific ability of the test taker. The score obtained from this test is used as the true score of this specific ability. However, according to Classical True Score (CTS) Measurement Theory the relationships between the observed test scores and factors affecting these scores should be taken into consideration. Therefore, this theory asserts that, first, the observed score involves both the true score which stems from the test taker’s ability and the error score which is due to the unrelated factors to the ability being tested (True Score Variance). Second, since error scores are random it leads to the the differences in test scores (Measurement Error). As a result, these two assumptions of CTS Measurement Theory help to identify true scores and error scores. Moreover, three approaches were developed in CTS model to estimate reliability: Internal consistency estimates, equivalence estimates and stability estimates.

2.5.1.2.1.1 INTERNAL CONSISTENCY

Although the test takers are expected to perform consistently when answering the test questions they may perform differently on different parts of the test. The test method facets are considered as the main causes of the varying performances of the test takers throughout the test. The test takers consistency of performance in the whole test is called ‘internal consistency’ (Bachman 1990). Alderson et al. (1995) call the internal consistency as inter-item consistency.

Internal consistency can be estimated through split-half method. In split-half method the test is divided into two equal parts and these two parts are considered as parallel versions. The measurement results obtained from these two parts are correlated. The reliability of the whole test depends on the extent that these two parts correlate (Heaton 1988; Bachman 1990; Alderson et al. 1995).

2.5.1.2.1.2 RATER CONSISTENCY

Objective marking of tests is of vital importance to provide the reliability of a test. However, especially in writing and speaking tests, test scores are marked subjectively. As a result, there may be inconsistent test scores. ‘Intra-rater reliability’ and ‘inter-rater reliability’ are two important concepts two provide the rater consistency.

2.5.1.2.1.3 INTRA-RATER RELIABILITY

Intra-rater reliability refers to the consistency of marks of the same rater when he or she gives the same test on two different occasions (Bachman 1990; Alderson et al. 1995). However, some problems may occur in rating situations. For example, Bachman (1990: 179) points out that “The sequence of scoring may lead to inconsistency not only in the rating criteria themselves but also the way which the rating criteria are applied”.

According to Bachman (1990), and Alderson et al. (1995) the best way to determine intra-rater reliability is through obtaining two independent marks from the same rater for each test. This procedure can be carried out by marking the tests once and then marking these tests once more. However, the tests should be randomized at a different time when next marking is done. When this procedure is completed the reliability between these marks can be measured by computing correlation coefficient or coefficient alpha.

2.5.1.2.1.4 INTER-RATER RELIABILITY

Different testers may obtain different results when they are asked to mark the same test. Alderson et al. (1995: 129) defines inter-rater reliability as “the degree of similarity between two or more examiners. Therefore, it is of vital importance for each examiner to match the standard all the time”. The reliability between these ratings can be measured by computing correlation coefficient or coefficient alpha as in estimating intra-rater reliability (Bachman 1990).

2.5.1.2.1.5 EQUIVALENCE (PARALLEL FORMS RELIABILITY)

The reliability of a test can also be measured by carrying out parallel forms of the test to the same group. That is, these tests should be identical in terms of their sampling, difficulty, length, rubrics, etc. In this approach, the scores obtained from two similar tests are compared. The test can be considered reliable if the correlation between the two tests is high. (Heaton 1988).

Bachman (1990) claims that when two tests measure the same language ability they can be considered parallel tests. In other words, the test taker’s score in both tests should be the same. However, Alderson et al. (1995) point out that it is not easy to carry out this process, because developing two identical tests is too difficult.

2.5.1.2.1.6 STABILITY (TEST-RETEST RELIABILITY)

One of the ways of measuring reliability of a test is stability or test-retest reliability. In this approach, the test is given to a group of students and then the test is given to the same group of students once more. The correlation between the two test scores are then computed and interpreted. If the students get the same scores on both tests the test is considered reliable. However, the test giver should provide that the test takers should stay the same in both test applications (Bachman 1990).

According to Alderson et al. (1995) and Heaton (1988) this approach is not practical. Because, the scores of two tests may change when the test takers get used to test method and format. In addition, personal factors such as motivation may affect the performance of some test takers. This problem can be solved through providing longer time interval between the test applications but this can lead to new problems because during this interval test takers may have changed.

2.5.1.2.2 GENERALIZABILITY THEORY

Generalizability Theory (G-Theory) was developed by Cronbach and his colleagues in order to identify and measure the effects of different factors on the test scores. Test users may identify these factors as abilities or sources of error by carrying out generalizability theory. In this theory, an obtained test score is considered as a sample. When interpreting this test score, the sample is generalized to other testing situations. In other words, generalizability is related to reliability. G-Theory is superior to CTS model in many respects. First, it helps the test developer to specify and examine different sources of error which cannot be identified in CTS model. Second, it helps the test user to measure the relative effects of variance in just one test application. The test user can make use of these measures to specify the methods which can increase the reliability of a given test (Bachman 1990).

2.5.1.2.3 ITEM RESPONSE THEORY

Item response theory (IRT) was developed as an alternative to CTS and G-theory. Because, Bachman (1990) and Alderson et al. (1995) claim that both CTS and G-theory have some drawbacks. This theory is also called as 'latent-trait' theory (Bachman 1990).

Alderson et al. (1995) point out that the relationship between the test taker and test characteristics is very strong. For example, when the results of a given test are analyzed the results are only true for this testing situation. In other words, these results cannot be generalized to the other test measures. Because, as Alderson et al. (1995: 89) put it:

> "If the items in a test have low facility values, the test may be difficult, or it may have been tried out on low-level students. If the facility values are high, the test may be easy, or it may have been given to highly proficient students. Because of this it is difficult to compare students who have taken different tests, or to compare items that have been tried out on different groups of students."

Therefore, two main points should be taken into consideration when interpreting the test results. First, the level of difficulty of the item and second, the individual's level of ability.

IRT is a useful for the test constructor in many respects. First, it can be used to determine not only the inappropriate items for a test but also inappropriate students for the testing group. Second, it helps to identify test bias. Third, it can be used to analyze the results of objective and subjective tests, and finally it is indispensible for computer adaptive testing. (Alderson et al. 1995).

2.5.2 VALIDITY

Validity as well as reliability is one of the essential elements in test development and test use. Reliability and validity concepts should be handled together since there is a strong relationship between them. Therefore, it is of vital importance to explain these two concepts clearly and identify the differences between them for appropriate test development and use.

Henning (1987) defines validity as follows:

> "Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term valid when used to describe a test should usually be accompanied by the preposition for. Any test than may be valid for some purposes, but not for others." (Henning 1987: 89 in Alderson et al. 1995: 170)

According to Henning's definition the purpose of the test is the basic requirement for the test to be valid. A test should be used for the intended purpose or should measure what it aims to measure in order to be valid. However, it is hard to say that a test is totally valid. Because, while a test is considered to be valid for some purposes it cannot be valid for the other purposes. Therefore, there are degrees of validity which may differ from purpose to purpose.

Similarly, Geneese et al. (1996: 62) mention the importance of the intended purpose' of the test in order to establish validity by their definition of validity: "Validity is the extent to which the information you collect actually reflects the characteristics or attribute you want to know about". Therefore, the more a test measures what it aims to measure the more valid it becomes.

There are many different types of validity. The measurement specialists classify validity into different types: Face validity, content validity, criterion validity and construct validity (Bachman 1990; Alderson et al. 1995; Hughes 2002).

The American Psychological Association (1985) asserts that there is a unitary concept of validity involving all these different approaches to validity and it defines the unitary concept of validity as follows:

> "Validity...is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of the test are validated, not the test itself."
> (American Psychological Association 1985: 9 in Bachman 1990: 236-237).

Alderson et al. (1995) claim that a test should be validated in many different ways. This can be accomplished by establishing as many different types of validity and collecting as much information as possible for each type of validity. This procedure is also called as 'evidential basis' for test interpretation and use by Messick (1980, 1989 in Bachman 1990: 242).

There is a strong relationship between test and the ethical values of the culture. Because, test use has consequences on the social structure of the society. Messick (1975, 1980, 1989) mentions the role of ethical values on test use and interpretation as follows:

> "The examination of validity has also traditionally focused on the types of evidence that need to be gathered to support a particular meaning or use. Given the significant role that testing now plays in influencing educational and social decisions about individuals, however, we can no longer limit our investigation of validity to collecting factual evidence to support a given interpretation or use. Since testing takes place in an educational or social context, we must also consider the educational and social consequences of the uses we make of tests. Examining the validity of test scores therefore a complex process that must involve the examination of both the evidence that supports that interpretation or use and the ethical values that provide the basis or justification for that interpretation or use."
> (Messick 1975, 1980, 1989 in Bachman 1990: 237).

Therefore, test development and use should not be implemented without considering the value systems of the society. Because, the decisions based on test scores affect the culture that the test is carried out. For this reason, involving the cultural values in making interpretations about test results is of great value for its contribution to the society.

2.5.2.1 TYPES OF VALIDITY

There are three basic types of validity: Content relevance, criterion relatedness and meaningfulness of construct. In fact, these different types of validity refer to different ways of measuring validity (Bachman 1990).

Genesee et al. (1996: 64) indicate that it is impossible to estimate validity directly. Because as they point out:

> "To assess the validity of information directly, you would have to be certain of the true state of affairs in order to compare it with the information you have collected. In the realm of human assessment, most of the qualities and the attributes evaluaters are interested in are not themselves subject to direct assessment. Thus, there is no direct way to know the true level of most human qualities or abilities that we are interested in. We have only indicators that allow us to make inferences about attributes or interest."

For this reason, the testers should make use of different methods to obtain information in order to determine the validity of assessment. According to Genesee et al. (1996) content relevance, criterion relatedness and construct validity are three basic types of validity.

2.5.2.1.1 FACE VALIDITY

Face validity is not considered as a scientific concept different from the three basic types of validity stated above. Because, the people who decide the face validity of a test are not specialists in measurement area. Students and non-expert test users can be counted among these people. The face validity concept became a more important issue when CLT (Communicative language testing) was first introduced in the language testing field. Because, a communicative language test should represent the real life situations (Alderson et al. 1995).

Hughes (2002) asserts that a test can be accepted to have face validity if it looks as if it measures what it is supposed to measure. According to Ingram (1977: 18 in Alderson et al. 1995: 172) "Face validity refers to the test's surface credibility or public acceptability" .

Face validity is important for a couple of reasons. First, a test lacking face validity may not be accepted by people working in language testing field (Hughes 2002). Second, if the test users do not think that a test is valid they may not take the test seriously for its purpose. Third, the face validity of the test affects the response validity of the test. For example, the test takers may perform better if they believe the test is valid and the vice versa (Alderson et al. 1995).

2.5.2.1.2 CONTENT VALIDITY

One of the essential parts of test validation is content validity. When the test's content is formed of the representative sample of the structures, language skills etc. that it is supposed to measure than it can be accepted to have content validity (Hughes 2002).

Genesee et al. (1996: 251) define content validity as " the extent to which a test provides an adequate representation (coverage) of the language domain it intends to test".

The difference between the face validity and the content validity depends on the judgment of these two validity concepts. In face validity it is not necessary to accept the judgment of other people even though these people's judgment is respected. However, in content validity the judgment of these people is accepted to be true since they are considered to be experts in language testing field (Alderson et al. 1995).

According to Bachman (1990) content validation can be classified into two types: content coverage and content relevance. In content coverage, the tasks of the tests should represent the behavioral domain. The content coverage refers to how much these tasks represent that domain. Specification of the ability domain and test method facets are two necessary elements of content relevance. Domain specification is important in defining constructs. On the other hand, Cronbach (1971) explains the importance of test method facets in test validation as follows:

> "a validation study examines the procedure as a whole. Every aspect of the setting in which the test is given and every detail of the procedure may have an influence on performance and hence on what is measured. Are the examiner's sex, status, and ethnic group the same as those of the examinee? Does he put the examinee at ease? Does he suggest that the test will affect the examinee's future, or does he explain that he is merely checking out the effectiveness of the instructional method? Changes in procedure such as these lead to substantial changes in ability-and personality,test performance, and hence in the appropriate interpretation of test scores... The measurement procedure being validated needs to be described with such clarity that other investigators could reproduce the significant aspects of the procedure themselves." (Cronbach 1971: 449 in Bachman 1990: 242).

Hughes (2002) points out that the test's content validity is important for several reasons. First, the test's content validity affects the overall validity of a test. The degree of overall validity increases when the test becomes more content valid. Second, the specified areas in the test specifications should be represented well in the test. Otherwise its backwash effect may be harmful. As a result, it decreases the content and the overall validity of the test. Therefore, writing full test specifications representing the areas to be tested is the best way to establish the content validity of the test.

2.5.2.1.3 CRITERION VALIDITY

"Criterion relatedness is the extent to which information about some attribute or quality assessed by one method correlates with or is related to information about the same or a related quality assessed by a different method" (Genesee et al. 1996: 66). For example, when a test takers' performance is measured with two different tests, the performance of the test taker refers to criterion.

According to Hughes (2002: 23) "there are two kinds of criterion-related validity: Concurrent validity and predictive validity. In concurrent validity, the the test and the criterion are administered at about the same time". On the other hand, predictive validity refers to the use of test scores to predict the testers' future behaviour (Bachman 1990; Alderson et al. 1995; Hughes 2002).

Genesee et al. (1996: 250) point out that "Criterion relatedness is shown by correlations between test scores and criterion measures. A criterion measure may be another test of the same ability whose validity is already well established". Proficiency tests such as IELTS and TOEFL are the most common examples to tests which are used for predictive validation (Alderson et al. 1995).

2.5.2.1.4 CONSTRUCT VALIDITY

Bachman (1990: 255) states that "construct validity concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs".

This concept was first formed in 1950s when the American Psychological Association decided to write a code of professional ethics focused on the appropriateness of psychological tests (Cronbach 1988 in Bachman 1990). In the following years, construct validity has become one of the main concepts contributing to the appropriate interpretation of test scores.

Ebel and Frisbie (1991) define the construct validity as follows:

> "The term construct refers to a psychological construct, a theoretical conceptualization about an aspect of human behavior that cannot be measured or observed directly. Examples of constructs are intelligence, achievement motivation, anxiety, achievement, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test is indeed measures the psychological construct the makers intend to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect from them to mean." (Ebel and Frisbie 1991: 108 in Alderson et al. : 183).

Similarly, Bachman (1990) states that the abilities to be measured cannot be directly observed. However, the ability estimates can be obtained according to the observed performance. Moreover, the abilites observed are considered to be theoretical, that is, these abilities are believed to affect the individuals' language use and performance on language tests. In construct validity, these abilities of the individuals are interpreted according to the test performance. Therefore, it is important to obtain evidence showing the relationships between test scores and abilities. Finally, it can be concluded that construct validity involves verifying and falsifying a scientific theory. Therefore, in order to prove or disprove a theory, construct validation makes use of methods such as logical analysis and empirical investigation.

2.6 CHAPTER SUMMARY

This chapter started with the description of basic terms regarding testing in language teaching. The uses of language tests in educational programs and the approaches developed for language testing so far were stated. The classification of language tests were indicated as well. Finally, the qualities of a language test were discussed in details

CHAPTER THREE
ETHICS IN LANGUAGE TESTING

3.0 INTRODUCTION

Test developers and specialists considered the validity and reliability qualities as the basic ethical issues in language testing field so far. However, in the last two decades language testing researchers began to focus on issues such as test standards, test bias, equity and ethics besides reliability and validity issues. Therefore, new models for ethical test use were developed.

In this chapter, first, the concepts of ethics, morality and fairness will be explained. Second, approaches to ethics will be presented. Third, ethical frameworks developed will be discussed. Fourth, the uses of tests in language programs and the role of stakeholders in test development will be discussed. Finally, the concepts of professionalism, test bias, and washback in language testing will be explained.

3.1 THE CONCEPT OF ETHICS

Ethics plays a vital role in each area of a person’s life. Ethics refers to standards of behavior that determine how human beings ought to act in the many situations in which they find themselves-as friends, parents, children, citizens, businesspeople, teachers, professionals, and so on.
(http://www.scu.edu/ethics/practicing/decision/framework.html).

House (1990: 91 in Davies 1997: 328) defines ‘ethics’ as “the rules or standards of right conduct or practice, especially the standards of profession”.

According to Scriven (1991: 134 in Davies 1997: 329) "ethics is the emperor of the social sciences, imperial because it refers to considerations that supervene above all others, such as obligations to science, prudence, culture and nation".

The test developers, test users, and the people engaged in language testing profession all have responsibilities to the profession and the society at large. According to O'Hear (1985: 277 in Davies 1997: 236) "participation in such groups is characterized by the recognition on the part of all of reciprocal duties, rights, loyalties and deserts, which are not a matter of individual choice or preference". Therefore, the people involved in the profession of language testing should act according to certain rules that this profession requires. In other words there are standards of behavior that the language testers should accept and obey. For this reason, the concept of 'normative ethics' should be defined. Two approaches were developed in normative ethics in order to follow norms: teleological approach and deontological approach. While, teleological approach is realistic, deontological approach is idealistic. The result is the most important concern of teleological approach. On the other hand, deontological approach aims to ensure fairness through considering values (Davies 1997). Stewart (1996 in Davies, 1997: 237) calls these approaches as "individual good and general good" respectively.

Rawls (1967: 221 in Davies, 1997: 237) points that these two approaches are equally important in terms of justice:

> "...first each person engaged in an institution or affected by it has an equal right to the most extensive liberty compatible with a like liberty for all: and second, inequalities as defined by the institutional structure or fostered by it are arbitrary unless it is reasonable to expect that the will work out to everyone's advantage and provided that the positions and offices to which they attach or from which they may be gained are open to all."

Therefore, 'ethics' comprises both principles and it should balance them. Each individual has rights and roles in a society. However, since he or she lives in a society, he or she has responsibilities to the society as well. For this reason, 'ethics' should to take into consideration both social justice (deontological approach) and individual justice (teleological approcah). Otherwise when one of these principles does not get the attention it deserves, it may totally disappear. This may have dire consequences to the society and the individual. For example, there may be disagreements between the society and the individual. Each party can fight for their cause. As a result, the clashes between the individual and the society are inevitable.

'Ethics' and 'ethical' are not the only terms used to express appropriate behavior and conduct. The term 'moral' is also widely used in every context in the society. Therefore, the relationship between these terms should be explained.

Webster's ninth new collegiate dictionary of the English Language (1994 edition) defines the terms 'moral' and 'ethical' as follows:

> "Moral and Ethical are both concerned with rightness and wrongness of actions and conduct, but moral is more often applied to the practice or actions of individuals, often specifically in sexual relations, ethical more often to theoretical of general questions of rightness, fairness or equity."

Davies (1997) claims that the term 'moral' refers to actions of individuals but the terms ethics/ethicality and morals/morality are still used instead of each other.

Concepts of 'fairness' and 'ethics' will be used interchangeably in this study since fairness addresses the overall consequences of ethics, or ethicality.

3.2 APPROACHES TO ETHICS

Five basic approaches were developed to identify the ethical standards. Because, it is difficult to decide on the principles to determine the ethical standards. The ethical standards may be based on religion, law, feelings, science and etc. However, the concerns such as; which rules of these concepts can be applied to what extent? Should there be common ethical values that each society agrees on? Does each culture have different moral values? If each culture has different moral values how can they create ethical standards and implement them? need to be answered in order to determine ethical standards for each society.

The ethicists and philosophers developed five approaches to ethics to solve these dilemmas: The utilitarian approach, the rights approach, the fairness and justice approach, the common good approach and the virtue approach.

3.2.1 THE UTILITARIAN APPROACH

The utilitarian approach aims to provide the greatest good and least harm for all the stakeholders of the society and the environment. The balance between the good and harm is the basic principle of this approach.

3.2.2 THE RIGHTS APPROACH

The rights approach suggests that the moral rights of the stakeholders is the most important aspect in ethics. That is, each individual is born with a dignity on their own; therefore, they are able and free to make decisions about their lives. However, when one is leading his or her own life, he or she should not intervene in other individuals lives. In other words, each individual should respect others' rights.

3.2.3 THE FAIRNESS OR JUSTICE APPROACH

The fairness or justice approach was developed by Aristotle and other Greek philosophers. According to this approach, all equals should be treated equally.

Moreover, the standards determined should be sensible. For instance, the more people work the more money they get. This situation is considered justified. However, there are discussions about this approach as well. For example, there are managers who are paid much more than the other workers. The fairness of this situation is questionable.

3.2.4 THE COMMON GOOD APPROACH

According to the common good approach since people live together in a community, each individual can contribute to the community, respect the others and live peacufully. Therefore, each one’s happiness is important and the system should provide the opportunities, and the necessities for each individual.

3.2.5 THE VIRTUE APPROACH

The virtue approach suggests that each individual should act at their very best in order to develop their humanity in every respect. The virtues such as honesty, courage, compassion, generosity, tolerance, love, fidelity, integrity, fairness, self-control that each one should possess and act according to them (http://www.scu.edu/ethics/practicing/decision/framework.html).

As it was stated above the philosophers searched for best of ‘ethics’ for the public good. Each approach values different aspects of ethics. On the other hand, developing a unitary ethics approach seems not possible. Because, what is ethical and what is not may change from person to person, from culture to culture as well as from situation to situation. Therefore, applying all the principles of these approaches is against the human nature. For this reason, each society should determine its ethical values considering the culture, history, and religion of itself. Moreover, each culture is valuable and has a part to play in world history. Furthermore, cultures interact each other; therefore, they have some common ethical values. As a result, each society can determine its ethical standards by taking into consideration both its own and the common ethical values.

3.3 LANGUAGE TESTING AND ETHICS

Language tests are used to make educational purposes such as selection, placement, diognosis and evaluation as it was discussed in chapter 1. Language testers carried out most of the research in order to identify the sources of unrelibility and invalidity to examine the fairness, in other words ethics. However, fairness cannot be limited to the issues of validity and reliability. That is, the use of language tests should be taken into consideration as well as test level. Since "tests are very powerful instruments which have big impact on the individuals, the programs and the society" (Shohamy 2000: 15).

Similarly, Bachman (1990: 279) points out that "Tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or of society at large." Pennycook (1994 in Hamp-Lyons 1997: 302) also claims that "language learning is not limited to classrooms, and its consequences are not only educational but also social and political".

Hulin et al. (1983 in Bachman 1990: 280) state that "the rights, interests of test takers, the decisions based on these tests made by the institutions responsible for testing, and the public interest should be taken into consideration by the test developers and test users to provide ethics in language testing".

Moreover, Kunnan (2000) points out that fairness concept should be discussed within the framework of social justice rather than fundemental fairness concerns such as reliability and validity.

The ethics of language test use has a wide coverage: The rights of test takers (secrecy, confidentiality, access to information etc.), the balance between individual rights and the values of society, the responsibilities of test developers and test users. Moreover, the impact of these issues changes from culture to culture and one testing context to another (Bachman 1990).

Furthermore, "test developers should take into account test takers of different races, gender, ethnic backgrounds or handicapping conditions when constructing tests to ensure fairness in language testing" (Kunnan, 2000: 1).

Shohamy (2000: 15-16) lists the questions regarding the use of language tests as follows:

> "How are tests being used by decision makers?
> How are scores being interpreted?
> Are language tests used according to their intended purposes?
> What are the consequences of tests?
> Are tests used fairly?
> Do tests create biases?
> What is the impact of tests on learning and teaching?
> Who are the users and instigators of language tests?"

Therefore, ethical language test use involves a lot of concerns. It is clear that appropriate testing practice is necessary. However, the ethical test practice cannot be limited to just reliability and validity concerns. As it was mentioned above test practices not only affect the society but also the individuals. Therefore, stakeholder involvement in language testing practice, considering values of society and individual rights of test takers in test development and use and using tests for intended purposes are of vital importance for providing ethics in language testing field. In the following sections, these concerns regarding ethical test use will be dealt.

3.4 THE USE OF LANGUAGE TESTS FOR UNINTENDED PURPOSES

Tests are very powerful instruments which have big impact on the individuals, the programs and the society. Tests were used for many different purposes other than measuring the language skills. In other words language tests also serve to fulfill political, educational purposes (Shohamy 2000).

Similarly, Pennycook (1994 in Hamp-Lyons 1997: 302) points out that "Language learning is not limited to classrooms, and its consequences are not only educational, but also social and political".

Schmeiser (1995: 1) states the importance of ethical test use and some unethical practices as follows:

> "Every profession has distinct ethical obligations to the public. These obligations include professional competency, integrity, honesty, confidentiality, objectivity, public safety and fairness, all of which are intended to preserve and safeguard public confidence. Those who are involved with assessment are unfortunately not immune to unethical practices. Abusing in preparing students to take tests as well as in the use and interpretation of test results have been widely publicized. Misuses of test data in high-stakes decisions, such as scholarship awards, retention/promotion decisions, and accountability decisions, have been reported all too frequently. Even claims made in advertisements about the success rates of test coaching courses have raised questions about truth in advertising."

As Schmeiser (1995) mentions above language testing as a profession forms an essential part of society and unfortunately unethical testing practices are common such as inappropriate use of high-stakes decisions. However, the unfair testing practices go beyond these.

Shohamy (2000: 17) states the inappropriate use of tests for political purposes as follows:

> "In the political levels tests are used to create de facto language policies (bureocratic goals), to raise the status of some languages and to lower the that of others, to control citizenship, to include, to exclude, to gatekeep, to maintain the power of the elite and to offer simplistic solutions to complex problems."

Spolsky (1997: 242) also points out that “tests have always been used as a means of political and social control since their invention”. For example, “Shibboleth Test in the Bible (Judges, 12: 4-6) was used to distinguish members of two enemy communities through their pronunciation differences” (McNamara 2005: 352).

The language tests were also used to change the system from Aristocracy to Meritocracy during American and French revolutions (Madaus and Kelleghan 1991 in Spolsky 1997).

Moreover, language tests were also used to solve political issues which could not be solved by policy making. For instance, President Clinton intoduced national tests to save the U.S. educational system (Shohamy 2000).

On the other hand, Spolsky (1997: 242) argues “the ‘gatekeeping function’ of tests which refers to the use of examination results to determine qualifications for positions or for training for positions”. For example, the main purpose of Chinese examination system was to help to emporer to select high-ranking officers who stay loyal to himself other than the other land owners.

Shohamy (2000) also points out that the use of tests for political purposes serves the needs of the people in power rather than the people who are affected in the testing process, teachers and test takers. Because, the teachers are supposed to follow orders and tests become a tool for the people in power to control the system. In addition, the people who are involved in the testing process have no rights in this system.

The dictation test which was implemented by the Australian Government is another example to the use of language tests for political purposes. The test was designed to prevent the immigrants enter Australia (McNamara 2005).

Similarly, the use of examinations for selecting personnel for Indian Civil Service was criticized for their inappropriate use in 1858. Because, according to the critics students were encouraged to cram. The term ‘cram’ refers to “memorizing examination results and answers” (New English Dictionary 1815). Later, the term ‘crammer’ appeared who aims to prepare students to pass the test rather than teaching them (Spolsky 1997).

Language tests are also used to control the syllabus and what happens in the classroom (Spolsky 1997; Shohamy 2000). For instance, the unethical use of a reading comprehension test. Although it was announced that the test aimed to measure the achievements of 4th and 5th grade students, it was carried out to prove the power of Ministry of Education on the system (Shohamy 2000).

Therefore, language tests were used for unintended purposes throughout the history. These purposes were not only political but also educational. Moreover, the learners were affected the most among other stakeholders. In order to provide ethics in language testing a number of language specialists developed models to enhance fairness in langauge testing field.

3.5 MODELS DEVELOPED FOR THE ETHICAL USE OF TESTS IN LANGUAGE PROGRAMS

Ethics, in other words fairness is a complex concept to define. As it was defined in the previous section, what is ethical and what is not ethical may change from person to person, culture to culture as well as one testing situation to another. Therefore, it is difficult to reach a consensus in the concept of ethics. Furthermore, the inappropriate use of tests is common virtually in every community. For this reason, a number of models were developed to construct ethical language tests addressing the needs of all stakeholders. These models are, Kunnan’s test fairness model, Bachman and Palmer’s test usefulness model, Messick’s ethical test model and McNamara’s ethical test model.

3.5.1 KUNNAN’S TEST FAIRNESS MODEL

Code of Fair Testing Principles in Education which was prepared by the Joint Committee on Practices in 1988 determined the standards for test developers and users. The code consists of four basic areas: Developing and selecting tests, interpreting scores, striving for fairness and informing test takers (Appendix A). Kunnan (2000) developed a model in the light of the issues addressed in the code. Kunnan (2000: 3) states that “validity, access to test and justice are the main concerns of fairness in language testing”.

Validity concern focuses on the construct validity of test-score interpretations for test takers having different characteristics such as culture, gender, race and etc. Content bias may give advantage to some test takers over others. Some test takers may perform differently than other test takers. For this reason, test developers and users should ensure that performance differences are the result of abilities rather than the other factors. Finally, insensitive language may lead to stereotyping of test taker groups.

Access concern refers to accessibility of tests for the test takers. The test developers and users should consider the financial and geographical aspects for test takers coming from far districts and having different incomes. As for personal aspect, accommodations for the disabled should be provided and sufficient information regarding the type of accommodations should be given to test takers in advance. Moreover, some test takers may be assessed according to the material they had the opportunity to learn some may not. Finally, familiarity with the equipment and testing conditions and acccess to test-taking equipment play a vital role in the performance of test takers.

Justice concern focuses on societal equity and legal challenges. Test takers belonging to different culture, nationality, gender, race, ethnicity and etc. might not be treated equally as the other test taker groups when attending a college or applying for a job. For this reason, new testing programs should be designed to provide equal opportunities for all test taker groups. Kunnan's test fairness model is shown below (Kunnan 2000) (see Table 1).

Main Concern	**Specific Focus**
Validity	Construct Validity
	Content and Format Bias
	Differential Item/Test Functioning
	Insensitive Language
	Stereotyping of Test Taker Groups
Access	Financial: Affordability
	Geographical: Location and Distance
	Personal : Accommodations for Disabled Persons
	Educational: Opportunity to Learn
	Equipment and Test Conditions
Justice	Societal Equity
	Legal Challenges

Table 1: Main Concerns of Fairness (Taken from Kunnan 2000: 3)

3.5.2 BACHMAN AND PALMER'S TEST USEFULNESS MODEL

Bachman et al. (1996) point out that a test should be used for the intended purposes. Therefore, test usefulness is the most important quality of a test. Reliability and validity were considered as the fundemental qualities of a language test by many language specialists (Alderson 1997; Shohamy 1997b in 1999). However, Heaton (1988) claims that there is a conflict between reliability and validity issues, because maximizing reliability minimizes validity and the vice versa. For this reason, Bachman et al. (1996: 18) propose that “test developers should provide the balance among the different test qualities varying from one testing situation to another. Therefore, different qualities should be taken into consideration to develop an ethical language test. He calls this as ‘test usefulness’.”

Test usefulness involves the basic qualities of a test. These are, “reliability, construct validity, authenticity, interactiveness, impact and practicality” (Bachman et al. 1996: 18). In order to design an ethical test these qualities should be provided as much as possible (Figure 1).

USEFULNESS= Reliability + Construct Validity + Authenticity + Interactiveness + Impact + Practicality

Figure 1: Test Usefulness (Taken from Bachman et al. 1996: 18)

Reliability and construct validity concepts were explained in the previous chapter. Therefore, the concepts of authenticity, interactiveness, impact and practicality will be explained in this part.

Authenticity refers to the level of relationship of the qualities of a given language test task to the characteristics of a target language use (TLU) task. The connection between characteristics of the TLU task and characteristics of the test task are shown in figure 2.

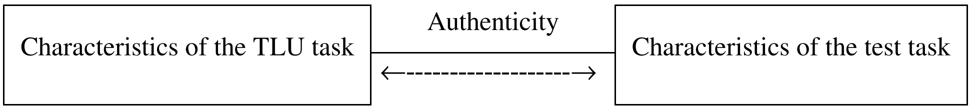
Characteristics of the TLU task
Authenticity
Characteristics of the test task

Figure 2: Authenticity (Taken from Bachman et al. 1996: 23)

Interactiveness refers to the degree to which the constructs to be assessed are involved in accomplishing the test task. Test taker’s language ability (language knowledge and strategic competence or metacognitive strategies), topical knowledge, and affective schemeta are the individual characteristics of the test taker (Figure 3).

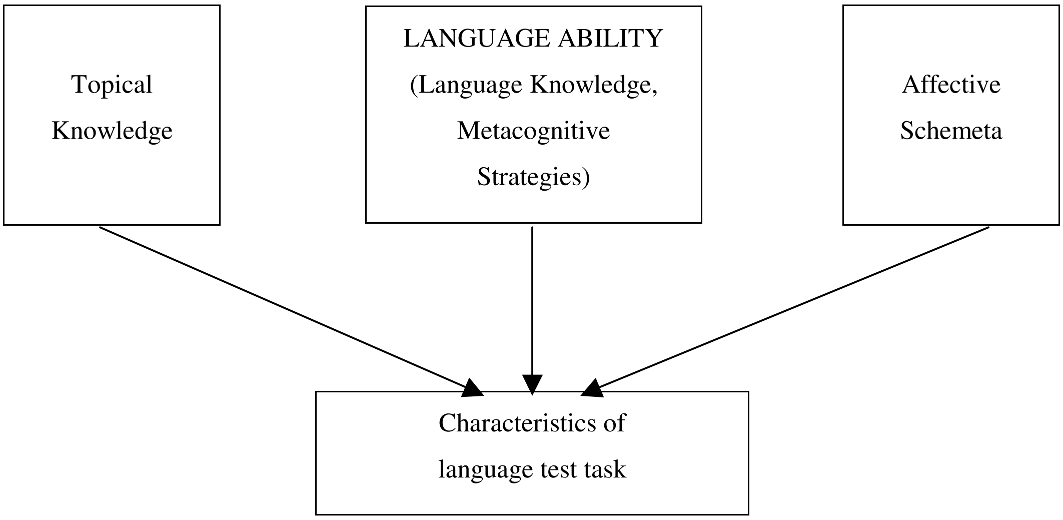
Topical Knowledge
LANGUAGE ABILITY (Language Knowledge, Metacognitive Strategies)
Affective Schemeta
Characteristics of language test task

Figure 3: Interactiveness (Taken from Bachman et al. 1996: 26)

Impact focuses on how test use affects the society, the education system and the other stakeholders related to education. The individual, societal and educational value systems should be taken into consideration in test development and test use. Test impact should be examined in two levels. While micro level refers to individuals affected by the specific test use, macro level is concerned about the society and the education system.

Practicality refers to the ways the test is administered in a particular testing situation. The resources required to develop and use of the test should be determined and provided in advance. These resources involve human resources, material resources and time (Bachman et al. 1996) (see Figure 4).

1. **Human Resources** (e.g. test writers, scorers or raters, and test administrators).
2. **Material Resources** Space (e.g. rooms for test development and test administration), Equipment (e.g. word processors, tape and video recorders, computers).
3. **Time** Development time (time from the beginning of the test development process to the reporting of scores from the first operational administration), Time for specific tasks (e.g. designing, writing, adminestering, scoring, analyzing)

Figure 4: Types of resources (Taken from Bachman et al. 1996: 37)

Bachman et al. (1996: 149) point out that “It is of vital importance to collect information related to evaluation of usefulness. Because, it helps to provide the essential qualities of a test which refer to ‘test usefulness’. The information can be collected in the initial stages as well as during the administration stage”. Bachman et al. (1996) designed a checklist to evaluate test usefulness (Appendix B).

3.5.3 MESSICK’S ETHICAL TEST MODEL

According to Messick (1980 in Bachman 1990: 281) “there are four areas to be taken into consideration in order to explain ethical use and interpretation of test results: Construct validity, value systems informing the particular test use, practical usefulness of the test and the consequences of test use to the educational system or society (washback, impact) using test results for a particular purpose” (see Figure 5).

Ethical Test= Construct validity + value systems informing the particular test use + practical usefulness of the test + washback and impact.

Figure 5: Messick’s ethical test model

3.5.4 MCNAMARA’S ETHICAL TEST MODEL

McNamara (2000) claims that there are two approaches regarding the political and social role of tests. The first one advocates that it is possible to make the language testing practice ethical and language testers are primarily responsible to carry out this goal. According to the second approach, language tests serve to maintain power and control. While the first approach is called ethical languge testing, the second one is called critical language testing. Ethical language test model of McNamara will be explained in this part.

McNamara (2000) asserts that ethical language testing have three main concerns: These are: accountability, washback and test impact. Accountability is about the responsibility of test developers and test users to test takers. Washback refers to the effects of tests in teaching and learning. These effects can be both positive and negative. However, ethical language testing requires positive washback. Finally, the effects of test may not be limited to just classroom. Tests may have wider effects in the community including the school which is called test impact (see Figure 6).

Concerns of Ethical language testing = Accountability + Washback + Impact

Figure 6: McNamara's concerns regarding ethical language testing

All these models have common principles. These are construct validity, washback and impact. Therefore, the language specialists attach importance to social consequences of test more than any other quality. In addition, since stakeholders form the most essential part of the society they are affected a great deal by these consequences. However, test takers are affected the most in all stakeholders. The individual rights of test takers should will be explained next part.

3.6 INDIVIDUAL RIGHTS OF THE TEST TAKER

Punch (1994 in Lynch 1997: 317-318) states that "consent, deception, privacy and confidentiality are the basic individual rights of a test taker."

Consent refers to have right to taking or not taking the test. However, the term 'informed consent' is about informing test takers regarding the reasons for testing and how the test results will be used.

Deception focuses on testing the individuals with the test types or strategies which test takers are not used to. Therefore, it questions if it is ethical to make decisions on the performances of test takers in these tests.

Privacy and confidentiality are concerned about if the test results can be used to insult test takers. That is, a test taker may not be able to enter the university or fail to obtain results necessary to get the job he wants. This argument is based on the principle that “test takers should not be harmed by the test” (Hamp-Lyons 1989: 13 in Lynch 1997: 318). On the other hand, counter argument suggests that tests serve to determine the differences of an ability. Therefore, tests show the existing differences rather than leading them.

Test takers are not the only stakeholders in language testing practice. The others are teachers, parents, government and official bodies and etc. Since the stakeholders are affected the most in language testing process. Their involvement should not be ignored. Therefore, next section examines the contribution of stakehoholder involvement in test development and use.

3.7 STAKEHOLDER INVOLVEMENT IN LANGUAGE TESTING PROCESS

The stakeholders comprise a wide range of people from different occupations. The language testers, teachers, parents, administrators, teacher educators, sponsors and funding bodies, government bodies, the public, various international and national examination authorities, members of working parties and curriculum committees, test takers, test administrators such as university admission officers interpreting test scores. The stakeholders can also be grouped in two main categories: The people who are decision makers and the people affected by those decisions. The former one have a more powerful role than the latter one in language test use (Rea-Dickens 1997).

The participation of stakeholders in the test development and test use process is considered one of the necessities of ethical test use. Therefore, three of the stakeholders will be discussed in the following section because they are considered as the most important ones.

3.7.1 LEARNERS

Learners (test-takers) are considered as the most important stakeholders. Because, learners are directly affected by the decisions made by the test users. For this reason, involving learners to testing process constitutes one of the fundemental responsibilities of test developers and users.

Hamp-Lyons (2000) states that in a learning environment the students are affected positively and negatively by their teachers' excessive encouragement or criticism. That is, the teacher's behavior towards errors of students, and the level of tasks determined by the teacher to assess performance of students affect them. As a result, students may consider themselves more successful or less successful than their real achievement. This can be solved through benchmarking teachers. Teacher benchmarking refers to ensure that teachers can carry out the appropriate standards both in teaching and testing the students. Furthermore, there is not only one model to teach and learn a foreign language. Therefore, there is not a specicific way to assess the progress, ability or achievement of a learner. For this reason, learners should have the opportunity to choose the learning styles, strategies as well as test types which are the most appropriate for their learning characteristics. This can be achieved by developing different types of test and giving learners the chance to select from these tests.

Involving learners to the testing process also requires learning about the views of them. However, Alderson and Clapham (1992) indicate that it is hard to make sure the views of learners and interpreting these views since the views of learners contradict each other. For this reason, data collecting regarding their views should be planned as a long term project in order to obtain meaningful results.

3.7.2 TEACHERS

Teachers are considered as one of the most important stakeholders. Since as Wall and Alderson (1993) state teachers have two fundemental roles: First, teachers are the informants in the research and development process and second they design assessment procedures and maintain educational standards. However, there are a number of questions to be answered with regard to teachers' role in the assessment process (Rea-Dickens 1997: 307):

> 1. How much control do teachers have of the assessment procedures and the tests they administer?
> 2. In cases where specific-purpose test instruments are required and not provided, to what extent have teachers been skilled so as to construct their own?
> 3. Are tests in textbooks relied on solely because there are no other sources or resources to draw upon?
> 4. Should teachers be expected or required to design their own tests?

Therefore, teacher education is necessary to overcome these problems. It will be useful to determine standards for an ideal teacher. Since, it is certain that they have to carry out the duties stated above and they are the most important decision makers to carry out these practices.

Furthermore, Hamp-Lyons (2000) states that each teacher has a different philosophy, personality and experience. Therefore, they do not have to stick to formal assessment methods. In other words, they can apply different teaching styles and strategies as well as testing styles in the classroom. As a result, there should be alternative scoring methods, so that testers could select the most appropriate method matching their philosophies.

Rea-Dickens (1997) points out that one of the important aspects of stakeholder involvement in curriculum development is teacher education in order to implement the necessities of the new curriculum. According to Spielman (1993) this approach helps the teachers to understand the assessment process better and practise with their own learners. Furthermore, it also enhances professional dialogue with teachers and encourages staff development. As a result, the teachers can make use of the assessment plan appropriately through critical dialogue and practical engagement. Furthermore, Rea-Dickens (1997) points out that if teachers are given opportunities through communication and making use of the materials to grasp the testing processes, they will become better test constructors. Moreover, learners can become better test takers.

Finally, autonomy of the teacher, teacher education and involvement of them to test design and use will not only contribute to the language testing process but also themselves to become better educators in every respect.

3.7.3 PARENTS

Participation of parents to testing process is necessary as well. Because, Rea-Dickens (1997) claims that parents should be kept informed and participate in the language testing process so that they can understand what assessment and testing involve and support their children.

Moreover, learners grow up in their familes. Therefore, parents know their children more than any one else. Parents are concerned about the future of their children. For this reason, they want to monitor their children’s progress and be informed about the events in the classroom. Most of the tests exclude the views of parents. Furthermore, they do not give information about the real performance of their children, because test scores may not always reflect the actual achievement of learners. For this reason, parents could be involved in the test design and receive instruction regarding test interpretation (Hamp-Lyons 2000).

On the other hand, the involvement of stakeholders to test development process and test use was argued. Weiss (1986: 144) states the concerns regarding stakeholder approach to language testing as follows:

> "The stakeholder approach holds modest promise. It can improve the fairness of the evaluation, democratise access to data, and equalise whatever power knowledge provides. However, it will not ensure that appropriate relevant information is collected nor increase the use of evaluation results."

Therefore, the role of stakeholders is under debate since the stakeholder approach might fail to contribute to the language learning as well as testing process. However, it is worth trying. Because, today democratization encourages the participation of everyone in making decisions. In addition, democratization of society starts at school. For this reason, stakeholder involvement will contribute a great deal in each area of education.

3.8 WASHBACK

"The impact of language tests on teaching and learning is called washback (backwash)" (McNamara 2000: 73). However, the term washback is not the only term formed to explain the relationship between testing and teaching and learning. The other terms are: curricular alignment, measurement driven instruction, and systemic validity (Hamp-Lyons 1997).

While 'curricular alignment' is a systems approach designed to develop and evaluate the curriculum (Dowd et. al 2007), 'measurement driven instruction' is an alternative method aimed at increasing the quality of education (Popham 1987 in Margheim 2001). Finally, Fredericksen and Collins (1989: 27 in Hamp-Lyons 1997:

295) state that “systemic validity is about the tests which bring about curricular and instructional changes encouraging the development of the cognitive skills that the test is designed to measure”.

Furthermore, Hamp-Lyons (1997: 297) indicates that “the term ‘washback’ is not used in the general education or educational measurement literature. The term ‘impact’ is used to refer, more broadly than is encompassed by ‘washback’, to the effects of high-stakes assessments”.

Alderson and Wall (1993: 121) accept that “washback can be positive or negative but question if washback is a concept which should be taken seriously or just a metaphor which encourages us to examine not only the role of tests in learning but also the relationships between teaching and testing”. Moreover, Hamp-Lyons (1997) states that the washback reflects the fears of teachers and how they avoid to carry out formal assessment.

Furthermore, the differences between test impact and washback were defined by Bachman and Palmer and McNamara: Test Impact refers to the effects of tests on macro-levels of education and society, and washback is concerned about the effects of language tests on micro-levels of language teaching and learning, i.e. inside the classroom (Bachman et al. 1996; McNamara 2000).

Bailey (1996: 261 in Hamp-Lyons 1997: 296) asserts that “positive washback is a primary goal for test developers and there are a number of factors contributing positive washback: Paralellism between tests and educational goals, genuineness of test tasks, greater self-assessment and learner autonomy and profile (detailed) score reporting.”

Alderson and Wall (1993: 116) also suggest that “the failure of a test to produce beneficial washback may not be due to problems in the test but to other forces which exist within society, education and schools, that might prevent washback from appearing”. Therefore, even though a test is considered to be appropriate when all the factors are taken into consideration both within the test itself and its implementation, the other factors such as forces in society might be the causes of negative washback.

The negative impact of tests which go beyond the classroom in the USA was stated in the following example:

> "There are two essential problems with standardized tests... First they fail to adequately measure important student learning. Even more important, their use has encouraged, or at least helped perpetuate, classroom practices that fail to provide high quality education, particularly for children from low income families. The reasons for this include:
>
> a) the multiple choice format
> b) norm-referencing
> c) making decisions using one test
> d) the use of these tests for accountability."
>
> (FairTest n.d.: 1 in Hamp-Lyons 1997: 288-289)

On the other hand, teachers may alleviate the negative effects of washback by changing assessment methods as Hill and Parry (1994b) state:

> "In the final analysis, a test forces students to engage in arbitrary tasks under considerable time pressure. It is for this reason that many educators have replaced testing...with [alternative] assessment practices that provide students with not only greater freedom to select the work on which they will be assessed but more extended periods of time in which to execute it."
>
> (Hill and Parry 1994b: 263 in Davies 1997: 335)

However, it is not certain that application of alternative assessment methods will provide positive washback. Hamp-Lyons (1997a) points out that although humanistic concerns led to the development of alternative assessment methods, she could not find real evidence when applying portfolio assessment to her classes. Instead, she proposes developing a logical model in order to examine the consequential validity of performance assessments (Hamp-Lyons 1997: 300).

On the other hand, Lynch (1997 in Lynch 1997: 324) argues that “if alternative assessment methods will be implemented to provide ethics, they will need to be validated with different procedures from those employed for traditional tests”.

Therefore, development of new methods to determine the validity of alternative assessment seems unavoidable. Moreover, if the positive washback cannot be achieved because of the factors within the society, and since the school is part of the society, the washback (backwash) studies should be focused on developing theories which fosters greater stakeholder involvement to the testing process.

Finally, the whole society is responsible in language testing process. But, the testers are primaly accountable for ethical testing practices because they are the ones who should provide positive washback when developing and using tests.

3.9 TEST BIAS

In language testing field language tests are constructed for particular purposes such as placement, diognosis, proficiency and etc. Tests have specific uses and they are carried out with specific groups of test takers. However, there may be other groups within these groups and there may be differences between them other than the language ability. These differences may affect the test performance of these groups as well as the validation process. This is called test bias (Bachman 1990).

According to Elder (1997: 261) there may be two reasons for the group differences:

> a) There is a real difference in the ability being tested (which may be attributed to factors outside the test-whether social, cultural or historical)
> b) There are confounding variables within the test (e.g., method effects, background knowledge) which systematically mask or distort the ability being tested.

Elder (1997) points out that confounding variables are responsible for test bias. Therefore, in order to prevent test bias the testers should determine these variables and try to minimize their effects in order to develop unbiased tests.

According to Bachman (1990: 255) bias is directly related to construct validity as he points out "Construct validity concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs".

The primary goal of tests is to measure an individual's language ability, however, for example, if the employment background of the test-taker affects his or her performance on the test, construct irrevelant variance might be thought to interfere in the measurement process (Messick 1989 in Wagner 2006). For this reason, a biased test cannot be considered a valid test.

Shohamy (1997) states that the studies showed that the 'test method' affects the performance of test takers as well. Test methods include the type of item, genre or testing tasks.

Shepard (1982; 1987 in Elder 1997: 263) points out that "bias detection requires the exercise of 'judgment' to determine whether observable differences in group performance are the result of measurement error or rather of real differences in the ability under test".

Since, learning therefore testing takes place in a learning environment where students gather who have different characterics, it is impossible to seperate the individuals in the classroom. Therefore, designing multiple types of tests and applying them may be a solution to the bias. The total score of these tests may facilitate obtaining more valid results and neutralize effects of test bias.

3.10 RESPONSIBILITY OF TESTERS IN LANGUAGE TESTING

Language testing primarily aims to provide validity. For this reason, what is valid can be considered ethical. As Hamp-Lyons (1989) points out ethics in language testing is a mixture of validity and backwash. She claims that the tester is responsible for the test use-the backwash, and therefore providing validity.

Davies (1997) points out that there should be limits regarding what can be achieved for a language tester. Therefore, he claims that a tester cannot be accountable for all consequences. In other words, a tester can be held accountable for the limited social consequences that he considers himself or herself responsible.

Bachman (2000) states his concerns about the responsibility of testers as follows:

> 1. Is it fair to expect practitioners to develop, administer and score tests when they have had little or no education or training in this?
> 2. Is it fair to require practitioners to write types of test items or tasks for which they have little or no training of experience?
> 3. Which applicants should be required to take a foreign language screening test for admission to a university? Who decides?
> 4. Of the students who have been admitted to the university, which ones should be required to take a foreign language placement test? Who decides?
> 5. Who is responsible for educating practitioners about fairness issues and considerations, and providing them the tools and knowledge needed to deal with these on a practical day-to-day basis? (Bachman 2000: 39-40)

As Hamp-Lyons (1997: 302) states "Language testers should accept responsibility for all those consequences we are aware of".

Spaan (2000) points out that it is impossible to develop totally fair tests but the test developers, test users and test takers can try to achive this by working together to make the test as valid, reliable and practical as possible. Therefore, stakeholders should work together in harmony by carrying out the duties they are required at their best.

Therefore, language testers are the ones who are primarily responsible for language testing practices. Stakeholder involvement and teacher education are essential to promote the fairness of the test. Stakeholders can be considered inspectors who can contribute to the language testing process.

3.11 STANDARDS FOR LANGUAGE TESTING

Test development and use is a process involving test preperation, test administration and interpreting test results. The test developers, administrators, and the people involved interpreting test results all have responsibilities to ensure fairness in each step of this process. However, the test administrators and the teachers are primarily responsible for conducting appropriate and fair tests.

Many factors should be taken into consideration in test development and use. These are practical, financial and political factors. Although the purpose of the test users may be the same, different testers can make use of different procedures and instruments due to different contexts. The term 'compromise' is used by Heaton (1988: 24 in Alderson et al. 1995) to define the give-and-take which is essential in test development process. Therefore, the testers should decide on the elements to include and exclude in test construction (Alderson et. al. 1995).

Bell (1994: 1) states that "any assessment practice that results in differences in assessment results that are not due to differences in student knowledge and skills affects the accuracy of the assessment itself, and thus undermines the decisions based upon those results". For this reason, it is of vital importance to determine the ethical standards for each step in the testing process to provide ethics in language testing.

Schmeiser (1995: 1) points out that many organizations developed standards for ethical test use. These organizations include Joint Committee on Testing Practices (1988), Joint Committee on Standards for Educational Evaluation (1988), National Association of College Admission Counselors (1988), American Federation of Teachers (1990), National Council on Measurement in Education (1990), National Education Association (1990).

Alderson et. al. (1995) also state the standards developed so far as follows: Standards for Educational and Psychological Testing (1985), Standards for Educational Testing Methods (1986), ETS Standards for Quality and Fairness (1987), Code of Fair Testing Practices in Education (1988), SEAC's Mandatory Code of Practice (1993), The ALTE Code of Practice (1994)

One of these standards is Code of Fair Testing Practices in Education which was developed by the Joint Committee on Testing Practices in 1988. This code addresses the responsibilities of professional test users and developers to test takers. The code focuses on the use of tests in education. However, it is not designed for classroom testing. In fact, it is intended to serve the needs of formal testing programs. The code states standards in four fields: Developing and selecting tests, interpreting scores, striving for fairness and informing test takers. The principles of Code of Fair Testing Practices in Education (1988) can be found in Appendix A.

The standards stated above mainly focuses on large scale assessments. However, the classroom testing which was administered by the teacher himself/herself should be taken into account in ethical language testing. Since, a large number of classroom tests are carried out at elementary, secondary, high schools and at universities. The Washington State Educational Assessment Program (WSEAP) developed ethical standards for test preperation and administration including the roles of teachers and principals (Appendix C).

3.12 CHAPTER SUMMARY

This chapter dealt with the concept of ethics, results of use of language tests for unintended purposes, models for an ethical test, the importance of stakeholders in the language testing process, washback, impact, responsibility of the test taker and standards for language testing.

CHAPTER FOUR
METHODOLOGY

4.0 INTRODUCTION

This chapter describes and discusses the methodology carried out in this study. It firstly discusses the rationale for why a questionnaire was used for data collection. Second, it restates the research questions and objectives of the study. Third, it describes the methodology including the research questions, pilot study, participants, data collection and analysis.

4.1 RATIONALE FOR QUESTIONNAIRE AS A DATA COLLECTING INSTRUMENT

There are many different types of data collecting methods involving questionnaire, interview, observation, and various experimental research design. The participants are asked the same questions under the same circumstances in survey studies which involve questionnaire, interview and observation. McMillan & Schumacher (1993) point out that 'questionnaire' is the most common way of collecting data. Because, it is easy and cheap to obtain data through a questionnaire and it saves time. However, Bell (1993: 75) states that "designing a questionnaire is very difficult in terms of selecting question type, writing the question, design, piloting, distribution and return of questionnaire". For this reason, the researcher should be careful in all stages of questionnaire design in order to obtain valid information and interprete the obtained data.

In this study the questionnaire is preferred as a data collecting instrument. Furthermore, the likert-type scale is employed. Because, likert-type scales are widely used in research and they enable the participants indicate their preferences through choosing the option which states their views the best (Oppenheim 1992). An example to a five-point likert scale is shown in figure 7. Moreover, a similar study was carried out in Poland to identify their views regarding a code of ethics for language test use. However, the items of the questionnaire implemented in Poland were open-ended. This study employs five point likert scale and open-ended questions to obtain data from the participants.

Strongly Agree	Agree	Neither agree or disagree	Disagree	Strongly Disagree
×				

Figure 7: An example to a likert-scale (McMillan & Schumacher 1993: 245)

4.2 OBJECTIVES AND RESEARCH QUESTIONS OF THE STUDY

This study aims to determine the views of instructors on ethics in language testing. Furthermore, this study aims to investigate the perceptions of instructors about a code of ethics or code of practice for language test use.

This study was carried out under the assumption that all the instructors believe that a code of ethics is necessary for language test use in Turkey and such a code should be formed considering the value systems of the Turkish culture.

In this study, it is also assumed that all the instructors agree on the basic concepts constituting the ethics concept and factors affecting ethics in language test use. The research questions of the study are as follows:

RQ 1: What is the test developers' and test users' level of making ethical choices in using English language tests?

RQ 2: What is the stakeholders' level of making ethical choices in using English language tests?

RQ 3: What is the stakeholders' level of responsibility in ethical use of English language tests?

RQ 4: 'Confidentiality' and 'access to information' are considered among the fundemental rights of test takers. What is the level of contribution of these rights to ethical use of English language tests?

RQ 5: What is the level of acceptance of language test givers to using language tests for non-intended purposes?

RQ 6: What is the level of awareness of language test givers to availability of a 'code of ethics' for language test use?

RQ 7: How much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests?

RQ 8: What is the level of acceptance of language test givers to unethical behaviors in language testing?

RQ 9: What do the test givers think about the content of a 'code of ethics' for language test use?

RQ 10: What do the test givers consider the unethical behaviors of a test giver?

RQ 11: What do the test givers consider the unethical behaviors of a test taker?

RQ 12: What do the test givers consider the qualities of a test giver?

RQ 13: What is the test givers level of considering value systems of the society in ethical use of English language tests?

RQ 14: Is there a significant diference between the views of the instructors and the members of the faculty holding PhD Degree with regard to;

a: What is the test developers' and test users' level of making ethical choices in using English language tests?

b: What is the stakeholders' level of making ethical choices in using English language tests?

c: What is the stakeholders' level of responsibility in ethical use of English language tests?

d: 'Confidentiality' and 'access to information' are considered among the fundemental rights of test takers. What is the level of contribution of these rights to ethical use of English language tests?

e: What is the level of acceptance of language test givers to using language tests for non-intended purposes?

f: How much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests?

g: How much do the test givers believe in the reasons for creating such a code for ethical use of language tests?

h: What is the level of acceptance of language test givers to unethical behaviors in language testing?

i: What do the test givers think about the content of a ‘code of ethics’ for language test use?

j: What do the test givers consider the unethical behaviors of a test giver?

k: What do the test givers consider the unethical behaviors of a test taker?

l: What do the test givers consider the qualities of a test giver?

m: What is the test givers level of considering value systems of the society in ethical use of English language tests?

4.3 DESIGN OF THE STUDY

The study consists of one pilot study and one main study. The pilot study and the main study are explained in the following sections.

4.3.1 PILOT STUDY

The pilot study was carried out to identify the possible problems of the data collecting instrument (questionnaire) of the study, correct the problems and make the necessary changes. The details regarding the pilot study are stated in the following section.

4.3.1.1 SETTING

The pilot study was conducted at ELT department in Çanakkale Onsekiz Mart University. Because, the researcher was holding post at the same department. The pilot study was carried out in a week during the spring semester of the 2006-2007 academic year.

4.3.1.2 PARTICIPANTS

The participants were three assistant professor doctors in ELT department. They were chosen since they were not only experienced instructors but also conducted a lot of theses up to now. Furthermore, the participants consist of one male and two females and all the participants were native speakers of Turkish.

4.3.1.3 INSTRUMENT AND PROCEDURES

The questionnaire was designed to collect data from the English language teachers about the ethics concept in language testing. Some items of the questionnaire were adapted from a questionnaire to collect information in order to develop a code of ethics for language test use from Polish examiners. The questionnaire is composed of 62 items. 33 questionnaire items were adapted from this questionnaire. The other items are created by the researcher. The questionnaire items adapted from this questionnaire are as follows:

Table 2: The questionnaire items adapted from a questionnaire carried out in Poland to develop a code of ethics for language test use.

Is there any need (necessity) of creating such a code of ethics? Please explain the reasons for your answer?
Yes, it would be useful
• to show the personal qualities of an examiner
• to explain unclear situations
• there should be some written document with regards to proper behaviour
No, it would not be useful
• Teachers on the whole are people of great moral standard and are conscious of their duties
• a 'code of professional ethics' exists
What should a code of ethics and professional conduct of an examiner contain?
• discipline
• norms of what is and what is not allowed
• examiners' duties
• what must be observed, rules of behaviour (general)
• rules of behaviour during the exam and ways of punishing dishonest examiners
What conduct of an examiner would you consider unethical?
• asking somebody else to mark the scripts an examiner was given (help)
• subjective marking
• being biased in assessment
• helping one's own students
• asking other examiners about some students' scripts, revealing results
• lack of confidentiality
What qualities should an examiner possess with regards to his/her personality?
• being objective
• honesty
• patience
• kindness
• unbiased
• punctuality
• discipline
• responsibility
• empathy
• understanding
• hardworking
• cooperative
• confidentiality

4.3.1.4 FINDINGS OF THE PILOT STUDY

The questionnaire in this study consists of three parts. In the first part, the definitions related to ethical concepts are stated. The second part of the questionnaire has 26 statements which reflect the opinions of English Language instructors about ethics concept in language testing. The final part has four questions written regarding personal information of the participants.

In the first part of the questionnaire, the terms reliability, validity, test bias, washback, test impact, test user, test developer and stakeholders are explained briefly. There are several reasons to define these terms. First, the participants will face these terms throughout the questionnare. Second, these specific terms are used widely in language testing field, and require specialized knowledge. Third, the participants of the study are not only the graduates of ELT department. The participants also involve graduates of English Language and Literature, American Language and Literature and Translation and Interpretation departments. Therefore, the graduates of other departments may have difficulty in understanding the statements appropriately while answering the questionnaire. Because, they may not have specialized knowledge in ELT field.

In the second part of the questionnaire, the questionnare items were a mixture of likert scale variables, and dichotomous questions but no open-ended question was included in this part of the questionnaire. There were 62 questions in total and 61 of them were 5-point likert scale type and 1 of them was a dichotomous question. The questionnaire was discussed with three assistant professors in ELT department to ensure not only the wording of the questions is clear and precise but also to provide face validity. For this reason, several changes were made to the overall design and the statements of the questionnaire. First, the order of parts was changed. Second, part 2 of the questionnaire was divided into four sections. Finally, some of the questions were excluded and rewritten to make them clear for the participants. The changes made to the questionnaire are as follows.

Part 1 in which the definitions were stated was located before the questions and their references were missing. Since most of the participants were graduates of ELT department, and it was assumed that they knew the specific terms used in the questionnaire the order of parts was reorganized and part 1 was located at the end of the questionnaire. Furthermore, references to the definitions were written.

Part 2 of the questionnaire was divided into 4 sections. The researcher decided that it would be easier for the participants to answer the questions when they were grouped in different sections. Because, the questions were addressing different aspects of the ethics concept. Section 1 includes general concepts regarding ethics in English test use. Section 2 is about the availability and content of 'a code of ethics' for language test use. The questions regarding unethical behaviors of a language test taker and test giver and the qualities of a language test giver were stated in section 3. Finally, the questions about the moral values of an individual were given in section 4.

Finally, some statements of the questionnaire were excluded and rewritten to make them clear for the participants. The differences between the original and the rewritten statements of the questionnaire are shown in the following table.

Table 3: The differences between the original and the rewritten statements of the questionnaire

Original Statements	Rewritten Statements
1. To what extent do you believe each of the following issues are important in language testing ethics?	
a. Reliability	**Part 1** 1. I believe that reliability is important in language testing ethics.
b. Validity	2. I believe that validity is important in language testing ethics.
c. Test bias	3. I believe that test bias is important in language testing ethics.

d. Washback	4. I believe that washback is important in language testing ethics.
e. Test Impact	5. I believe that test impact is important in language testing ethics.
3. Test scores cannot be interpreted without considering value systems of the society.	**Part 4** 1. Value systems of the society affect language test use in education
4. Individual rights and differences of test takers should be considered by test developers and users to ensure fairness in language testing.	2. Individual rights of test takers should be considered by test developers and users to ensure fairness in language testing.
	3. Individual differences of test takers should be considered by test developers and users to ensure fairness in language testing.
5. Language tests serve to maintain power and control.	NONE
7. It is appropriate to use tests for political purposes.	NONE
8. Stakeholder involvement in test use can contribute to the language testing process.	6. I believe that the stakeholder involvement in test preperation can contribute to the language testing process.
	7. I believe that the stakeholder involvement in test practice can contribute to the language testing process.
	8. I believe that the stakeholder involvement in test evaluation can contribute to the language testing process.
	9. I believe that the stakeholder involvement in making decisions can contribute to the language testing process.
9. It is impossible for a tester to be accountable for all possible consequences of language test use.	10. I believe that a tester should be accountable for all possible consequences of language test use.
13. To what extent do you believe each of the following parties are responsible for ethical test use in language programs?	
a. Test takers	11. I believe that test takers are responsible for ethical test use in language programs.
b. Families	12. I believe that families are responsible for ethical test use in language programs

c. Policy makers		13. I believe that policy makers are responsible for ethical test use in language programs		
d. Test users		14. I believe that test users are responsible for ethical test use in language programs.		
e. Test developers		15. I believe that test developers are responsible for ethical test use in language programs.		
14. To what extent do you believe each of the following parties are important with regard to rights of test takers in language testing ethics?				
a. Secrecy		NONE		
b. Confidentiality		**Part 1** 16. I believe that confidentiality is important with regard to the rights of test takers in language testing ethics		
c. Access to information		17. I believe that the right to access to information is important with regard to the rights of test takers in language testing ethics.		
18. There is a 'code of ethics' for language test use in Turkey.		**Part 2** 1. There is a 'code of ethics' for language test use in Turkey.		
YES	NO	YES	I DO NOT KNOW	NO
26. What should be the qualities of a language test administrator?		**Part 3** 3. What should be the qualities of a language test giver?		
a. objectivity		• objective		
b. honesty		• honest		
c. patience		• patient		
d. unbiased		• unbiased		
e. discipline		• disciplined		
f. responsibility		• responsible		
g. empathy		• empathetic		
h. kindness		• kind		
i. punctuality		• punctual		
j. understanding		• understanding		
k. hardworking		• hardworking		
l. ability to cooperate		• cooperative		

The data collecting instrument of the study (questionnaire) was developed by the researcher in the light of the literature review carried out by himself (Appendix D). Moreover, some statements of the questionnaire were adapted from a questionnaire which was designed by the British Council in Poland in order to develop a code of ethics for Polish examiners. This questionnaire is stated in Appendix E and the summary of the questionnaire responses is stated in Appendix F.

The researcher explained the aim of the research to the participants before giving the questionnaire and reminded them that the data collected will be kept secret and they will only be used for educational purposes. The questionnaires were delivered to the participants by the researcher and the questionnaires were completed in about ten days.

The data collected through the questionnaire in this study were analyzed by the program Statistical Package for the Social Sciences (SPSS).

4.3.1.5 IMPLICATIONS FOR THE MAIN STUDY

In the pilot study some statements of the questionnaire were rewritten and some statements were excluded to ensure the validity of their validity by the researcher.

4.3.2 MAIN STUDY

Main study was carried out after necessary changes were made on the questionnaire items.

4.3.2.1 SETTING

The main study was conducted at Çanakkale Onsekiz Mart University. Because, the aim of the research was to investigate the ethical concerns in language testing at university level, and the researcher was working as an instructor at Çanakkale Onsekiz Mart University. Therefore, Çanakkale Onsekiz Mart University was the most convenient university to carry out the research.

The main study was carried out in two weeks during the spring semester of 2006-2007 academic year.

4.3.2.2 PARTICIPANTS

The participants were assistant professors working in ELT department and instructors teaching English at various faculties in Çanakkale Onsekiz Mart University. All the assistant professors participated in the study were graduates of ELT department. However, the instructors participated in the study were graduates of different departments. The total number of participants and the distribution of participants with regard to the title and graduated department are shown in the following tables.

Table 4: Title distribution of the participants of the questionnaire in the main study

Title	**Number**
Assistant Professor	5
Instructor	23
Total	28

Table 5: Department distribution of the participants of the questionnaire in the main study

Department	**Number**
English Language Teaching	22
English Language and Literature	2
American Culture and Literature	1
Translation and Interpretation of English	3
Total	28

4.3.2.3 PROCEDURES FOR DATA COLLECTION

In the main study the data were collected through the data collecting instrument, questionnaire**.**

The questionnaires were delivered to the participants by the researcher himself or sent via e-mail. The aim of the questionnaire was explained to the participants in the first section of the questionnaire called 'purpose'. Moreover, the participants were told that the data collected would be kept secret and would only be used for educational purposes. Two weeks were allocated for participants to answer the questionnaire. Although, there were 45 participants to answer the questionnaire 28 of them returned them.

4.3.2.4 PROCEDURES FOR DATA ANALYSIS

The data collected through the questionnaire were analyzed by the SPSS program.

4.4 CHAPTER SUMMARY

This chapter described the rationale of why questionnaire was selected as a data collecting instrument. Next, the aim of the study and the research questions were stated. After that the pilot study and its findings were presented. Finally, the methodology of the main study was stated.

CHAPTER FIVE
FINDINGS

5.0 INTRODUCTION

This chapter deals with the findings of the data obtained from the main study. First, the research questions will be set and than the findings and the results of the analyses will be stated in the light of the research questions.

5.1 FINDINGS OF THE MAIN STUDY

The main study aims to investigate the ethical concerns in language testing at university level. Furthermore, it aims to investigate the perceptions of instructors about a code of ethics or code of practice for language test use.

The theoretical information regarding the study and the research questions to be dealt were explained in detail in the methodology part. Therefore, the findings of the study obtained through the research instrument (questionnaire) are explained in this chapter.

5.2 RESULTS OF THE QUESTIONNAIRE

RQ 1: What is the test developers' and test users' level of making ethical choices in using English language tests?

In order to find out the test developers' and test users' level of making ethical choices in using English language tests, descriptive statistics of general ethical concepts were carried out and means were calculated. The mean values to the general ethical concepts of the test developers and test users are shown in the following table.

Table 6: Descriptive statistics of the test developers' and test users' level of making ethical choices in using English language tests.

Items of the questionnaire	N	$\overline{X}$	SD
1. I believe that reliability is important in language testing ethics.	28	4.8214	,3900
2. I believe that validity is important in language testing ethics.	28	4.8214	,4756
3. I believe that test bias is important in language testing ethics.	28	4.2500	,7005
4. I believe that washback is important in language testing ethics.	28	4.2857	,6587
5. I believe that test impact is important in language testing ethics.	28	4.3214	,7724

According to the items of the questionnaire, there are five basic concepts constituting the ethics in language testing: Reliability, validity, test bias, washback and test impact. The mean values to the statements show that the participants think that the reliability and the validity are the most important concepts in language testing ethics (mean: 4.8214). In addition, test impact (mean: 4.3214) and washback (mean: 4.2857) are also important in language testing ethics. Although test bias (mean: 4.2500) has the least mean score of all, the participants believe that it is an important concept to provide ethics in language testing. Therefore, the findings suggest that all these concepts are essential elements of language testing ethics.

RQ 2: What is the stakeholders' level of making ethical choices in using English language tests?

In order to measure the stakeholders level of contribution to making ethical choices in language testing ethics descriptive statistics were carried out and means were calculated. The stakeholders level of contribution to making ethical choices in language testing ethics are shown in the following table.

Table 7: Descriptive statistics of the stakeholders level of making ethical choices in using English language tests.

Items of the questionnaire	N	$\overline{X}$	SD
6. I believe that the stakeholder involvement in test preperation can contribute to the language testing process.	28	3.9643	,7927
7. I believe that the stakeholder involvement in test practice can contribute to the language testing process.	28	3.6071	,9165
8. I believe that the stakeholder involvement in test evaluation can contribute to the language testing process.	28	3.5714	,8357
9. I believe that the stakeholder involvement in making decisions can contribute to the language testing process.	28	4.0714	,8133

The results of the descriptive statistics show that stakeholder involvement in making decisions (mean: 4.0714), test preperation (mean: 3.9643), test practice (mean: 3.6071), and test evaluation (mean: 3.5714) will contribute to the language testing process. According to the findings stakeholder involvement in making decisions can contribute to the language testing process the most. This might indicate that participants believe that decision making affects the test takers' lives more than the other steps in language testing process since all stakeholders are affected by the decisions made. Therefore, stakeholders have a right to involve in making decisions about individuals.

RQ 3: What is the stakeholders' level of responsibility in ethical use of English language tests?

In order to find out the stakeholders' level of responsibility in using English language tests descriptive statistics were carried out and means were calculated. The following table indicates the descriptive statistics regarding different types of stakeholders' level of responsibility in ethical use of English tests.

Table 8: Descriptive statistics of different types of stakeholders' level of responsibility in ethical use of English tests.

Items of the questionnaire	N	$\overline{X}$	SD
10. I believe that a tester should be accountable for all possible consequences of language test use.	28	4.2857	,9759
11. I believe that test takers are responsible for ethical test use in language programs.	28	4.0357	1,1701
12. I believe that families are responsible for ethical test use in language programs	28	2.5000	1,2019
13. I believe that policy makers are responsible for ethical test use in language programs	28	4.0000	,9759
14. I believe that test users are responsible for ethical test use in language programs.	28	4.0000	1,0709
15. I believe that test developers are responsible for ethical test use in language programs.	28	4.5000	1,1127

As it is seen in table 8 the different types of stakeholders mentioned in the study are testers, test takers, families, policy makers, test users and test developers. The findings indicate that the participants believe that testers (mean: 4.2857), test takers (mean: 4.0357), policy makers (mean: 4.0000), test users (mean: 4.0000) and test developers (mean: 4.5000) should be held accountable for the ethical use of English tests. Test developers are considered as the most responsible stakeholders in ethical test use by the participants. However, participants do not believe that families should be considered among the stakeholders responsible for ethical use of English tests (mean: 2.5000). In addition to this, while 6 out of the 28 (21.4 %) participants think families are responsible in ethical test use and 6 out of the 28 (21.4 %) participants do not have any idea, 16 out of 28 (57.1 %) participants are opposed to this (see Appendix G).

RQ 4: 'Confidentiality' and 'access to information' are considered among the fundemental rights of test takers. What is the level of contribution of these rights to ethical use of English tests?

When the descriptive statistics were carried out and means were calculated to find out the level of contribution of the rights 'confidentiality' and 'access to information' it is seen that participants believe that both confidentiality, (mean: 4.5357) and access to information (mean: 4.4286) contribute a great deal to ethical use of English tests. The following table states the results regarding the level of contributon of 'confidentiality' and 'access to information' to ethical use of English language tests.

Table 9: Descriptive statistics of the level of contribution of the rights 'confidentiality' and 'access to information' to ethical use of English tests.

Items of the questionnaire	N	$\overline{X}$	SD
16. I believe that confidentiality is important with regard to the rights of test takers in language testing ethics	28	4.5357	,6372
17. I believe that the right to access to information is important with regard to the rights of test takers in language testing ethics.	28	4.4286	,9201

RQ 5: What is language test givers level of acceptance to using language tests for non-intended purposes?

With the purpose of finding out the level of acceptance of language test givers to using English language tests for non-intended purposes descriptive statistics were carried out and means were calculated. The language test givers level of acceptance to using English language tests for non-intended purposes are shown in the following table.

Table 10: Descriptive statistics of language test givers level of acceptance to using language tests for non-intended purposes

Items of the questionnaire	N	$\overline{X}$	SD
18. Language tests are used for particular purposes such as assessment, diognosis etc. I believe that it is appropriate to use them for non-intended purposes. e.g. introducing a new education policy.	28	2.4286	1,1684

The results in the table indicate that most of the participants believe that it is inappropriate to use English test for non-intended purposes (mean: 2.4286).

RQ 6: What is the language test givers level of awareness to availability of a 'code of ethics' for language test use?

When the descriptive statistics were carried out and means were calculated to find out the language test givers level of awareness to availability of a 'code of ethics' for language test use it was found out that most of the participants do not believe that a code of ethics exists in Turkey (mean: 2.2857). That is, 4 out of 28 (14.3 %) participants think that there is a 'code of ethics' for language test use in Turkey and 14 out of 28 (50 %) participants do not have any idea if there is a 'code of ethics' for language test use in Turkey. Furthermore, 9 out of 28 (32.1 %) participants think that there is not a 'code of ethics' for language test use in Turkey (see Appendix G). The following table indicates the descriptive statistics regarding test givers level of awareness to availability of a 'code of ethics' for language test use.

Table 11: Descriptive statistics of language test givers level of awareness to availability of a 'code of ethics' for language test use.

Items of the questionnaire	N	$\overline{X}$	SD
1. There is a 'code of ethics for language test use' in Turkey.	28	2.2857	,8545

RQ 7: How much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests?

In order to answer this research question and to find out the reasons for creating such a code the descriptive statistics were carried out and means were calculated. The results are stated in the following table.

Table 12: Descriptive statistics of how much test givers believe in the necessity of creating a code or codes for ethical use of language tests and reasons for creating a code of ethics.

Items of the questionnaire	N	$\overline{X}$	SD
2. A 'code of ethics for language test use' is necessary to ensure fairness in language testing. Because;	28	4.4286	,8357
• appropriate behaviours should be stated in a written document.	28	4.4286	,7902
• unclear situations should be explained	28	4.6429	,6215
• the personal qualities of an examiner should be shown	28	2.7857	1,2578

The findings of the descriptive statistics show that participants believe that a 'code of ethics for language test use' is necessary to ensure fairness in language testing (mean: 4.4286). However, just 1 out of 28 (3.6 %) participants do not think such a code is necessary to ensure fairness in language testing (see Appendix G). Moreover, participants believe that such a code is necessary because appropriate behaviours should be stated in a written document (mean: 4.4286). Similarly, they think that a code of ethics is necessary because unclear situations should be explained (mean: 4.6429). However, nearly half of the participants do not think that the personal qualities of an examiner should be shown in order to create a code of ethics (mean: 2.7857). Therefore, it can be concluded that while stating the unclear situations is considered as the most important reason for creating such a code, showing personal qualities of an examiner is not supported by the participants.

RQ 8: What is the level of acceptance of language test givers to unethical behaviors in language testing?

In order to find out the level of acceptance of language test givers to unethical behaviors in language testing were carried out and means were calculated. The results are stated in the following table.

Table 13: Descriptive statistics of the level of acceptance of language test givers to unethical behaviors in language testing.

Items of the questionnaire	N	$\overline{X}$	SD
• A 'code of ethics for language test use' is not necessary for fairness in language testing, because teachers are honest people and aware of their responsibilities.	28	2.4643	1,1701
• A 'code of ethics for language test use' is not necessary for fairness in language testing, because there is already a code of professional ethics.	28	2.6071	1,2274

As it is seen in the table 13, most of the participants do not believe that teachers are honest and aware of their responsibilities (mean: 2.4643). Similarly, they do not agree on that there is already a code of professional ethics in language testing (mean: 2.6071). Furthermore, 35.7 % of participants are undecided in these issues (see Appendix G).

RQ 9: What do the test givers think about the content of a 'code of ethics' for language test use?

Next, the test givers thoughts about the content of a 'code of ethics' for language test use were sought through descriptive statistics The results are stated in the following table.

Table 14: Descriptive statistics regarding the test givers thoughts about the content of a 'code of ethics' for language test use.

Items of the questionnaire	N	$\overline{X}$	SD
5. What should 'a code of ethics for language test use' include?			
• rules regarding what is and what is not allowed	28	4.6786	,4756
• examiners' responsibilities.	28	4.3214	1,0203
• rules regarding how to punish dishonest examiners	28	3.8571	1,2971
• appropriate rules of behaviour during the exam.	28	4.5714	,6901
• discipline	28	3.8571	1,2084

The table shows that a code of ethics for language test use' should include rules regarding what is and what is not allowed (mean: 4.6786), examiners' responsibilities (mean: 4.3214), rules regarding how to punish dishonest examiners (mean: 3.8571), appropriate rules of behaviour during the exam (mean: 4.5424), and discipline (mean: 3.8571). Therefore, rules regarding what is and what is not allowed and appropriate rules of behaviour during the exam are the most important elements that should be included in a code of ethics. However, discipline and rules concerning how to punish dishonest examiners are not considered as important as the other statements by the participants.

RQ 10: What do the test givers consider the unethical behaviors of a test giver?

In order to determine the unethical behaviors of a test giver mean values were calculated. The following table indicates the findings.

Table 15: **Descriptive statistics regarding the test givers thoughts about the unethical behaviors of a test giver**

Items of the questionnaire	N	$\overline{X}$	SD
1. What do you consider to be the unethical behaviors of a test giver?			
• asking someone else to mark the testee's paper.	28	4.2143	1,2280
• lack of confidentiality	28	4.4286	1,0338
• marking the papers subjectively	28	4.6786	,8630
• helping one's own students	28	4.5000	1,1706

The results of the descriptive statistics indicate that subjective marking of the papers and helping one's own students are considered as the most important unethical behaviors of a test giver for the participants (mean: 4.6786, mean: 4.5000 respectively). Furthermore, participants believe that lack of confidentiality is an unethical behavior of a test giver (mean: 4.4286). However, asking someone else to mark the testee's paper is the least important inappropriate behavior of a test giver among the other statements (mean: 4.2143).

RQ 11: What do the test givers consider the unethical behaviors of a test taker?

Descriptive statistics were carried out in order to determine what the test givers consider the unethical behaviors of a test taker. The results are shown in the following table.

Table 16: Descriptive statistics regarding the test givers thoughts about the unethical behaviors of a test taker

Items of the questionnaire	N	$\overline{X}$	SD
2. What do you consider the unethical behaviors of a test taker?			
• cheating during the exam	28	4.7857	,7868
• helping other student(s) during the exam	28	4.6429	,8698
• interrupting other students during the exam	28	4.4286	,9974
• unpunctuality (not arriving at the exam venue on time).	28	3.9643	1,1380

According to the findings stated in the table above, cheating and helping other student(s) during the exam are considered as the most important unethical behaviors of a test taker by the participants (mean: 4.7857, mean: 4.6429 respectively). Participants also believe that interrupting other students during the exam is an unethical behavior of a test taker (4.4286). However, unpunctuality is not seen as important unethical behavior of a test taker when compared to the others (mean: 3.9643).

RQ 12: What do the test givers consider the qualities of a test giver?

In order to find out what the test givers consider the qualities of a test giver descriptive statistics were carried out. The findings are stated in the following table.

Table 17: Descriptive statistics regarding the test givers thoughts about the qualities of a test giver.

Items of the questionnaire	N	$\overline{X}$	SD
3. What should be the qualities of a language test giver?			
• objective	28	4.9286	,3780
• honest	28	4.8214	,3900
• patient	28	4.3929	,7860
• unbiased	28	4.9286	,2623
• disciplined	28	4.2857	1,0838
• responsible	28	4.7500	,5182
• empathetic	28	4.2143	,8759
• kind	28	4.0714	1,0862
• punctual	28	4.6786	,6118
• understanding	28	4.2500	,9670
• hardworking	28	4.1071	1,0306
• cooperative	28	4.5000	,9230

The findings indicate that a language test giver should be objective (mean: 4.9286), honest (mean: 4.8214), patient (mean: 4.3929), unbiased (mean: 4.9286), disciplined (mean: 4.2857), responsible (mean: 4.7500), empathetic (mean: 4.2143), kind (mean: 4.0714), punctual (mean: 4.6786), understanding (mean: 4.2500), hardworking (mean: 4.1071), and cooperative (mean: 4.5000). Objectivity, being unbiased, honesty and responsibility are considered as the most important qualities of a test giver by the participants. On the other hand, being kind, hardworking, disciplined and empathetic are not accepted as valuable as the other stated qualities of a test giver.

RQ 13: What is the test givers' level of considering value systems of the society in ethical use of English language tests?

Descriptive statistics were carried out in order to find out the test givers' level of considering value systems of the society in ethical use of English language tests. The following table indicates the results.

Table 18: Descriptive statistics of the test givers level of considering value systems of the society in ethical use of English language tests

Items of the questionnaire	N	$\overline{X}$	SD
1. Value systems of the society affect language test use in education.	28	4.1429	,8909
2. Individual rights of test takers should be considered by test developers and users to ensure fairness in language testing.	28	4.3571	,7310
3. Individual differences of test takers should be considered by test developers and users to ensure fairness in language testing.	28	3.7857	1,1339
4. There are different standards for different examiners.	28	3.2143	1,2869
5. The moral values vary from one culture to another.	28	3.8929	1,3427
6. The moral values vary from one situation to another.	28	3.1786	1,3892
7. The moral values vary from one individual to another.	28	3.3571	1,3113
8. Adopting the 'code of ethics for language test use' of another country is useful, since different cultures have common ethical values.	28	2.7857	1,3705

According to the findings stated in table 18, considering the individual rights of the test takers is considered as the most important factor to provide ethics in language testing (mean: 4.3571). Participants also believe that value systems of the society affect language test use in education a great deal (mean: 4.1429). Furthermore, individual differences of test takers should be taken account by test developers and users according to participants (mean: 3.7857). They also state that the moral values vary from one culture to another, from one situation to another and from one individual to another (mean: 3.8929, mean: 3.1786, mean: 3.3571 respectively). Therefore, it might be concluded that the culture is the most effective element on moral values. In addition, participants think that there are different standards for different examiners (mean: 3.2143). However, most of the participants do not agree that adopting the 'code of ethics for language test use' of another country is useful. As a result, since participants believe there are different standards for different examiners, a code of ethics is necessary to determine the standards for all examiners. Furthermore, the cultural values of Turkish society should be taken into consideration when creating a code of ethics for language test use.

RQ 14: Is there a significant difference between the views of the instructors and the members of the faculty holding PhD degree with regard to;

a) **What is the test developers' and test users' level of making ethical choices in using English language tests?**

Table 19: T-test of test developers' and test users' level of making ethical choices in using English language tests.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4,4957	,33504	-,138	26	,229
Asst. Prof. Dr.	5	4,5200	,46043			

$P > 0.05$

As can be seen in the table above there is no significant difference between the views of instructors and faculty members holding PhD degree (mean: 4.4957, mean: 4.5200 respectively). Therefore, holding a PhD degree and being an instructor does not make any difference concerning test developers' and test users' level of making ethical choices in using English language tests (p > 0.5).

b: What is the stakeholders' level of making ethical choices in using English language tests?

Table 20: T-test of the stakeholders' level of making ethical choices in using English language tests.

Title	N	$\bar{X}$	SD	t	df	Sig.
Instructor	23	3,8370	,63339	,599	26	,832
Asst. Prof. Dr.	5	3,6500	,62750			

P > 0.05

As for question 14/b, there is no significant difference between the views of instructors and faculty members holding PhD degree (mean: 3.8370, mean: 3.6500 respectively). Therefore, there is no relationship between holding a PhD degree and being an instructor regarding the stakeholders' level of making ethical choices in using English language tests (p > 0.5).

c: What is the stakeholders' level of responsibility in ethical use of English language tests?

Table 21: T-test of the stakeholders' level of responsibility in ethical use of English language tests.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	3,7536	,54093	-,662	26	751
Asst. Prof. Dr.	5	3,9333	,59628			

P > 0.05

In question 14/c, results obtained through t test show that instructors and faculty members holding PhD degree agree on the stakeholders' level of responsibility in ethical use of English language tests (mean: 3.7536, mean: 3.9333 respectively). Therefore, holding a PhD degree and being an instructor does not make any difference concerning the stakeholders' level of responsibility in ethical use of English language tests (p > 0.5).

d:'Confidentiality' and 'access to information' are considered among the fundemental rights of test takers. What is the level of contribution of these rights to ethical use of English language tests?

Table 22: T-test of the level of contribution of rights ('Confidentiality' and 'access to information') of test takers to ethical use of English language tests.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4.4348	,75835	-,747	26	,213
Asst. Prof. Dr.	5	4,7000	,44721			

P > 0.05

As it can be seen in table 22 there is no significant difference between the views of instructors and faculty members holding PhD degree (mean: 4.4348, mean: 47000 respectively). Therefore, the participants share the same views regarding the level of contribution of 'confidentiality' and 'access to information' rights to ethical use of English language tests (p > 0.5).

e: What is the level of acceptance of language test givers to using language tests for non-intended purposes?

Table 23: T-test of the level of acceptance of language test givers to using language tests for non-intended purposes.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	3,5652	1,19947	-,059	26	,801
Asst. Prof. Dr.	5	3,6000	1,14018			

P > 0.05

As for question 14/e, the findings obtained through t test indicate that the views of instructors and faculty members holding PhD overlap regarding the level of acceptance of language test givers to using language tests for non-intended purposes (mean: 3.5652, mean: 36000 respectively). That is, there is no relationship between holding a PhD degree and being an instructor regarding the level of acceptance of language test givers to using language tests for non-intended purposes (p > 0.5).

f: How much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests?

Table 24: T-test of how much do the test givers believe in the necessity of creating a code or codes for ethical use of language tests.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4,3913	,89133	-,499	26	,234
Asst. Prof. Dr.	5	4,6000	,54772			

P > 0.05

As can be seen in the table above, there is no significant difference between the views of instructors and faculty members holding Phd degree regarding the necessity of creating a code or codes for ethical use of language tests (mean: 4.3913, mean: 4.6000). Therefore, holding a PhD degree and being an instructor does not make any difference concerning the necessity of creating a code or codes for ethical use of language tests (p > 0.5).

g: How much do the test givers believe in the reasons for creating such a code for ethical use of language tests?

Table 25: T-test of how much do the test givers believe in the reasons for creating such a code for ethical use of language tests.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	3,9420	,59163	-,197	26	,534
Asst. Prof. Dr.	5	4,0000	,62361			

P > 0.05

In question 14/g, findings obtained through t test show that instructors and faculty members holding PhD degree share the same views regarding the reasons for creating such a code for ethical use of language tests (mean: 3.9420, mean: 40000 respectively). That is, there is not a significant difference between the views of the instructors and faculty members holding a PhD degree concerning the reasons for creating such a code for ethical use of language tests (p > 0.5).

h: What is the level of acceptance of language test givers to unethical behaviors in language testing?

Table 26: T-test of the level of acceptance of language test givers to unethical behaviors in language testing.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	2,4783	,98256	-,700	26	,350
Asst. Prof. Dr.	5	2,8000	,57009			

P > 0.05

As for question 14/h, there is no significant difference between the views of instructors and faculty members holding PhD degree (mean: 2.4783, mean: 2.8000 respectively). Therefore, the participants have similar views regarding the level of acceptance of language test givers to unethical behaviors in language testing (p > 0.5).

i: What do the test givers think about the content of a 'code of ethics' for language test use?

Table 27: T-test of the test givers thoughts about the content of a 'code of ethics' for language test use.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4,2522	,61561	-,094	26	,169
Asst. Prof. Dr.	5	4,2800	,50200			

P > 0.05

In question 14/i, results obtained through t test indicate that instructors and faculty members holding PhD degree agree on the content of a 'code of ethics' for language test use (mean: 4.2522, mean: 4.2800 respectively). That is, there is no difference between the views of instructors and faculty members holding PhD degree concerning the content of a 'code of ethics' for language test use (p > 0.5).

j: What do the test givers consider the unethical behaviors of a test giver?

Table 28: T-test of the test givers thoughts about the unethical behaviors of a test giver.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4,4348	,91147	-,271	26	,617
Asst. Prof. Dr.	5	4,5500	,51235			

$P > 0.05$

As it is seen in table 28, there is no significant difference between the views of instructors and faculty members holding PhD degree regarding the unethical behaviors of a test giver (mean: 4.4348, mean: 4.5500 respectively). Therefore, holding a PhD degree and being an instructor does not make any difference concerning the the unethical behaviors of a test giver ($p > 0.5$).

k: What do the test givers consider the unethical behaviors of a test taker?

Table 29: T-test of the test givers thoughts about the unethical behaviors of a test taker.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4,4239	,87072	-,436	26	,443
Asst. Prof. Dr.	5	4,6000	,41833			

$P > 0.05$

As for question 14/k, results obtained through t test indicate that instructors and faculty members holding PhD degree share the same views regarding the unethical behaviors of a test taker (mean: 4.4239, mean: 4.600 respectively). That is, there is no difference between holding a PhD degree and being an instructor concerning the the unethical behaviors of a test taker ($p > 0.5$).

l: What do the test givers consider the qualities of a test giver?

Table 30: T-test of the test givers thoughts about the qualities of a test giver.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	4,5072	,41584	,345	26	,506
Asst. Prof. Dr.	5	4,4333	,52507			

P > 0.05

As can be seen in the table above, there is no significant difference between the views of instructors and faculty members holding PhD degree regarding the qualities of a test giver according to the findings (Mean: 4. 5072, mean: 4. 4333 respectively). Therefore, holding a PhD degree and being an instructor does not make any difference concerning the the qualities of a test giver (P > 0.05).

m: What is the test givers level of considering value systems of the society in ethical use of English language tests?

Table 31: T-test of the test givers level of considering value systems of the society in ethical use of English language tests.

Title	N	$\overline{X}$	SD	t	df	Sig.
Instructor	23	3,7065	,71966	1,914	26	,460
Asst. Prof. Dr.	5	3,0500	,54199			

P > 0.5

In question 14/m, results obtained through t test indicate that instructors and faculty members holding PhD degree have similar views regarding test givers level of considering value systems of the society in ethical use of English language tests (Mean: 3.7065, mean: 3.0500 respectively). Therefore, holding a PhD degree and being an instructor does not make any difference concerning test givers level of considering value systems of the society in ethical use of English language tests ($p > 0.5$).

5.3 CHAPTER SUMMARY

This chapter has given the main findings and statistical analysis of the study conducted.

CHAPTER SIX
DISCUSSION, CONCLUSION AND IMPLICATIONS

6.0 INTRODUCTION

In this chapter, firstly aim of the study will be stated. Then the findings of the study will be discussed in the light of the literature review. Next, implications of the study will be discussed. Finally, suggestions for further research will be made.

6.1. AIM OF THE STUDY

This study aims at determining the views of ELT Department instructors on ethical issues regarding English language test use. Moreover, it also investigates the roles of stakeholders in ethical test use and the importance of a 'code of ethics' in language testing.

6.2 SUMMARY OF THE METHODOLOGY

A questionnaire was implemented for the purpose of the study. The questionnaire was formed of items to elicit evidence regarding the ELT instructors' opinions of ethical test use at university level. The questionnaire consists of four parts. First part is about the basic ethical qualities of test reliability, validity, test bias, washback, impact and stakeholder involvement in testing process. Second part refers to a code of ethics in language testing and the content of this code. Third part is about the qualities of a test giver. The fourth part is regarding the value systems of the society and individual characteristics of test takers.

6.3 DISCUSSIONS

The purpose of this study is to determine the perceptions of ELT instructors regarding the ethical test use in language programs and a code of ethics in language testing.

The first part of the questionnaire focus on the basic qualities of a language test, the stakeholders roles and responsibilities in ethical use of language tests, the rights of test takers and use of tests for non-intended purposes.

The findings of the RQ 1 indicate that the participants agree on that reliability, validity, washback, test bias and test impact are the fundemental qualities of a language test. Similarly, the ethical language test models submitted by Bachman et al. (1996), Kunnan (2000), and McNamara (2000) state that these are the most important qualities that a language test should possess in order to be considered ethical.

According to the results obtained from the RQ 2 regarding the stakeholders role in ethical language test use indicate that stakeholder involvement will contribute a great deal to all stages of test development and use. Therefore, stakeholders (test users, test developers, policy makers, test takers and test givers) form an essential part of ethical language test use.

The findings of the RQ 3 state that participants consider the stakeholders are responsible for ethical language test use. However, 57.1 % of participants do not believe that families are responsible for ethical test use in language programs. Therefore, they do not consider families as one of the stakeholders that may contribute to the language testing process. This can be the result of the social structure of Turkish society. For example, in Turkey, families do not tend to donate to official institutions and expect financial aid from the government. This attitude towards the government is prevalent in Turkey unlike developed countries such as the USA. Test practices are implemented at school environment and families generally do not want to interfere in the processes carried out at school or official institutions in Turkey. For this reason, participants do not believe that parents are responsible for ethical language testing process.

According to the findings of RQ 4 the participants attach importance to individual rights (confidentiality- 92.8 %, and access to information- 78.6 %) of the test takers. Similarly, the findings of the study carried out in Poland which aims at learning the perceptions of Polish examiners about the content of a code of ethics in language testing overlap the findings of this study. The Polish test givers (27 participants out of 96- the highest score among 18 different responses) believe that 'not treating the work confidentiality' is the most important unethical behavior (see Appendix F).

The findings of the RQ 5 indicate that the use of tests for non-intended purposes is not considered appropriate (60.8 %) by Turkish teachers.

Part two of the questionnaire is about the existence of a code of ethics for ethical test use, necessity of creating such a code, and the content of a code of ethics.

The findings of the RQ 6 show that 50 % of the participants have no idea if there is a 'code of ethics for language test use' in Turkey. 32.1 % of the participants do not think there is such a code and only 17.9 % of them agree that there is a code of ethics for language test use. However, there is not a code of ethics for language testing in Turkey. Therefore, 67.9 % of participants are not aware if such a code exists or not. However, each teacher should follow the developments in education and know about the laws and legislations regarding his or her profession. This can also be the result of the application of codes in Turkey. Because, although there are laws, and codes in Turkey they are not implemented thorougly as it is in other countries such as the USA.

When the participants were asked if such a code is necessary in RQ 7, 85.7 % of the participants state that such a code is necessary. The findings also indicate that showing appropriate behaviors and explaining unclear situations are of vital importance to create a code of ethics. However, just 28.6 % of participants believe that personal qualities of an examiner should be shown in such a code and 42.9 % of participants are opposed to this idea in Turkey. These findings are supported by the findings of the study carried out in Poland about the content of a code of ethics. 84 % of Polish teachers state that they need such a code. Furthermore, the statement 'unclear situations should be explained' is the most important reason for Polish examiners (6 participants) as for the Turkish colleagues (92.8 %). In addition, the statement 'appropriate behaviors should be stated in a written document' is one of the most supported statements by Polish examiners (4 participants). Finally, the statement 'the personal qualities of an examiner should be shown' was only stated by 2 Polish examiners which is one of the least supported statements (Appendix F). As for this final statement, the participants perceptions about 'personal qualities' of a test giver should be examined. Because, in RQ 12 the qualities of a test giver were asked to the participants and objectivity, honesty, patience and etc. (see Table 16) were stated as these qualities. For this reason, why they are opposed to showing personal qualities of an examiner in a code of ethics should be discussed. There may be a number of reasons leading these results. First, the participants might have thought that it is not necessary because whatever the personal qualities of a test giver are each test giver would treat each student equally in language test use. Second, the nature of teaching profession might have led this because, teachers want to be autonomous in the classroom and school. For this reason, the participants might have considered this question as violating their confidentiality and private life since they may not want to share their personal qualities with other people. Third, what is right changes person to person, situation to situation. Therefore, personal qualities of an examiner may change from person to person and situation to situation as well. For this reason, 'stating teachers' personal qualities' might not affect the test givers' testing practices.

The findings of the RQ 8 regarding why a code of ethics is not necessary showed that 21.4 % of participants believe that there is already a code of ethics. However, there is not such a code in Turkey. It should be questioned whether these participants consider the discipline rules as a code of ethics. Moreover, 4.6 % of Polish examiners share the same views as the Turkish counterparts. The difference between the percentages might mean that Polish examiners attach more importance to the issues about teaching profession. The responses given to the statement 'teachers are honest people and aware of their responsibilities' indicate that 50 % of the participants in Turkey do not agree with this statement. The findings of the study carried out in Poland overlap with the findings of this study. It can be concluded that a code of ethics should also contain the responsibilities of a teacher and such a code is necessary not only for language test use but also for all testing practices.

The RQ 9 is about the content of a code of ethics. Both the Turkish participants (100 %) and Polish examiners (25 %) stated that 'rules regarding what is and what is not allowed' is the most important concern in a code of ethics. Turkish examiners indicated that 'discipline' is the least important issue to be included in a code of ethics (64.3 %). Similarly, discipline (2 participants), professionalism (2 participants), exam regulations (1 participant), kinderstube (2 participants), directions of what an examiner should not do (2 participants) and qualities of an examiner (2 participants) are the least supported statements to be written in a code of ethics by Polish examiners.

The RQ 10 is regarding the unethical behaviors of a test giver. The findings of this research question state that Turkish test givers believe that 'marking the papers subjectively' (92.8 %) is the most important responsibility of an examiner. The findings of this study overlap with the findings of the study carried out in Poland. Polish examiners indicate that subjective marking of the papers is one of the most important unethical behavior of a test giver (8 participants). Furthermore, when the Polish examiners were asked 'what are the duties of an examiner with regards to a student he/she assesses' they stated that 'being objective' is the most important duty of an examiner (29 participants). According to statement 'not treating the work confidentially' the Polish test givers (27 participants) and Turkish counterparts agree on this statement (85.4 %). Finally, the findings of the statements 'asking someone else to mark the testee's paper' and 'helping one's own students' indicate that these behaviors are considered unethical both by Turkish and Polish test givers.

When the participants are asked the unethical behaviors of a test taker in RQ 11, the findings state that 'cheating during the exam' is the most important concern of them (96.4 %). Furthermore, participants agree on that 'helping and interrupting the other students during the exam' should be considered as unethical behaviours of a test taker. However, they believe the least important one is 'unpunctuality' of a test taker (67.9 %).

The findings of the RQ 12 which is regarding the qualities of a test giver overlap the findings of the study carried out with Polish examiners. While 'honesty' (100 %), and 'objectivity' (96.4 %) are the fundemental qualities of a test giver for Turkish examiners 'objectivity' (18 participants), and 'honesty' (14 participants) are the basic qualities of a test giver for Polish examiners as well.

The RQ 13 is about the test givers' level of considering value systems of the society. The participants state that consideration of individual rights of the test takers by test developers and test users to ensure fairness in language testing is the most important concern (85.7 %). Therefore, the participants attach importance to individual rights of test takers. Moreover, they indicate that value systems of the society affect language test use (82.2 %). In addition, more than half of the participants think that moral values vary from culture to culture (78.6 %), situation to situation (50 %) and individual to individual (57.2 %). Also, most of the participants disagree to the statement 'adopting the code of ethics of another country is useful, since different cultures have common ethical values' (42.9 %). On the other hand, 35.7 % of participants agree to this statement. Therefore, moral values of the Turkish culture should be taken into consideration when creating a code of ethics for language test use. Moreover, 46.7 % of participants state that there are different standards for different examiners. Therefore, the participants believe that subjectivity in language test use is common among examiners. For this reason, the standards should be determined for language test use and this can be accomplished through creating a code of ethics.

Finally, findings of the RQ 14 indicate that there is not a significant difference between the views of instructors and faculty members holding PhD degree with regard to the questionnaire items.

6.4 CONCLUSIONS

The first part of the study showed that the participants consider the reliability, validity, test bias, washback and test impact as the fundemental qualities of a language test. Moreover, they think that stakeholder involvement in all stages of test development and use (test preperation, test practice, test evaluation and making decisions) would contribute to the language testing process. Furthermore, the participants believe that the stakeholders (test givers, test takers, policy makers, test users and developers) should be held accountable for ethical test use in language programs. However, they do not believe that families are responsible for ethical test use in language programs. When the testers were asked about the level of contribution of rights of test takers (confidentiality and right to access to information) they agreed that these rights are of vital importance in ethical language testing. Finally, the participants think that using language tests for non-intended purposes is not appropriate.

The second part of the study indicated that participants are not aware of that if a code of ethics for language test use exists or not. However, they believe that such a code is necessary. When they are asked the reasons for the necessity of a code of ethics they stated that appropriate behaviors should be stated in a written document and unclear situations should be explained. On the other hand, most of the participants believe that showing the personal qualities of an examiner should not be counted among the reasons to create such a code. According to the question regarding why a code of ethics is not necessary, half of the participants indicated that the teachers are not honest people and they are not aware of their responsibilities. Furthermore, nearly half of them opposed to the idea that there is already a code of ethics.

The third part of the questionnaire showed that participants believe that rules regarding what is and what is not allowed, examiner's responsibilities, rules about how to punish dishonest examiners, appropriate rules of behavior during the exam and discipline should be included in a code of ethics for language test use. Participants also stated that asking someone else to mark the test taker's paper, lack of confidentiality, marking the papers subjectively and helping one's own students are the unethical behaviours of a test giver. On the other hand, participants believe that cheating, helping other students, interrupting other students during the exam and unpunctuality are the unethical behaviors of a test taker. Finally, they think that a language test giver should be objective, honest, patient, unbiased, disciplined, responsible, empathetic, kind, punctual, understanding, hardworking and cooperative to be considered a good examiner.

The last part of the questionnaire was about the value systems of the society in ethical language test use. The participants stated that value systems of the society affect language test use. Furthermore, individual rights and differences should be taken into consideration to ensure fairness in testing. The participants opposed to adopting the code of ethics of another country because they believe that each culture has different ethical values and moral values vary from situation to situation and individual to individual. In addition, most of the participants indicated that each examiner has different standards.

Finally, the study also shows that there is not a significant difference between the views of instructors and members of the faculty holding PhD regarding the questionnaire items.

When all the results are taken into account developing a code of ethics in language testing is necessary, and this process should be implemented through the participation of all stakeholders. Furthermore, although the Turkish and Polish testers agree on the basic test and examiner qualities, the ethical values peculiar to Turkish culture are of vital importance in creating a code of ethics for language test use.

6.5 IMPLICATIONS

Only a few studies were carried out regarding the code of ethics in language testing field so far. Furthermore, no study was carried out in Turkey on this issue. In the light of the study carried out, it will be useful to create a code of ethics to determine the standards for language test use. Schmeiser (1995: 3) points out that "codes serve to increase the awareness of ethical practice among their memberships and to promote ethical uses of assessment in various contexts: teaching, counseling, evaluation, and research". Therefore, language teaching which involves assessment practices is one of the fields in which a code should be formed. However, a code for language testing does not exist in Turkey. In fact, there are academic ethics boards in Turkish universities such as Yıldız University Academic Ethics Board and student discipline regulations which involve discipline rules for teaching practices and exams. 'Student Discipline Regulations' of Eastern Mediterrenean University is an example to the student discipline regulations. While academic ethics board refers to actions of academicians such as plagiarism and piracy, student discipline regulations are mainly concerned about the discipline rules that the students should obey in the university campus. Furthermore, Ministry of National Education has regulations such as 'Reward and Discipline Regulation of the Secondary Education Institutions'. However, these regulations do not address the ethical issues involved in language testing other than the discipline rules to be implemented during an examination.

Language tests mainly serve to make decisions about individuals (Bachman et al. 1996). Therefore, the test takers are affected the most among the other stakeholders. For example, the unethical testing practices might have dire consequences regarding the future of individuals. For this reason, the standards for language test use or a code of ethics should be created to ensure fairness in language testing field.

Writing of a code of ethics for language test use is a demanding process since test developers, test users, the families, and the official bodies have stakes in language test use. Therefore, the purpose, the content, and the enforcement of a code of ethics are of vital importance.

The process of creating the a model code of ethics for language test use is as follows:

Step 1: Collecting information regarding the testing practices of the different countries. The countries to be selected should have codes or regulations regarding the ethical test use in language programs. This step involves consulting the experts who carry out both large-scale and classroom examinations. The experts should involve the experienced teachers, teacher trainers, and professionals. The examination systems and the implication of the codes should be examined thoroughly.

Step 2: The relationship between the value systems of these countries and the effects of these value systems on creating the code of ethics should be examined in all aspects. For example, to what extent the value systems affect the language testing process? Do different countries implement different ethical codes? If these codes are different what are the reasons? Is the code of ethics applied locally or to all institutions across the country?

Step 3: The purpose of the code should be clearly identified in a policy statement. For example, the aim of the code of ethics in language testing is to provide ethical language practices in educational contexts.

Step 4: The content of the code is of vital importance as well. The code should be prepared considering the value systems of the Turkish society and the needs of the stakeholders involved in language testing process. Furthermore, it should determine the responsibilities of test developers, test users and test takers.

Step 5: The code might be prepared with the participation of a wide range of stakeholders ranging from learners to the government bodies. However, it might not be necessary to involve all the stakeholders. For example, the stakeholders may include test developers, test users, test takers, families and policy makers.

Step 6: Each university having ELT department should carry out a study to learn the perceptions of test developers, test users and test givers regarding a code of ethics. A joint committee should be formed including representatives from universities and policy makers.

Step 7: The rules and principles of a code of ethics should be determined so that a draft code could be prepared. The responsibilities of test takers, test givers, test developers, the school administrators and policy makers should be clearly defined in the code.

Step 8: There should be two codes. First code should be intended for large-scale assessments and second code should address the needs of classroom assessments. Code of Fair Testing Practices in Education (Appendix A) intended for large-scale assessments and Ethical Testing Standards (Appendix C) prepared by Washington Educational Research for classroom assessments are examples to these codes.

Step 9: Each university and high school should constitute an examination board including trained examiners and testing professionals.

Step 10: The code should be enforced by the administrators of the universities. Furthermore, it should be enforced in the national level and the principles for enforcement of the code should also be included as a written statement in the code. However, the official bodies should cooperate with the school and university administrators in implementing the code.

Step 11: The code might be revised in every five years according to the recent developments in language testing. The joint committee formed of representatives from universities and policy makers might improve the code.

6.6 SUGGESTIONS FOR FURTHER RESEARCH

Further research may study the content of a 'code of ethics' for language test use in Turkey. Further study may also be carried out with the participation of the instructors of ELT departments across the country. Moreover, since stakeholder involvement is very important in language testing process, the study can be carried out with the participation of students as well as instructors.

6.7 CHAPTER SUMMARY

The aim of the study and the summary of the methodology were presented in the beginning of this chapter. Then, the discussions and conclusions were shown in the light of the data obtained through the analyses of the findings. Finally, implications and suggestions for further research were stated.

REFERENCES

A FRAMEWORK FOR THINKING ETHICALLY

1988 Issues in Ethics, I, 2
(Retrieved Dec. in 2006 from
http://www.scu.edu/ethics/practicing/decision/framework.html)

AERA, APA, NCME.

1999 Standards for Educational and Psychological Testing.
Washington, DC: American Educational Research Association.

ALDERSON, J. and C. CLAPHAM

1992 “Applied Linguistics and Language Testing: A Case Study of the ELTS Test.” Applied Linguistics, 13: 149-167.

ALDERSON, J. and D. WALL

1993 “Does Washback Exist?”
Applied Linguistics, 14: 115-129.

ALDERSON, J., C. CLAPHAM and D. WALL.

1995 Language Test Construction and Evaluation.
Cambridge: Cambridge University Press

ALDERSON, J. C.

1997 “Ethics and Language Testing.”
Paper presented at The Annual TESOL Convention. Orlando, Florida.

AMERICAN PSYCHOLOGICAL ASSOCIATION

1985 Standards for Educational and Psychological Testing.
Washington, DC: American Psychological Association.

BACHMAN, Lyle

1990 Fundamental Considerations in Language Testing.
1st ed. Oxford, NY: Oxford University Press.

BACHMAN, L. and A. PALMER.

1996 Language Testing in Practice.
1st ed. Oxford, NY: Oxford University Press.

BAILEY, Kathleen

1996 "Working for Washback: A Review of the Washback Concept in Language Testing." Language Testing, 13: 257-279.

BELL, Gregory

1994 "The Test of Testing: Making Appropriate and Ethical Choices in Assessment." ERIC Digest. (Retrieved Sep. 10, 2005, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/25/18/4f.pdf.)

BELL, Gregory

1994 "Making Appropriate & Ethical Choices in Large-scale Assessments: A Model Policy Code." ERIC Digest. (Retrieved in Oct. 15, 2005, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/0a/46.pdf)

BELL, Judith

1993 Doing Your Research Project
BUCKINGHAM: Open University Press.

CODE OF FAIR TESTING PRACTICES IN EDUCATION

1988 Washington D.C.: Joint Committee on Testing Practices.
(Retrieved in Oct. 19, 2006, from http://www.apa.org/science/fairtestcode.html)

CODE OF ETHICS FOR POLISH EXAMINERS

2004 "Questionnaire: A Code of Ethics and Professional Conduct for an Examiner"

(Retrieved Dec. in 2006 from http://elt.britcoun.org.pl/elt/forum/engtrans1.htm)

CODE OF ETHICS FOR POLISH EXAMINERS

2004 "Summary of Questionnaire Responses"

(Retrieved Dec. in 2006 from http://elt.britcoun.org.pl/elt/forum/qresults.htm)

CRONBACH, L. J.

1971 "Validity" in THORNDIKE. : 443-597.

CRONBACH, L. J.

1988 "Construct Validation After Thirty Years" in ROBERT L. LINN (Ed.) Intelligence: Measurement, Theory and Public Policy.

Urbana, III.: University of Illionis Press: 147-171,

DAVIES, Alan

1990 Principles in Language Testing

Oxford, England: Blackwell Publishing

DAVIES, Alan

1997 "Introduction: The Limits of Ethics in Language Testing."

Language Testing, 14: 235-241.

DAVIES, Alan

1997 "Demands of Being Professional in Language Testing."

Language Testing, 14: 328-339.

DOWD, S. and J. BATTLES

2007 " Curriculum Development and Alignment"
Instructional Design, Vol: 67: 3. (Retrieved in July 2007 from
https://www.asrt.org/media/Pdf/ForEducators/2_InstructionalDesign/2.2CurriculumDev.pdf

EBEL, R. and D. FRISBIE.

1991 Essentials of Educational Measurement.
5th edition. Englewood Cliffs, NJ: Prentice-Hall

FAIR TEST N.D.

"Implementing performance assessment : a guide to classroom, school and system reform."Fairtest: The National Center for Fair and Open Testing
(web site: fairtest@aol.com).

FREDERICKSON, Norm

1984 "The Real Test Bias: Influences of Testing on Teaching and Learning."
American Psychologist, 39: 193-202

FREDRICKSEN, J. and A. COLLINS.

1989 "A Systems Approach to Educational Testing."
Educational Researcher, 18: 27-32.

FULCHER, Glenn

1999 "Ethics in Language Testing"
TAE SIG Newsletter. Special Conference Issue, I, 1.

GENESEE, F. and J. UPSHUR.

1996 Classroom-based Evaluation in Second Language Education.
1st ed. Cambridge: Cambridge University Press.

HAMP-LYONS, Liz

1989 “Language Testing and Ethics.”

Prospect, 5: 7-15

HAMP-LYONS, Liz

1997 “Washback, Impact and Validity: Ethical Concerns.”

Language Testing, 14: 295-303.

HAMP-LYONS, Liz

1997a “Developing a Consequential Validation for Portfolio Assessment-based Writing Courses: College Level.”

Paper presented at The National Council of Teachers of English Seminar, Conflict and Consensus: Exploring Diversity and Standards in the Portfolio Movement. New Orleans, LA, January.

HAMP-LYONS, Liz

2000 “Fairness in Language Testing” in A. J. KUNNAN (Ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida : 16-19

Cambridge, UK: Cambridge University Press.

HEATON, J. B.

1988 Writing English Language Tests.

2nd edition. London: Longman.

HENNING, Grant

1987 A Guide to Language Testing.

Cambridge, Mass: Newbury House.

HILL, C. and K. PARRY.

1994b "Assessing English Language and Literacy Around the World." in C. HILL and K. PARRY (eds). From Testing to Assessment: English as an International Language, London: Longman, 253-271.

HOUSE, Ernest

1990 "Ethics of Evaluation Studies." in H. WAHLBERG and G. HAERTEL (eds.) The International Encyclopedia of Educational Evaluation.
Oxford: Pergamon Press, 91-94

HUGHES, Arthur

2002 Testing for Language Teachers.
2nd ed. Cambridge: Cambridge University Press.

HULIN, C. L. and F. DRASGOW and C. K. PARSONS.

1983 Item Response Theory : Application to Psychological Measurement.
Homewood, Ill. : Dow Jones-Irwin.

INGRAM, Elizabeth

1977 "Basic Concepts in Testing." in J.P.B. ALLEN and A. DAVIES (eds.)
Testing and Experimental Methods. Oxford: Oxford University Press.

KUNNAN, Antony

1999 "Recent Developments in Language Testing."
Annual Review of Applied Linguistics, 19: 235-253.
Retrieved in November 2006 from
(http://journals.cambridge.org/download.php?file=%2FAPL%2FAPL19%2FS0267190599190123a.pdf&code=d2bfdc4bbed0d5cea22b52fd9ff5084a)

KUNNAN, Antony (Ed.)

2000 Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium
Orlando, Florida. Cambridge, UK: Cambridge University Press.

LINN, R. L.

1983"Testing and Instruction: Links and Distinctions."
Journal of Educational Measurement, 20: 179-189.

LYNCH, Brian

1997 "The Ethical Potential of Alternative Assessment."
Paper presented at The 31st Annual Convention of Tesol, Orlando, FL.

LYNCH, Brian

1997 "In Search of the Ethical Test"
Language Testing, 14: 315-327.

MADAUS, G.F. and T. KELLEGHAN

1991 "Student Examination Systems in the European Community: Lessons for the United States." Contractor Report Submitted to the Office of Technology Assessment, United States Congress.

MARGHEIM, D.E.

2001 "Teacher Beliefs About the Outcomes of High-Stakes Testing and Measurement-Driven Instruction in Virginia's Public Schools."
Dissertation Submitted to the Faculty of the Virginia Polytechnic Institute and State University in Partial Fulfillment of the Requirements fot the Degree of Doctor of Education in Educational Leadership and Policy Studies.
(Retrieved in July 2007 from http://scholar.lib.vt.edu/theses/available/etd-12052001-211701/unrestricted/MargheimChap1.pdf)

MCMILLAN, H.J. and S. SCHUMACHER

1993 Research in Education: A Conceptual Introduction.

New York: Harper Collins College Publishers.

MCNAMARA, Tim

2000 Language Testing.

Oxford: Oxford University Press.

MCNAMARA, Tim

2005 “21st Century Shibboleth: Language Tests, Identity and Intergroup Conflict.” Language Policy, Ⅳ, 4, November. (Retrieved in Dec. 2006 from http://www.springerlink.com/content/f86u4l74u2m8h426/fulltext.pdf)

MESSICK, Samuel

1975 “The Standard Problem: Meaning and Values in Measurement and Evaluation.” American Psychologist, 30: 955-966.

MESSICK, Samuel

1980 “Test Validity and the Ethics of Assessment.”

American Psychologist, 35: 1012-1027

MESSICK, Samuel

1989 “Validity” in LINN. : 13-103

NEW ENGLISH DICTIONARY

1815 Edition

O’HEAR, Anthony

1985 What Philosophy Is.

Harmondsworth: Penguin Books.

OPPENHEIM, A.N

1992 Questionnaire Design, Interviewing and Attitude Measurement.
London: Printer.

PENNYCOOK, Alastair

1994 The Cultural Politics of English as a World Language.
London: Longman.

POWER, M.A.(ed)

1999 “Ethical Standards In Testing: Test Preperation and Administration”
WERA Professional Publications Vol: 1 (Revised 2001)
(Retrieved Dec. in 2006 from
http://www.weraweb.org/pages/publications/WERA_Test_Ethics.pdf)

PUNCH, Maurice

1994 “Politics and Ethics in Qualitative Research.” in N.K. DENZIN and Y.S. LINCOLN (eds). Handbook of Qualitative Research. Thousand Oaks, CA: Sage, 83-97

RAWLS, John

1967 “Distributive Justice.” in P. LASLETT and W. RUNCIMAN (eds). Philosophy, Politics and Society (3rd series). Oxford: Blackwell

STEWART, R.M. (Ed.)
“Distributive Justice.” Readings in Social and Political Philosophy
New York: Oxford University Press, 219-234.

REA-DICKINS, Pauline

1997 “So, Why Do We Need Relationships with Stakeholders in Language Testing? A View from the UK.”
Language Testing, 14: 304-314.

REWARD AND DISCIPLINE REGULATION OF THE SECONDARY EDUCATION INSTITUTIONS

2007 (Retrieved June, 07, 2007, from
(http://mevzuat.meb.gov.tr/html/22188_0.html).

SCHMEISER, C. B.

1995 "Ethics in Assessment."
ERIC Digest. Retrieved September 12, 2005,
from http://www.ericdigests.org/1996-3/in.htm.)

SCRIVEN, Michael.

1991 Evaluation Thesaurus (4th edn).
Newbury Park, CA: Sage.

SHEPARD, Lorrie

1982 "Definitions of Bias." in R.A. BERK (Ed.), Handbook of Methods for Detecting Test Bias, Baltimore, MD: Johns Hopkins University Press, 9-30.

SHEPARD, Lorrie

1987 "The Case for Bias in Tests of Achievement and Scholastic Aptitude." in S. MOGDIL and C. MOGDIL (Eds), Arthur Jensen: Consensus and Contreversy. New York: Falmer Press, 177-190.

SHOHAMY, Elena

1997 "Testing Methods, Testing Consequences: Are They Ethical? Are they fair?" Language Testing, 14: 340-349.

SHOHAMY, Elena

2000 "Fairness in Testing." in A. J. KUNNAN (Ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida : 15-19.
Cambridge, UK: Cambridge University Press.

SPAAN, Mary

2000 "Enhancing Fairness Through a Social Contract." in A. J. KUNNAN (Ed.), Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida : 35-38. Cambridge, UK: Cambridge University Press.

SPIELMAN, S. J.

1993 "Adapting Benchmarks to French Immersion Education."
Orbit, 24: 25-28.

SPOLSKY, Bernard

1997 "The Ethics of Gatekeeping Tests: What Have We Learned in a Hundred Years?" Language Testing, 14: 242-248.

STANLEY, Julian

1971 "Reliability" in THORNDIKE, 1981: 356-442

STEWART, R. M. (ed)

1996 Readings in Social and Political Philosophy.
New York: Oxford University Press.

STUDENT DISCIPLINE REGULATIONS OF EASTERN MEDITERRENEAN UNIVERSITY

2007 (Retrieved February 11, 2007, from http://registrar.emu.edu.tr/tuzuk/7-%20Ogrenci%20Disiplin%20Yonetmeligi.pdf)

VANSICKLE, Timothy

2003 "Types and Uses of Tests."
ERIC Digest. (Retrieved Oct. 25, 2006, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/57/74.pdf)

WAGNER, Elvis

2006 "Can the Search for 'Fairness' Be Taken Too Far?" Teachers college, Columbia University Working Papers in TESOL & Applied Linguistics, VI, 2. (Retrieved in Dec. in 2006 from http://journals.tc-library.org/templates/about/editable/pdf/Wagner%20Forum.pdf)

WALL, D. and J. ALDERSON

1993 "Examining Washback: The Sri Lankan Impact Study." Language Testing, 10: 41-70.

WEBSTER'S NINTH NEW COLLEGIATE DICTIONARY OF THE ENGLISH LANGUAGE

1994 Edition

WEISS, Carol

1986 "The Stakeholder Approach to Evaluation." in E.R. HOUSE (Ed). New Directions in Educational Evaluation. Lewes: Falmer Press, 145-157.

YILDIZ UNIVERSITY ACADEMIC ETHICS BOARD

2007 (Retrieved July 11, 2007, from http://www.aek.yildiz.edu.tr/index.htm)

APPENDICES

APPENDIX A
CODE OF FAIR TESTING PRACTICES IN EDUCATION

A. Developing and Selecting Appropriate Tests

TEST DEVELOPERS	TEST USERS
Test developers should provide the information and supporting evidence that test users need to select appropriate tests.	Test users should select tests that meet the intended purpose and that are appropriate for the intended test takers.
A-1. Provide evidence of what the test measures, the recommended uses, the intended test takers, and the strengths and limitations of the test, including the level of precision of the test scores.	A-1. Define the purpose for testing, the content and skills to be tested, and the intended test takers. Select and use the most appropriate test based on a thorough review of available information.
A-2. Describe how the content and skills to be tested were selected and how the tests were developed.	A-2. Review and select tests based on the appropriateness of test content, skills tested, and content coverage for the intended purpose of testing.
A-3. Communicate information about a test's characteristics at a level of detail appropriate to the intended test users.	A-3. Review materials provided by test developers and select tests for which clear, accurate, and complete information is provided.
A-4. Provide guidance on the levels of skills, knowledge, and training necessary for appropriate review, selection, and administration of tests.	A-4. Select tests through a process that includes persons with appropriate knowledge, skills, and training.
A-5. Provide evidence that the technical quality, including reliability and validity, of the test meets its intended purposes.	A-5. Evaluate evidence of the technical quality of the test provided by the test developer and any independent reviewers.
A-6. Provide to qualified test users representative samples of test questions or practice tests,	A-6. Evaluate representative samples of test questions or practice tests, directions, answer

directions, answer sheets, manuals, and score reports.	sheets, manuals, and score reports before selecting a test.
A-7. Avoid potentially offensive content or language when developing test questions and related materials.	A-7. Evaluate procedures and materials used by test developers, as well as the resulting test, to ensure that potentially offensive content or language is avoided.
A-8. Make appropriately modified forms of tests or administration procedures available for test takers with disabilities who need special accommodations.	A-8. Select tests with appropriately modified forms or administration procedures for test takers with disabilities who need special accommodations.
A-9. Obtain and provide evidence on the performance of test takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed.	A-9. Evaluate the available evidence on the performance of test takers of diverse subgroups. Determine to the extent feasible which performance differences may have been caused by factors unrelated to the skills being assessed.

B. Administering and Scoring Tests

TEST DEVELOPERS	TEST USERS
Test developers should explain how to administer and score tests correctly and fairly.	Test users should administer and score tests correctly and fairly.
B-1. Provide clear descriptions of detailed procedures for administering tests in a standardized manner.	B-1. Follow established procedures for administering tests in a standardized manner.
B-2. Provide guidelines on reasonable procedures for assessing persons with disabilities who need special accommodations or those with diverse linguistic backgrounds.	B-2. Provide and document appropriate procedures for test takers with disabilities who need special accommodations or those with diverse linguistic backgrounds. Some accommodations may be required by law or regulation.

B-3. Provide information to test takers or test users on test question formats and procedures for answering test questions, including information on the use of any needed materials and equipment.	B-3. Provide test takers with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing.
B-4. Establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting.	B-4. Protect the security of test materials, including respecting copyrights and eliminating opportunities for test takers to obtain scores by fraudulent means.
B-5. Provide procedures, materials and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, provide adequate training for scorers.	B-5. If test scoring is the responsibility of the test user, provide adequate training to scorers and ensure and monitor the accuracy of the scoring process.
B-6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.	B-6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.
B-7. Develop and implement procedures for ensuring the confidentiality of scores.	B-7. Develop and implement procedures for ensuring the confidentiality of scores.

C. Reporting and Interpreting Test Results

TEST DEVELOPERS	TEST USERS
Test developers should report test results accurately and provide information to help test users interpret test results correctly.	Test users should report and interpret test results accurately and clearly.
C-1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.	C-1. Interpret the meaning of the test results, taking into account the nature of the content, norms or comparison groups, other technical evidence, and benefits and limitations of test results.

C-2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.	C-2. Interpret test results from modified test or test administration procedures in view of the impact those modifications may have had on test results.
C-3. Specify appropriate uses of test results and warn test users of potential misuses.	C-3. Avoid using tests for purposes other than those recommended by the test developer unless there is evidence to support the intended use or interpretation.
C-4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.	C-4. Review the procedures for setting performance standards or passing scores. Avoid using stigmatizing labels.
C-5. Encourage test users to base decisions about test takers on multiple sources of appropriate information, not on a single test score.	C-5. Avoid using a single test score as the sole determinant of decisions about test takers. Interpret test scores in conjunction with other information about individuals.
C-6. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.	C-6. State the intended interpretation and use of test results for groups of test takers. Avoid grouping test results for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use. Report procedures that were followed in determining who were and who were not included in the groups being compared and describe factors that might influence the interpretation of results.
C-7. Provide test results in a timely fashion and in a manner that is understood by the test taker.	C-7. Communicate test results in a timely fashion and in a manner that is understood by the test taker.
C-8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.	C-8. Develop and implement procedures for monitoring test use, including consistency with the intended purposes of the test.

D. Informing Test Takers

Test developers or test users should inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores.

D-1. Inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers.

D-2. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether a test should be taken—including indications of any consequences that may result from not taking the test (e.g., not being eligible to compete for a particular scholarship) —and whether there is an available alternative to the test.

D-3. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid.

D-4. Provide test takers or their parents/guardians with information about responsibilities test takers have, such as being aware of the intended purpose and uses of the test, performing at capacity, following directions, and not disclosing test items or interfering with other test takers.

D-5. Inform test takers or their parents/guardians how long scores will be kept on file and indicate to whom, under what circumstances, and in what manner test scores and related information will or will not be released. Protect test scores from unauthorized release and access.

D-6. Describe procedures for investigating and resolving circumstances that might result in canceling or withholding scores, such as failure to adhere to specified testing procedures.

D-7. Describe procedures that test takers, parents/guardians, and other interested parties may use to obtain more information about the test, register complaints, and have problems resolved.

Code of Fair Testing Practices in Education (2004).
Washington, DC: Joint Committee on Testing Practices.
http://www.apa.org/science/fairtestcode.html

APPENDIX B
CHECKLIST FOR EVALUATING TEST USEFULNESS

Questions for logical evaluation of usefulness	Extent to which quality is satisfied	Explanation of how quality is satisfied
Reliability		
1. To what extent do characteristics of the test setting vary from one administration of the test to another?		
2. To what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
3. To what extent do characteristics of the test input vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
4. To what extent do characteristics of the expected response vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
5. To what extent do characteristics of the relationship between input and response vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
Construct Validity		
Clarity and appropriateness of the construct definition, and the appropriateness of the task characteristics with respect to the construct definition.		
6. Is the language ability construct for this test clearly and unambiguously defined?		
7. Is the language ability construct for this test relevant to the purpose of the test?		
8. To what extent does the test task reflect the construct definition?		
9. To what extent do the scoring procedures reflect the construct definition?		

10. Will the scores obtained from the test help us to make the desired interpretations about test takers' language ability?		
Possible sources of bias in the test characteristics		
11. What characteristics of the test setting are likely to cause different test takers to perform differently?		
12. What characteristics of the test rubric are likely to cause different test takers to perform differently?		
13. What characteristics of the test input are likely to cause different test takers to perform differently?		
14. What characteristics of the expected response are likely to cause different test takers to perform differently?		
15. What characteristics of the relationship between input and response are likely to cause different test takers to perform differently?		
Authenticity		
16. To what extent does the description of tasks in the TLU domain include information about the setting, input, expected response, and relationship between input and response?		
17. To what extent do the characteristics of the test task correspond to those of TLU tasks?		
Involvement of the test takers' topical knowledge		
18. To what extent does the task presuppose the appropriate area or level of topical knowledge, and to what extent can we expect the test takers to have this area or level of topical knowledge?		
Suitability of test tasks to the personal characteristics of the test takers		
19. To what extent are the personal characteristics of the test takers included in the design statement?		
20. To what extent are the characteristics of the test tasks suitable for test takers with the specified personal characteristics?		
Involvement of the test takers' language knowledge		
21. Does the processing required in the test task involve		

a very narrow range or a wide range of areas of language knowledge?		
Involvement of language functions in the test tasks		
22. What language functions, other than the simple demonstration of language ability, are involved in processing the input and formulating a response?		
Involvement of the test takers' metacognitive strategies		
23. To what extent are the test tasks independent?		
24. How much opportunity for strategy involvement is provided?		
Involvement of the test takers' affective schemeta in responding to the test tasks		
25. Is this test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best?		
Impact		
Impact on individuals		
Impact on test takers		
26. To what extent might the experience of taking the test or the feedback received affect characteristics of the test takers that pertain to language use (such as topical knowledge, perception of the target language use situation, areas of language knowledge, and use of strategies) ?		
27. What provisions are there for involving test takers directly, or for collecting and utilizing feedback from test takers, in the design and development of the test?		
28. How relevant, complete, and meaningful is the feedback that is provided to test takers?		
29. Are decision procedures and criteria applied uniformly to all groups of test takers?		
30. How relevant and appropriate are the test scores to the decisions to be made?		
31. Are the test takers fully informed about the procedures and criteria that will be used in making decisions?		

32. Are these procedures and criteria actually followed in making the decisions?		
Impact on teachers		
33. How consistent are the areas of language ability to be measured with those that are included in teaching materials?		
34. How consistent are the characteristics of the test and test tasks with the characteristics of teaching and learning activities?		
35. How consistent is the purpose of the test with the values and goals of teachers and of the instructional program?		
Impact on society and education systems		
36. Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?		
37. To what extent do the values and goals of the test developer coincide or conflict with those of society and the education system?		
38. What are the potential consequences, both positive and negative, for society and the education system, of using the test in this particular way?		
39. What is the most desirable positive consequence, or the best thing that could happen, as a result of using the test in this particular way, and how likely is this to happen?		
40. What is the least desirable negative consequence, or the worst thing that could happen, as a result of using the test in this particular way, and how likely is this to happen?		
Practicality		
41. What type and relative amounts of resources are required for: (a) the design stage, (b) the operationalism stage, and (c) the adinistration stage?		
42. What sources will be available for carrying out (a), (b), and (c) above?		

A checklist for evaluating usefulness (Bachman et al. 1996: 150-155)

APPENDIX C

ETHICAL STANDARDS IN TESTING: TEST PREPERATION AND ADMINISTRATION

The following individuals participated in a series of seminars in 1998-99 during which these standards were developed.

Jim Nelson
Seminar Facilitator and Writer
WERA Member Emeritus
Gig Harbor, WA

Linda Elman
Director of Research & Evaluation
Central Kitsap School District

Gordon Ensign Jr.
Director of Assessment (Retired)
Commission on Student Learning

Jill Hearne
Educational Consultant
Seattle

Bev Henderson
Curriculum Coordinator
Kennewick School District

Audrian Huff
Principal, Fairwood Elem. School
Kent School District

Wally Hunt
Supervisor, Title I/Learning
Assistance Program
Office Supt. of Public Instruction

Jerry Litzenberger
Director, Graduate Follow up Study
Snohomish, WA

Duncan MacQuarrie
Director of Curriculum and Assessment
Office Supt. Of Public Instruction

Steve Siera
Director, Research & Assessment
Kent School District

Bob Silverman
Senior WASL Analyst
Office Supt. of Public Instruction

Donna Smith
Principal, Terminal Park Elem. School
Auburn School District

Ric Williams
Director, Evaluation and Research
Everett Public Schools

WASHINGTON EDUCATIONAL RESEARCH ASSOCIATION ETHICAL STANDARDS TEST PREPARATION AND ADMINISTRATION

IT IS APPROPRIATE AND ETHICAL TO:

1. Communicate to students, parents and the public what any test does and does not do, when and how it will be administered, and how the results may be appropriately used.

2. Teach to the Essential Learning Requirements (WA. state curriculum standards) at each grade level so that students will learn the skills and knowledge they need to accurately show what they know and can do.
3. Incorporate all subject area objectives into the local curriculum throughout the year including, but not limited to, the objectives of the tests to be administered.
4. Review skills, strategies, and concepts previously taught.
5. Teach and review test-taking and familiarization skills that include an understanding of test characteristics independent of the subject matter being tested.
6. Use any test preparation documents and materials prepared by the test-maker, the Office of the Superintendent of Public Instruction or the Commission on Student Learning.
7. Read and discuss the test administration manual with colleagues.
8. Schedule and provide the appropriate amount of time needed for the assessment.
9. Take appropriate security precautions before, during and after administration of the test.
10. Include all eligible students in the assessment.
11. Actively proctor students during tests, keeping them focused and on task.
12. Seek clarification on issues and questions from the administrative team responsible for ethical and appropriate practices.
13. Avoid any actions that would permit or encourage individuals or groups of students to receive scores that misrepresent their actual level of knowledge and skill.

BEFORE THE TEST - IT IS INAPPROPRIATE AND UNETHICAL TO:

1. Use any test preparation material that promises to raise scores on a particular test by targeting skills or knowledge from specific test items, and does not increase students' general knowledge and skills. Materials which target the general skills tested may be appropriate if they reflect school or district priorities and best practices.
2. Limit curriculum and instruction only to those skills, strategies, and concepts included on the test.
3. Limit review to only those areas on which student performance was low on previous tests.
4. "Cram" test material just before the tests are given.
5. Train students for testing using locally developed versions of national norm-referenced tests.
*6. Reveal all or any part of secure copyrighted tests to students, in any manner, oral or written, prior to test administration.
*7. Copy or otherwise reproduce all or any part of secure or copyrighted tests.
*8. Review or provide test question answers to students.
*9. Possess unauthorized copies of state tests.

DURING THE TEST - IT IS INAPPROPRIATE AND UNETHICAL TO:

1. Read any parts of the test to students except where indicated in the directions.
2. Define or pronounce words used in the test.
3. Make comments of any kind during the test, including remarks about quality or quantity of student work, unless specifically called for in the administration manual.
4. Give "special help" of any kind to students taking the test.
5. Suggest or "coach" students to mark or change their answers in any way.
6. Exclude eligible students from taking the test.
*7. Reproduce test documents for any purpose.

** It is illegal under state statute to conduct or assist in carrying out any of the items marked with *. (Penalties may range from fines to dismissal, or even withdrawal of certification. [RCW 28A.230.190. Acts of Unprofessional Conduct, WAC 180-87-050])*

AFTER THE TEST - IT IS INAPPROPRIATE AND UNETHICAL TO:

1. Make inaccurate reports, unsubstantiated claims, inappropriate interpretations, or otherwise false and misleading statements about assessment results.
*2. Erase or change student answers.

** It is illegal under state statute to conduct or assist in carrying out any of the items marked with *. (Penalties may range from fines to dismissal, or even withdrawal of certification. [RCW 28A.230.190. Acts of Unprofessional Conduct, WAC 180-87-050])*

Many of the issues regarding ethical assessment practice are in the hands of the classroom teacher, but a significant number of these issues must be addressed through administrative practice.

GUIDELINES FOR TEST PREPARATION AND ADMINISTRATION

The Teacher's Role:

Students will do their best on tests if they find an encouraging and supportive atmosphere, if they know that they are well prepared, and that with hard work they will perform well. To create a situation that will encourage students to do their best, teachers should:

1. Attend workshops on test administration.
2. Develop an assessment calendar and schedule and share it with students and parents.
3. Prepare students well in advance for assessment by teaching test-wiseness skills independent of the subject matter being tested. Teach and review test familiarity that includes an understanding of how to use the test booklets and answer sheets, item response strategies, time management, listening, and following directions.
4. Develop a list of which and how many students will be tested and when. Determine which students will require special accommodations.
5. Develop a list of students who will be exempted from testing and the reason for the exemption. This list must be reviewed and approved by the principal or test administration committee. Parents must be notified and alternative assessments must be identified.
6. Develop plans for the administration of makeup tests for students absent during the scheduled testing period.
7. Prepare and motivate students just before the test.
8. Prepare to administer the test, with sufficient materials available for all students to be tested.
9. Prepare classrooms for the test. Arrange for comfortable seating where students will not be able to see each other's test materials but will be able to hear test directions. Eliminate posters or other materials that may be distracting or contain information that could be used to help students answer test items.
10. Alert neighboring teachers to the testing schedule and ask their help in achieving optimal testing conditions and in keeping noise levels to a minimum.
11. Arrange for a separate supervised area for those students who finish early and may cause a distraction for other students.
12. Read the test administration manual carefully, in advance. Administer the test according to directions.
13. Meet with proctors and discuss their duties and responsibilities. Carefully and actively proctor the test.
14. Arrange for appropriate breaks and student stress relievers.
15. Follow the rules for test security and return all test materials to the test administrator.

The Principal's Role

There are a number of things the principal can do to enhance the testing atmosphere in the school.
1. Inform both students and parents about what each test does and does not do, when and how it will be administered, and how the results will be reported and used. Indicate the importance of tests for students, staff, and the school. Stress the importance of school attendance on the scheduled testing dates.
2. Encourage the implementation of appropriate test-wiseness teaching and review. Teaching test familiarity skills should be independent of subject matter being tested. Discourage subject matter drill and practice solely for the test.

3. Let parents know about upcoming tests and what they can do to encourage their children's performance.
4. Work with teachers to develop a building testing schedule. Attempt to maximize the efficiency of the building's physical layout and resources.
5. Pay careful attention to school schedules during the testing period. Avoid planning assemblies, fire drills, maintenance, etc., during the testing period.
6. Develop a plan to keep tests and answer sheets secure before and after administration, and ensure that all are returned properly.
7. Arrange, where possible, for teachers to have proctoring help in administering tests. Ensure that tests are carried out according to ethical and legal practice.
8. Provide a handbook or policy statement such as this one to all involved with test administration spelling out proper and improper testing procedures.
9. Create a process to check out any suspicions or allegations of cheating. Document all steps taken.
10. Require detailed written explanations about why a student was not tested or the reason a score was not figured into a school's average.
11. Encourage teachers' participation in workshops and inservice sessions on assessment.
12. Ensure that all students are tested. Review all test accommodations, including exclusion, as a last resort, made for students with special needs. Ensure that accommodations/exclusions are consistent with specific testing program guidelines, and that appropriate accommodations are available as needed.
13. Ensure that there are no interruptions in classrooms during the testing period, including custodial tasks, intercom calls, delivery of messages, etc.
14. Work with the test coordinator and classroom teachers to schedule and staff makeup days for students who miss all or parts of the test. This might include bringing in a substitute or finding other ways to creatively use building staff to administer makeups in an appropriate setting.
15. Share test results with all staff. Staff members need to work together to ensure that the testing process is a smooth one. School improvement is a team effort.

http://www.wera-web.org/pages/publications/WERA_Test_Ethics.pdf

APPENDIX D
QUESTIONNAIRE ON ETHICS IN LANGUAGE TESTING

Dear Colleague,

This questionnaire is designed to learn the perceptions of instructors in the Department of Foreign Languages at Çanakkale Onsekiz Mart University about ethics concept and attitudes toward ethical issues in language testing.

There are no wrong answers in this questionnaire. The data collected will be used to carry out more ethical and appropriate language tests and to develop a code of ethics for language testing in education.

I would appreciate it if you would take the time to answer the following questions. Thank you for your cooperation.

Instructor Anıl Ceylan
Çanakkale Onsekiz Mart University

PERSONAL INFORMATION

Graduated from...Department

TITLE:

a) Instructor b)Language Specialist c)Asst. Prof. Dr. d)Assoc. Prof. Dr. e)Prof.Dr.

LENGTH OF SERVICE (Years):

a) 0-5 b) 6-10 c)10-15 d) 16-20 e)21-25 f)Other..........

Instructions: Please, mark the response indicating your level of agreement with each of the following statements based on your experiences, opinions or perceptions as shown in the table below.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
×				

QUESTIONS

PART 1

	Strongly Agree				Strongly Disagree
1. I believe that reliability is important in language testing ethics.	5	4	3	2	1
2. I believe that validity is important in language testing ethics.	5	4	3	2	1
3. I believe that test bias is important in language testing ethics.	5	4	3	2	1
4. I believe that washback is important in language testing ethics.	5	4	3	2	1
5. I believe that test impact is important in language testing ethics.	5	4	3	2	1
6. I believe that the stakeholder involvement in test preperation can contribute to the language testing process.	5	4	3	2	1
7. I believe that the stakeholder involvement in test practice can contribute to the language testing process.	5	4	3	2	1
8. I believe that the stakeholder involvement in test evaluation can contribute to the language testing process.	5	4	3	2	1
9. I believe that the stakeholder involvement in making decisions can contribute to the language testing process.	5	4	3	2	1
10. I believe that a tester should be accountable for all possible consequences of language test use.	5	4	3	2	1
11. I believe that test takers are responsible for ethical test use in language programs.	5	4	3	2	1
12. I believe that families are responsible for ethical test use in language programs	5	4	3	2	1
13. I believe that policy makers are responsible for ethical test use in language programs	5	4	3	2	1

14. I believe that test users are responsible for ethical test use in language programs.	5	4	3	2	1
15. I believe that test developers are responsible for ethical test use in language programs.	5	4	3	2	1
16. I believe that confidentiality is important with regard to the rights of test takers in language testing ethics	5	4	3	2	1
17. I believe that the right to access to information is important with regard to the rights of test takers in language testing ethics.	5	4	3	2	1
18. Language tests are used for particular purposes such as assessment, diognosis etc. I believe that it is appropriate to use them for non-intended purposes. e.g. introducing a new education policy.	5	4	3	2	1

PART 2

1. There is a 'code of ethics for language test use' in Turkey.	YES		I DO NOT KNOW		NO
2. A 'code of ethics for language test use' is necessary to ensure fairness in language testing.	5	4	3	2	1
3. A 'code of ethics for language test use' is necessary for fairness in language testing, because:	Strongly Agree				Strongly Disagree
• appropriate behaviours should be stated in a written document.	5	4	3	2	1
• unclear situations should be explained	5	4	3	2	1
• the personal qualities of an examiner should be shown	5	4	3	2	1
• if there is another/are others, please specify.	....				
4. A 'code of ethics for language test use' is not necessary for fairness in language testing, because:	Strongly Agree				Strongly Disagree
• teachers are honest people and aware of their responsibilities.	5	4	3	2	1

• there is already a code of professional ethics.	5	4	3	2	1
• if there is another/are others, please specify.	....				
5. What should 'a code of ethics for language test use' include?	Strongly Agree				Strongly Disagree
• rules regarding what is and what is not allowed	5	4	3	2	1
• examiners' responsibilities.	5	4	3	2	1
• rules regarding how to punish dishonest examiners	5	4	3	2	1
• appropriate rules of behaviour during the exam.	5	4	3	2	1
• discipline	5	4	3	2	1
• if there is another/are others, please specify.	....				

PART 3

1. What do you consider to be the unethical behaviors of a test giver?	Strongly Agree				Strongly Disagree
• asking someone else to mark the testee's paper.	5	4	3	2	1
• lack of confidentially	5	4	3	2	1
• marking the papers subjectively	5	4	3	2	1
• helping one's own students	5	4	3	2	1
• if there is another/are others, please specify.	....				
2. What do you consider the unethical behaviors of a test taker?	Strongly Agree				Strongly Disagree
• cheating during the exam	5	4	3	2	1
• helping other student(s) during the exam	5	4	3	2	1
• interrupting other students during the exam	5	4	3	2	1
• unpunctuality (not arriving at the exam venue on time).	5	4	3	2	1
• If there is another/are others, please specify.	....				
3. What should be the qualities of a language test	Strongly				Strongly

giver?	Agree				Disagree
• objective	5	4	3	2	1
• honest	5	4	3	2	1
• patient	5	4	3	2	1
• unbiased	5	4	3	2	1
• disciplined	5	4	3	2	1
• responsible	5	4	3	2	1
• empathetic	5	4	3	2	1
• kind	5	4	3	2	1
• punctual	5	4	3	2	1
• understanding	5	4	3	2	1
• hardworking	5	4	3	2	1
• cooperative	5	4	3	2	1

PART 4

	Strongly Agree				Strongly Disagree
1. Value systems of the society affect language test use in education.	5	4	3	2	1
2. Individual rights of test takers should be considered by test developers and users to ensure fairness in language testing.	5	4	3	2	1
3. Individual differences of test takers should be considered by test developers and users to ensure fairness in language testing.	5	4	3	2	1
4. There are different standards for different examiners.	5	4	3	2	1
5. The moral values vary from one culture to another.	5	4	3	2	1
6. The moral values vary from one situation to another.	5	4	3	2	1
7. The moral values vary from one individual to another.	5	4	3	2	1
8. Adopting the 'code of ethics for language test use' of another country is useful, since different cultures have common ethical values.	5	4	3	2	1

DEFINITIONS

The terms ‘ethics’, ‘morality’, and ‘fairness’ used in this questionnaire refer to ethical conduct in language test development, use, and interpretation.

The ethical conduct may include:

-The responsibilities of test developers, users, and takers.
-The unethical behaviors of test developers, test users, and test takers.
-The relationship between individual and public morality and language test use.
-Codes of ethics and codes of practice developed for ethical test use.

Test: Test (psychological or educational) is a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual (Carroll, 1968, p.46, cited in Bachman, 1990, p.20). In the language testing area, language tests are constructed for particular purposes such as placement, diognosis, proficiency, etc. (Bachman, 1990).

Reliability: Reliability is the consistency of the measures across different times, test forms, raters, and other characteristics of the measurement context (Bachman, 1990, p. 24).

Validity: A test is said to be valid to the extent that it measures what it is supposed to measure (Henning, 1987, p. 89, cited in Alderson et al. 1995, p.170)

Test Bias: Tests have specific uses and they are carried out with specific groups of test takers. However, there may be other groups within these groups and there may be differences between them other than language ability. These differences may affect the test performance of these groups as well as the validation process. This is called test bias (Bachman, 1990).

Washback: The impact of language tests on teaching and learning is called “washback” (McNamara, 2000).

Test Impact: Tests may have wider effects in the community including the school which is called “test impact” (McNamara, 2000).

Test user: Test users are people and agencies that select tests, administer tests, commission test development services, or make decisions on the basis of test scores (Code of Fair Testing Practices in Education, 1988)

Test developer: Test developers are people and organizations that construct tests, as well as those that set policies for testing programs (Code of Fair Testing Practices in Education, 1988)

Stakeholders: Stakeholders are the language testers, teachers, parents, administrators, teacher educators, sponsors and funding bodies, government bodies, the public, various international and national examination authorities, members of working parties and curriculum committees, test takers, test administrators such as university admission officers interpreting test scores (Rea-Dickens, 1997).

APPENDIX E

QUESTIONNAIRE: 'A CODE OF ETHICS AND PROFESSIONAL CONDUCT FOR AN EXAMINER'

The object of the following questionnaire is to gather opinions on a code of ethics for examiners. Please fill it in giving short answers.

- Is there any need (necessity) of creating such a code? Please explain the reasons for your answer ....
- What should a code of ethics and professional conduct of an examiner contain? ....
- What conduct of an examiner would you consider unethical? ....
- What qualities should an examiner possess with regard to …?

a. His/her personality ....
b. His/her professional life ....

- What are the duties of an examiner with regards to ...?

a. A student he/she assesses..
b. Regional Examination Board...
c. Other examiners..

Your remarks on examiners' behaviour during the Matura exam (May 2002) ....

Thank you for your co-operation!

PERSONAL INFORMATION

- EDUCATION..
- HOW LONG HAVE YOU BEEN A TEACHER

0 –5		6 –10		11 – 15		16 – 20		21 – 25	

- WHAT TYPE OF SCHOOL DO YOU TEACH IN

Vocational school	Technical school	Liceum	Other

- GENDER: F, M

- ARE YOU A TRAINED EXAMINER?

Yes		No	

- WHEN AND WHERE WERE YOU TRAINED?

Place		Time	

- IN THE SCHOOL YEAR 2001/2002 I PARTICIPATED IN:

Internal exams

Yes		No	

External exams

Yes		No	

Produced in Poland by British Council © 2004. The United Kingdom's international organisation for educational opportunities and cultural relations. We are registered in England as a charity.

http://elt.britcoun.org.pl/elt/forum/engtrans1.htm

APPENDIX F
SUMMARY OF QUESTIONNAIRE RESPONSES OF 'A CODE OF ETHICS AND PROFESSIONAL CONDUCT FOR AN EXAMINER'

Out of 100 questionnaires distributed at different conferences 62 were returned to the author

PERSONAL INFORMATION RESULTS

- EDUCATION: university diploma - 57 (93%)

- HOW LONG HAVE YOU BEEN A TEACHER?

0 -5	**1.6%**	6 -10	**21%**	11 -15	**21%**	16 – 20	**16. 4%**	21 – 25	**36%**

- WHAT TYPE OF SCHOOL DO YOU TEACH IN?

Vocational school	Technical school	*Liceum*	Other
10 %	**23%**	**76%**	**10%**

- GENDER: F - 52 (**84 %**) M - 10 (**16%**)

- ARE YOU A TRAINED EXAMINER?

Yes	**93%**

- IN THE SCHOOL YEAR 2001/2002 I PARTICIPATED IN:

- Internal exams

Yes	**32%**

- External exams

Yes	**66%**

Is there any need (necessity) of creating such a code? Please explain the reasons for your answer?

YES **52** NO **10**

YES:

It would be useful

to show the personal qualities of an examiner - 2

to demonstrate appropriate behaviour - 3

to explain unclear situations - 6

there should be some written document with regards to proper behaviour - 4

if one knows the duties it is easier to accept them and take decisions, to make the work of an examiner more efficient – 2

would warn against temptation, the same for all examiners - 2

will eliminate possible imprecise marking, justifications seems self-evident, to draw attention to both aspects (professional and ethical) of an examiner's work - 2

new examination & new tasks for examiners, to increase objective marking - 2

for somebody who hasn't got it - 2

some examiners' problems might be solved using this code - 4

to learn what the tasks are before you decide to become an examiner, because of the unethical behaviour of teachers and students who cheated, to make the duties of an examiner clear - 4

to make people realise the most obvious things, to set clear rules for examiner selection, to make parents and students realise that examiners are objective, competent and honest. If you don't speak about it and it has not been expressed somehow then it is not obligatory, examiners must realise their responsibilities and trust, you cannot do without rights and duties - 2

NO:

Teachers on the whole are people of great moral standard and are conscious of their duties - 4

because a set of rules will change nothing, you are either a pedagogue or it's substitute, in fact teachers who teach and assess already use it - 2

a 'code of professional ethics' exists - 2

What should a code of ethics and professional conduct of an examiner contain?

Directions of what an examiner should not do, what attitude he should present - 2

professionalism of an examiner - 2

his Kinderstube - 2

discipline - 2

how to co-operate with your boss, what to do in specific situations - 5

norms of what is and what is not allowed - 13

description of what to do in difficult, unexpected situations - 2

specifying examiners code of ethics - 3

examiners' duties - 8

what must be observed, rules of behaviour (general) - 6

elements which are in other codes, examiners' rights, detailed list of duties, general and specific rules of the work for OKE, qualities of an examiner - 2

rules of behaviour during the exam and ways of punishing dishonest examiners - 4

exam regulations (like for CU local syndicate - FCE, CAE, CPE)

What conduct of an examiner would you consider unethical?

Asking somebody else to mark the scripts an examiner was given (help), incorrect marking (too little time) - 8

inaccuracy - 3

carelessness in script marking - 3

opposite of conscientious in decoding students' scripts - 2

not treating the work confidentially - 27

subjective marking - 8

corruption - 6

tolerance of cheating - 3

being biased in assessment - 7

lowering marks of students in a different school, increasing marks of students in one's own school - 3

helping one's own students - 5

inflexibility, lack of firmness, behaviour which is unethical, unreliability, not keeping to the instructions, unprofessional behaviour - 3

no effort in trying to understand the students' way of thinking, influencing other examiners, unjust marking, assessment not according to the marking scheme - 4

asking other examiners about some students' scripts, revealing results - 4

irresponsibility - 2

emotional behaviour - 5

disturbing students - 2

being hard on students - 2

What qualities should an examiner possess with regards to …?

- **His/her personality**

being objective – 18

reliability - 17

honesty – 14

precision – 12

patience - 10

kindness - 10

conscientiousness - 8

unbiased - 7

unemotional - 6

justice - 6

openness – 5

endurance - 4

punctuality - 4

ability to organise one's time – 3

discipline - 3

firmness - 3

responsibility - 3

systematic - 3

need to improve one's abilities - 2

independence - 2

empathy - 2

assertiveness - 2

stressproof – 2

ability to confess one's mistakes, precision, carefulness, understanding, ability to co-operate, erudition, credibility, hard-working, concentration, respect for students, modesty - all once only

His/her professional life

professional - 26

precise – 3

involved, experienced in working at school - 4

experienced in preparing candidates for public exams - 9

in marking matura tests - 2

responsible - 7

reliable - 11

improving one’s abilities - 8

erudition, punctual - 2

honest - 6

well organised at work - 3

objective - 6

open to changes - 3

lack of routine, able to co-operate, creative, dutiful - 5

curious, just, self-demanding, conscientious - 5

unbiased –

systematic – 2

What are the duties of an examiner with regard to …?

A student he/she assesses

being objective - 29

reliable assessment of students knowledge and skills - 17

justice - 9

precision - 2

reliability - 9

being friendly - 5

responsibility – 2

precise counting of points –

being open - 2

having good intentions - 2

observing instructions – 2

honesty –

none with regards to students but with regards to the scripts - using objective criteria, unbiased - 7

kindness - 9

assessment using marking scheme - 2

sensitivity, concentration, interested in work, demanding precise answers, respectful to students, creating appropriate (friendly) conditions for students - all once only

Regional Examination Board

co-operation, the best performance of an examiners’ duties – 13 loyalty –

responsibility - 7

discipline - 3

professionalism –

reliability - 16

easy to keep in touch –

precision - 5

doing things on time - 24

ability to signal doubts and problems - 4

honesty - 5

treating the work confidentially - 2

precision in using marking scheme, fill in documents properly - 2

observing rules - 2

participation in meetings, and conferences - 2

keeping in touch with OKE - 2

conscientiousness – 2

Other examiners

co-operative - 22

helpful (if necessary) - 19

understanding - 2

friendly - 12

able to learn, kind - 3

discreet - 2

exchange experiences - 12

open - 5

respectful to other examiners - 2

assertive - 2

loyalty - 6

discipline, honesty - 4

punctuality, solidarity, keeping established rules - all once only

Remarks on examiners' behaviour during the Matura exam (May 2002)

marking was done in an appropriate (great) atmosphere - 3

great friendliness - 2

help from the part of other examiners, full co-operation, feeling uncertain, being objective, observing instructions - 3

All once only

nothing against
examiners were reliable in their work
knew the instructions
co-operated with co-ordinator eagerly
maths examiners were perfect in their work
constructive meetings and discussions
everybody worked actively
everything was OK
results were better than we expected
very professional

APPENDIX G

DESCRIPTIVE STATISTICS OF THE QUESTIONNAIRE ITEMS

Table 6: Descriptive statistics of the test developers' and test users' level of making ethical choices in using English language tests.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
1. I believe that reliability is important in language testing ethics.	28	23	82.1	5	17.9	------	-----	-----	-----	-----	-----	4.8214
2. I believe that validity is important in language testing ethics.	28	24	85.7	3	10.7	1	3.6	-----	-----	-----	-----	4.8214
3. I believe that test bias is important in language testing ethics.	28	11	39.3	13	46.4	4	14.3	-----	-----	-----	-----	4.2500
4. I believe that washback is important in language testing ethics.	28	11	39.3	14	50.0	3	10.7	-----	-----	-----	-----	4.2857
5. I believe that test impact is important in language testing ethics.	28	13	46.4	12	42.9	2	7.1	1	3.6	-----	-----	4.3214

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 7: Descriptive statistics of the stakeholders level of making ethical choices in using English language tests.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
6. I believe that the stakeholder	28	8	28.6	11	39.3	9	32.1	-----	-----	-----	-----	3.9643

involvement in test preperation can contribute to the language testing process.												
7. I believe that the stakeholder involvement in test practice can contribute to the language testing process.	28	5	17.9	10	35.7	10	35.7	3	10.7	-----	-----	3.6071
8. I believe that the stakeholder involvement in test evaluation can contribute to the language testing process.	28	4	14.3	10	35.7	12	42.9	2	7.1	-----	-----	3.5714
9. I believe that the stakeholder involvement in making decisions can contribute to the language testing process.	28	8	28.6	16	57.1	2	7.1	2	7.1	-----	-----	4.0714

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 8: **Descriptive statistics of different types of stakeholders' level of responsibility in ethical use of English tests.**

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
10. I believe that a tester should be accountable for all possible consequences of language test use.	28	14	50	11	39.3	1	3.6	1	3.6	1	3.6	4.2857
11. I believe that test takers are responsible for ethical test use in language programs.	28	12	42.9	10	35.7	3	10.7	1	3.6	2	7.1	4.0357
12. I believe that families are responsible for ethical test use in language programs	28	2	7.1	4	14.3	6	21.4	10	35.7	6	21.4	2.5000
13. I believe that policy makers are responsible for ethical test use in language programs	28	7	25.0	9	32.1	9	32.1	3	10.7	----	-----	4.0000
14. I believe that test users are responsible for ethical test use in language programs.	28	12	42.9	8	28.6	6	21.4	1	3.6	1	3.6	4.0000
15. I believe that test developers are responsible for ethical test use in language programs.	28	14	50	8	28.6	3	10.7	2	7.1	1	3.6	4.5000

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 9: Descriptive statistics of the level of contribution of the rights 'confidentiality' and 'access to information' to ethical use of English tests.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
16. I believe that confidentiality is important with regard to the rights of test takers in language testing ethics	28	17	60.7	9	32.1	2	7.1	-----	-----	-----	-----	4.5357
17. I believe that the right to access to information is important with regard to the rights of test takers in language testing ethics.	28	19	67.9	3	10.7	5	17.9	1	3.6	-----	-----	4.4286

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 10: Descriptive statistics of language test givers level of acceptance to using language tests for non-intended purposes

Statement		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
18. Language tests are used for particular purposes such as assessment, diognosis etc. I believe that it is appropriate to use them for non-intended purposes. e.g. introducing a new education policy.	28	2	7.1	3	10.7	6	21.4	11	39.3	6	21.4	2.4286

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 11: Descriptive statistics of language test givers level of awareness to availability of a 'code of ethics' for language test use.

Statements		YES		I DO NOT KNOW		NO		$\overline{X}$
	N	n	f	n	f	n	f	
1. There is a 'code of ethics for language test use' in Turkey.	28	4	14.3	14	50.0	9	32.1	2.2857

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 12: Descriptive statistics of how much test givers believe in the necessity of creating a code or codes for ethical use of language tests and reasons for creating a code of ethics.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
2. A 'code of ethics for language test use' is necessary to ensure fairness in language testing.	28	17	60.7	7	25.0	3	10.7	1	3.6	----	-----	4.4286
3.Because;												
• appropriate behaviours should be stated in a written document.	28	16	57.1	9	32.1	2	7.1	1	3.6	-	-----	4.4286
• unclear situations should be explained	28	20	71.4	6	21.4	2	7.1	-	-----	-	-----	4.6429
• the personal qualities of an examiner should be shown	28	3	10.7	5	17.9	8	28.6	7	25.0	5	17.9	2.7857

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 13: Descriptive statistics of the level of acceptance of language test givers to unethical behaviors in language testing.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\bar{X}$
4. A ‘code of ethics for language test use’ is not necessary for fairness in language testing, because;	N	n	f	n	f	n	f	n	f	n	f	
• teachers are honest people and aware of their responsibilities.	28	2	7.1	2	7.1	10	35.7	7	25.0	7	25.0	2.4643
• there is already a code of professional ethics.	28	2	7.1	4	14.3	10	35.7	5	17.9	7	25.0	2.6071

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 14: Descriptive statistics regarding the test givers thoughts about the content of a ‘code of ethics’ for language test use.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\bar{X}$
5. What should ‘a code of ethics for language test use’ include?	N	n	f	n	f	n	f	n	f	n	f	
• rules regarding what is and what is not allowed	28	19	67.9	9	32.1	-----	-----	-----	-----	-----	-----	4.6786
• examiners’ responsibilities.	28	16	57.1	8	28.6	2	7.1	1	3.6	1	3.6	4.3214
• rules regarding how to punish dishonest examiners	28	12	42.9	7	25.0	4	14.3	3	10.7	2	7.1	3.8571
• appropriate rules of behaviour during the exam.	28	18	64.3	9	32.1	1	3.6	-----	-----	-----	-----	4.5714
• discipline	28	11	39.3	7	25.0	7	25.0	1	3.6	2	7.1	3.8571

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 15: **Descriptive statistics regarding the test givers thoughts about the unethical behaviors of a test giver**

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
1. What do you consider to be the unethical behaviors of a test giver?	N	n	f	n	f	n	f	n	f	n	f	
• asking someone else to mark the testee's paper.	28	18	64.3	3	10.7	3	10.7	3	10.7	3	10.7	4.2143
• lack of confidentiality	28	19	67.9	5	17.9	2	7.1	1	3.6	1	3.6	4.4286
• marking the papers subjectively	28	23	82.1	3	10.7	1	3.6	-----	------	1	3.6	4.6786
• helping one's own students	28	22	78.6	3	10.7	-----	------	1	3.6	2	7.1	4.5000

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 16: Descriptive statistics regarding what the test givers consider the unethical behaviors of a test taker

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
2. What do you consider the unethical behaviors of a test taker?	N	n	f	n	f	n	f	n	f	n	f	
• cheating during the exam	28	25	89.3	2	7.1	------	------	----	------	1	3.6	4.7857
• helping other student(s) during the exam	28	22	78.6	4	14.3	1	3.6	----	------	1	3.6	4.6429

• interrupting other students during the exam	28	19	67.9	4	14.3	4	14.3	----	------	1	3.6	4.4286
• unpunctuality (not arriving at the exam venue on time).	28	12	42.9	7	25.0	6	21.4	2	7.1	1	3.6	3.9643

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 17: Descriptive statistics regarding the test givers thoughts about the qualities of a test giver.

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
3. What should be the qualities of a language test giver?	N	n	f	n	f	n	f	n	f	n	f	
• objective	28	27	96.4	-----	------	1	3.6	------	------	--	------	4.9286
• honest	28	23	82.1	5	17.9	-----	------	------	------	--	------	4.8214
• patient	28	16	57.1	7	25.0	5	17.9	------	------	--	------	4.3929
• unbiased	28	26	92.9	2	7.1	-----	------	------	------	--	------	4.9286
• disciplined	28	17	60.7	5	17.9	4	14.3	1	3.6	1	3.6	4.2857
• responsible	28	22	78.6	5	17.9	1	3.6	------	------	--	------	4.7500
• empathetic	28	12	42.9	12	42.9	2	7.1	2	7.1	--	------	4.2143
• kind	28	14	50	5	17.9	6	21.4	3	10.7	--	------	4.0714
• punctual	28	21	75	5	17.9	2	7.1	------	------	--	------	4.6786
•understanding	28	15	53.6	7	25.0	4	14.3	2	7.1	--	------	4.2500
• hardworking	28	13	46.4	8	28.6	4	14.3	3	10.7	--	------	4.1071
• cooperative	28	20	71.4	4	14.3	2	7.1	2	7.1	-	------	4.5000

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)

Table 18: Descriptive statistics of the test givers level of considering value systems of the society in ethical use of English language tests

Statements		5(SA)		4(A)		3(N)		2(D)		1(SD)		$\overline{X}$
	N	n	f	n	f	n	f	n	f	n	f	
1. Value systems of the society affect language test use in education.	28	11	39.3	12	42.9	3	10.7	2	7.1	-----	-----	4.1429
2. Individual rights of test takers should be considered by test developers and users to ensure fairness in language testing.	28	14	50.0	10	35.7	4	14.3	-----	-----	-----	-----	4.3571
3. Individual differences of test takers should be considered by test developers and users to ensure fairness in language testing.	28	9	32.1	9	32.1	6	21.4	3	10.7	1	3.6	3.7857
4. There are different standards for different examiners.	28	4	14.3	9	32.1	9	32.1	1	3.6	5	17.9	3.2143
5. The moral values vary from	28	11	39.3	11	39.3	2	7.1	-----	-----	4	14.3	3.8929

one culture to another.												
6. The moral values vary from one situation to another.	28	5	17.9	9	32.1	5	17.9	4	14.3	5	17.9	3.1786
7. The moral values vary from one individual to another.	28	5	17.9	11	39.3	5	17.9	3	10.7	4	14.3	3.3571
8. Adopting the 'code of ethics for language test use' of another country is useful, since different cultures have common ethical values.	28	3	10.7	7	25.0	6	21.4	5	17.9	7	25.0	2.7857

Definitions: SA(Strongly Agree), A(Agree), N (Neutral), D(Disagree), SD(Strongly Disagree)