

145544

TESTERS' PERCEPTIONS OF THE TEST DEVELOPMENT PROCESS
AND TEACHERS' AND TESTERS' ATTITUDES
TOWARDS THE RESULTING ACHIEVEMENT TESTS
AT MUĞLA UNIVERSITY SCHOOL OF FOREIGN LANGUAGES

The Institute of Economics and Social Sciences
of
Bilkent University

by

ELİF AYDIN

In Partial Fulfillment of the Requirements for the Degree of
MASTER OF ARTS

in

THE DEPARTMENT OF
TEACHING ENGLISH AS A FOREIGN LANGUAGE
BILKENT UNIVERSITY
ANKARA

June 2004

BILKENT UNIVERSITY
INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES MA THESIS
EXAMINATION RESULT FORM

JUNE 28, 2004

The examining committee appointed by the Institute of Economic and Social
Sciences for the thesis examination of the MA TEFL student

Elif Aydın

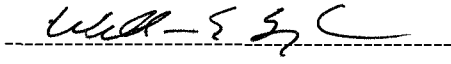
has read the thesis and has decided that the thesis of the student is satisfactory.

Title: Testers' Perceptions of the Test Development Process and
Teachers' and Testers' Attitudes towards the Resulting
Achievement Tests at Muğla University School of Foreign
Languages

Thesis Supervisor: Dr. Bill Snyder
Bilkent University, MA TEFL Program

Committee Members: Dr. Julie Mathews Aydınlı
Bilkent University, MA TEFL Program
Assistant Professor Gölge Seferoğlu
Middle East Technical University, Department of Foreign
Language Education

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



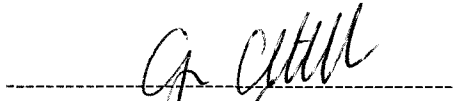
(Dr. Bill Snyder)
Supervisor

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



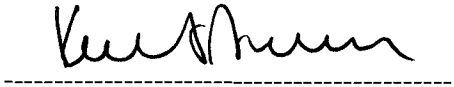
(Dr. Julie Mathews Aydın)
Examining Committee Member

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



(Assistant Professor Gölge Seferoğlu)
Examining Committee Member

Approval of the Institute of Economics and Social Sciences



(Prof. Dr. Kürşat Aydoğan)
Director

ABSTRACT

TESTERS' PERCEPTIONS OF TEST DEVELOPMENT PROCESS AND TEACHERS' AND TESTERS' ATTITUDES TOWARDS THE RESULTING ACHIEVEMENT TESTS AT MUĞLA UNIVERSITY SCHOOL OF FOREIGN LANGUAGES

Elif Aydın

M.A., Department of Teaching English as a Foreign Language

Supervisor: Dr. Bill Snyder

Co-Supervisor: Dr. Julie Mathews-Aydınlı

June 2004

This study investigated testers' perceptions of test development process at Muğla University School of Foreign Languages and teachers' and testers' attitudes towards the resulting achievement tests. 30 English teachers and five testers who currently work at Muğla University School of Foreign Languages participated in this study.

Two data collection instruments were employed in this study. First, the teachers and testers were given a questionnaire. Second, interviews with five testers and five randomly chosen teachers were carried out after the analysis of the questionnaires.

Analysis of data revealed that both teachers and testers working for Muğla University School of Foreign Languages have positive attitudes towards the tests used in the institution. Participants stated that the tests closely match with teaching practices and provide useful feedback on the learning process. The data also pointed

to weaknesses such as the lack of explicit learning objectives in the curriculum, the possible negative effects of the tests on teaching and learning, the lack of test specifications for the test design process and the limited use of authentic tasks in the tests. The data gained from the interviews indicated that there is no active cooperation between the teachers and testers in the test preparation process. The findings of the questionnaires showed that most of the teachers do not want to work in the test unit, and the testers have complaints about having excessive working hours, and not being rewarded for their work in the testing unit.

The results of this study suggest that explicit curriculum objectives are needed at Muğla University School of Foreign Languages in order to achieve better testing practices. The testing work may be more attractive in the institution if the testers are given financial support or their teaching hours are reduced.

Key Words: Achievement Tests, Test Specifications, Content Validity, Washback

ÖZET

MUĞLA ÜNİVERSİTESİ YABANCI DİLLER YÜKSEK OKULUNDAKİ SINAV SORUMLULARININ SINAV GELİŞİM SÜRECİNE YÖNELİK ALGILARI VE, ÖĞRETMENLERİN VE SINAV SORUMLULARININ UYGULANAN BAŞARI SINAVLARINA YÖNELİK TUTUMLARI

Elif Aydın

Yüksek Lisans, Yabancı Dil Olarak İngilizce Öğretimi Bölümü

Tez Yöneticisi: Dr. Bill Snyder

Ortak Tez Yöneticisi: Dr. Julie Mathews-Aydınlı

Haziran 2004

Bu çalışma Muğla Üniversitesi Yabancı Diller Yüksek Okulundaki sınav sorumlularının sınav gelişim sürecine yönelik algılarını, ve öğretmenlerin ve sınav sorumlularının uygulanan başarı sınavlarına yönelik tutumlarını incelemiştir. Bu çalışmaya, halen Muğla Üniversitesi Yabancı Diller Yüksek Okulunda çalışan 30 İngilizce öğretmeni ve beş sınav görevlisi katılmıştır.

Bu çalışmada iki veri toplama aracından faydalanılmıştır. İlk olarak, öğretmenlere ve sınav sorumlularına anket verilmiştir. İkinci olarak da, anketlerin veri analizinin ardından, beş sınav sorumlusu ve rastgele seçilen beş öğretmen ile mülakatlar yapılmıştır.

Veri analizi Muğla Üniversitesi Yabancı Diller Yüksek Okulunda çalışan sınav sorumluları ve öğretmenlerin kurumda kullanılan başarı sınavlarına karşı olumlu tutumlar geliştirdiğini ortaya koymuştur. Katılımcılar sınavların öğretim uygulamalarıyla yakından uyum sağladığını ve öğrenim üzerine yararlı geridönüt

sağladığını belirtmişlerdir. Veri, ayrıca müfredatta açık öğrenim amaçlarının eksikliği, sınavların öğretim ve öğrenim üzerindeki mümkün olabilecek olumsuz etkileri, sınav tasarım süreci için hazırlanan sınav tarif namesinin eksikliği, ve sınırlı sayıda otantik sınav sorularının kullanımı gibi zayıf noktaları ortaya çıkarmıştır. Mülakatlardan elde edilen veri, sınav hazırlama sürecinde öğretmenler ve sınav sorumluları arasında aktif işbirliği olmadığını ortaya koymuştur. Anketlerin bulguları, öğretmenlerin çoğunun sınav biriminde çalışmak istemediğini, ve sınav sorumlularının çok yoğun çalışma saatleri, sınav biriminde çalışmalarının karşılığını alamadıkları konusunda şikayetleri olduğunu göstermiştir.

Bu çalışmanın sonuçları, daha iyi sınav uygulamaları elde etmek için, Muğla Üniversitesi Yabancı Diller Yüksek Okulu'nda açık müfredat amaçlarına ihtiyaç olduğunu öne sürmektedir. Sınav sorumluları mali açıdan desteklendirildiği ve ders saatleri azaltıldığı takdirde, kurumda sınav hazırlama çalışması daha cazip hale getirilebilir.

Anahtar Kelimeler: Başarı Sınavları, Sınav Tarifnamesi, İçerik Geçerliliği, Sınavın Öğretim Sürecine olan Etkisi

ACKNOWLEDGEMENTS

I would like to thank and express my deepest gratitude to my thesis advisor, Dr. Bill Snyder, for his genuine interest, his contributions, invaluable guidance and patience throughout the preparation of my thesis.

Special thanks to Dr. Julie Mathews-Aydınlı for her assistance, and contributions throughout the preparations of my thesis. I would like to express my thanks to Dr. Martin Endley and Dr. Kimberly Trimble for their support and guidance throughout the year.

I am gratefully indebted to Assistant Professor Şevki Kömür who is the former director of Muğla University School of Foreign Languages because he encouraged me to attend the MA TEFL Program. I owe much to Assistant Professor Mustafa Kinsız who is the current director of Muğla University School of Foreign Languages because he supported my thesis research.

I would like to thank my colleagues at Muğla University School of Foreign Languages who participated in this study.

Many thanks to my friends in the MA TEFL Program for the wonderful relationships we shared.

Finally, I am grateful to my family for their continuous encouragement and support throughout the year and for their love throughout my life.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZET	vi
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF TABLES	xiii
CHAPTER I: INTRODUCTION	1
Introduction	1
Background of the Study	2
Statement of the Problem	5
Research Questions	6
Significance of the Study	7
Key Terminology	7
Conclusion	8
CHAPTER II: LITERATURE REVIEW	9
Introduction	9
Achievement Tests	9
Test Development Process	13
Construction of Test Specifications	14
Item-Writing Process	19
Evaluation of Language Tests	24
Validity	24
Test Validation Process	25

Internal Validation	26
Face Validation	26
Content Validation	27
Construct Validation	28
Reliability	29
Authenticity	30
Practicality	33
Interactiveness	34
Washback	34
Studies Conducted on Achievement Tests	39
Conclusion	41
CHAPTER III. METHODOLOGY	42
Introduction	42
Setting	42
Participants	44
Instruments	46
Questionnaires	46
Teachers' Questionnaire	46
Testers' Questionnaire	47
Interviews	49
Data Collection Procedures	50
Data Analysis	51
Conclusion	52
CHAPTER IV. DATA ANALYSIS	53
Introduction	53

Data Analysis Procedures	53
Testers' Perceptions of the Test Development Process	55
Cooperation during the Test Development Process	55
Testers' Perceptions of the Test Preparation Process	56
Testers' Perceptions of the Item Writing Process	60
Teachers' and Testers' Attitudes towards the Achievement Tests	61
Teachers' and Testers' General Opinions about the Tests	62
Teachers' Opinions on Preparing the Tests for Their Classes	63
Cooperation between the Teachers and Testers	65
Teachers' and Testers' Perceptions of the Content of the Tests	70
Teachers' and Testers' Perceptions of the Test Tasks	74
Teachers' and Testers' Perceptions of the Test Items and Instructions	76
Relationship between Learning and the Tests	79
Teachers' and Testers' Perceptions of Effect of Tests on Teaching	81
Teachers' and Testers' Perceptions of Working in the Test Office	83
Teachers' and Testers' Additional Comments on Testing Practices	84
Conclusion	89
CHAPTER V: CONCLUSION	91
Introduction	91
Overview of the Study	91
Discussion of Findings	93
Pedagogical Implications	98
Limitations of the Study	99
Implications for Further Research	100
Conclusion	101

REFERENCE LIST 102

APPENDIX A. TEACHERS’ QUESTIONNAIRE 107

APPENDIX B. TESTERS’ QUESTIONNAIRE 111

APPENDIX C. INTERVIEW SCHEDULE (FOR TEACHERS) 116

APPENDIX D. INTERVIEW SCHEDULE (FOR TESTERS) 117

APPENDIX E. INFORMED CONSENT FORM FOR INTERVIEWS 118

APPENDIX F. SAMPLE INTERVIEW 119

APPENDIX G. ANALYSIS OF SAMPLE INTERVIEW 121

APPENDIX H. ANALYSIS OF OPEN-ENDED QUESTIONS 122



LIST OF TABLES

<u>Table</u>	<u>Page</u>
1. Weighting of the Students' Assessment Criteria	43
2. Background Information about Teachers	44
3. Background Information about Testers	45
4. Distribution of Questions on the Teachers' Questionnaire	46
5. Distribution of Questions on the Testers' Questionnaire	48
6. Issues Covered in the Interviews	49
7. Items Related to Cooperation in the Test Development Process	55
8. Items Related to the Test Preparation Process	57
9. Items Related to the Item Writing Process	60
10. Teachers' and Testers' General Opinions about the Tests	62
11. Teachers' Opinions on Preparing the Tests for Their Classes	64
12. Items Related to the Cooperation between the Teachers and the Testers	65
13. Items Related to the Content of the Tests	70
14. Items Related to the Test Tasks	74
15. Items Related to the Test Items and Instructions	77
16. Items Related to the Relationship between Learning and the Tests	80
17. Items Related to Effect of the Tests on Teaching	81
18. Teachers' and Testers' Perceptions of Working in the Test Office	83
19. Teachers' Questionnaire Section III	85
20. Testers' Questionnaire Section IV	87

CHAPTER I: INTRODUCTION

Introduction

Tests are most commonly used in many language programs as a form of assessment. When tests are used systematically to determine whether the learners have attained the objectives of the program, test results may also be used to evaluate the effectiveness of teaching practices and the program as well (Bachman, 1990; Bachman & Palmer, 1996; Brown, 1996; Cohen, 1994; Genesee & Upshur, 1996).

In large educational institutions, test developers or testers are generally responsible for the preparation of tests which aim to serve the goals and objectives of the language program. The process of test development needs to be clear to stakeholders in order to ensure that the purpose and the content of the tests closely coincide with teaching practices and the program goals. In situations in which teachers are not actively involved in the test development process, they may be provided with opportunities to make comments about the tests (Davies & Pearse, 2000). Since teachers know best about teaching practices, they may evaluate the adequacy of the tests in terms of the course content and instructional objectives. They may also determine which objectives have been met by the learners and where changes might have to be made based on the test results (Weir, 1995). Such feedback may lead to appropriate changes in the curriculum design, thus contributing to curriculum development (Brown, 1996).

The aim of this study is to investigate testers' perceptions of the test development process at Muğla University School of Foreign Languages, and to explore teachers' and testers' attitudes towards the resulting achievement tests. The study looks at the ongoing test development process and identifies whether any discrepancies exist between teachers' and testers' attitudes towards the achievement tests. 30 English teachers and five testers participated in this survey study. Data was collected through two parallel questionnaires distributed to teachers and testers, and follow up interviews carried out with five testers and five randomly chosen teachers.

Background of the Study

Language tests are a set of tasks attempted by learners in order to demonstrate their language abilities and knowledge of language. These tests generally fall into five categories: proficiency, placement, diagnostic, aptitude and achievement tests. The focus of this study is on achievement tests because the research was conducted to investigate testers' and teachers' attitudes towards these tests as used at Muğla University School of Foreign Languages.

McNamara (2000) states that "tests which are closely associated with the process of instruction are achievement tests" (p. 5). In most educational settings, achievement tests are designed with reference to specific objectives of a course or curriculum in order to measure the learners' mastery of these objectives. Such tests are considered to be criterion-referenced because the learners' test scores are compared with the level of mastery achieved, rather than the other learners who take the same test (Brown, 1996). Achievement tests are well suited to provide systematic feedback on achievement level of learners in the language learning, the effectiveness of teaching practices and the language program itself (Bachman, 1990; Bachman & Palmer, 1996; Bailey, 1998; Brown, 1996; Brown & Hudson, 2002; Cohen, 1994;

Davies & Pearse, 2000; Genesee & Upshur, 1996; McNamara, 2000; Spolsky, 1995; Weir, 1995). Teachers may use the results of achievement tests to evaluate the effectiveness of the instruction in terms of the teaching techniques and teaching practices which take place in the class (Genesee & Upshur, 1996). Additionally, the results of achievement tests may lead to meaningful changes or adaptations to the syllabus or curriculum (Brown, 1996). Because the content of these tests are often based on the curriculum, it may be possible for the stakeholders to determine which specific objectives have been met and where changes might have to be made in the curriculum (Bachman, 1996; Brown, 1996; Weir, 1995).

In institutions with a large population of teachers and learners, it is not often a matter of personal choice for teachers when to assess and how to assess their learners. Test developers or testers are responsible for the preparation of the tests which aim to serve the purpose of the language program in such institutions. Testers are expected to develop a representative sample of what the students have been taught and what they have been expected to learn in a course or in curriculum (Alderson et al., 1995; Brown, 1996; Cohen, 1994). It is important for testers to specify the range of language learning appropriately according to teaching practices and learning activities. It may be possible for testers to improve quality of tests by analyzing and comparing the content of the tests with the course content and the objectives in the curriculum, or gathering judgments of other stakeholders in the program about the tests (Alderson et al., 1995; Brown & Hudson, 2002).

In the cases where tests do not match with the course content and curriculum, problems are likely to arise due to the lack of content relevance (Alderson et al, 1995; Bachman & Palmer, 1996; Brown, 1996; Cohen, 1994; Davies, 1990). To Bachman and Palmer (1996), these problems may affect negatively teaching,

learning and the ongoing program. Teachers may not obtain accurate information about their learners' actual performance and their progress in the language learning process, and they may not be able to evaluate the effectiveness of the ongoing instruction. Tests which lack content relevance may also lead administrators to make inappropriate decisions about the curriculum since they may not obtain accurate feedback on effectiveness of the program.

To ensure that the tests serve the goals of the program in which they are used, testers need to demonstrate to the other stakeholders that the tests are carefully organized. Bachman and Palmer (1996) suggest that "careful planning of test development process is crucial in all language testing situations, for three reasons. First, it assures that the test will be useful for its intended purpose... Second, it tends to increase accountability.... Then, it increases the amount of satisfaction test developers experience..." (p. 86). The test development process should be carefully planned because this assures that the tests will be useful serving their expected purposes as intended. The starting point in the test development process should ideally be the construction of test specifications (Alderson et al., 1995; Bachman & Palmer, 1996; Brown & Hudson, 2002; Davidson & Lynch, 2002; Lynch & Davidson, 1994; McNamara, 2000). These specifications are generally the statement of goals and objectives taken from the curriculum and syllabus, description of test tasks, sample items, and specifications for the responses or specified scoring criteria (Brown & Hudson, 2002; Davidson & Lynch, 2002).

The second step in the test preparation is the item-writing process in which the actual test tasks and items are selected and written by test writers. In any situation, test writers need to consider carefully the level of the learners, the learners' familiarity with the items, avoidance of ambiguity, avoidance of irrelevant

information, independence of the items, the item organization, revision of the items, and time allotment while writing the test items (Alderson & Clapham, 1995; Brown, 1996; Brown & Hudson, 2002; Brown, 2004; Cohen, 1994; Huerta-Macias, 1995; Genesee & Upshur, 1996; Hughes, 1989; Kenyon, as cited in Kunnan, 1998; Ransom & Santa, 1999; Rudner & Schafer, 2001; Weir, 1995). Genesee and Upshur (1996) suggest that the item-writing process can proceed only when a clearly agreed upon set of objectives is available. When the objectives are explicitly defined in the curriculum, it will be possible for testers to specify the content appropriately in test specifications. Then, testers may select and write appropriate test tasks and items with reference to well-defined curriculum objectives, thus ensuring the content relevance of the tests.

Tests used systematically in educational programs need to be evaluated in terms of their quality by the stakeholders in order to ensure the tests serve the goals and objectives of the educational program. Bachman and Palmer (1996) suggest a model of test usefulness in order to help the stakeholders evaluate the quality of tests. This test usefulness model is concerned with the qualities of “validity, reliability, authenticity, practicality, interactiveness and impact, or washback” (p. 17). These properties provide a basis for evaluation of language tests in the contexts where the tests are being used.

Statement of the Problem

Teachers may have either negative or positive attitudes towards the tests produced when they are not actively involved in the testing procedures. Teachers may have negative attitudes towards the tests when the tests do not match with the teaching objectives and learning activities which take place in the classroom. In this situation, teachers may not make appropriate decisions about the ongoing instruction

and their learners' progress in the language learning process. It is important for testers to plan the test development in such a way that learners, teachers and administrators can see that the tests serve their intended purposes, have content validity, and that the results are credible or reliable. If tests closely coincide with the learning objectives, teachers may feel confident in relying on the tests to support what they teach in the class and what their learners are expected to learn. Teachers should be provided with opportunities to make comments on the tests or evaluate the effectiveness of the tests. They may offer suggestions for modifications needed or appropriate changes. It is useful for testers to work with teachers cooperatively in order to avoid any mismatch between the tests and teaching practices.

At Muğla University School of Foreign Languages, teachers use achievement tests periodically in order to obtain feedback on their learners' progress in the language learning. These tests are prepared by testers who also currently work as teachers in the institution. Since other teachers are not actively involved in the test development process, they may have different perceptions of the tests from testers. This study looks at testers' perceptions of test development process at Muğla University School of Foreign Languages, and explores teachers' and testers' attitudes towards the resulting achievement tests.

Research Questions

1. What are the testers' perceptions of the test development process at Muğla University School of Foreign Languages?
2. What are the teachers' attitudes towards the resulting achievement tests used Muğla University School of Foreign Languages?
3. What are the testers' attitudes towards the resulting achievement tests used at Muğla University School of Foreign Languages?

4. Are there any discrepancies between the teachers' and testers' attitudes towards the achievement tests used at Muğla University School of Foreign Languages?

Significance of the Study

Many studies have been conducted in the field on different aspects of achievement tests, including stakeholders' perceptions of these tests, and the validity or reliability of achievement tests. The test development process also has been of interest to researchers in the field of English Language Teaching. However, little research has been done on testers' perceptions of test development process. Although teachers are central to their learners' assessment process, few studies investigated teachers' attitudes towards tests (Aksan, 2001; Dalyan, 1990; Kuntasal, 2001; Serpil, 2000). This study explored testers' perceptions of test development process including teachers' attitudes towards the resulting tests. The results of this study provided additional empirical evidence for teachers' attitudes towards achievement tests, and contributed to the literature on testers' perceptions of test development process.

At the local level, this study explored testers' perceptions of test development process at Muğla University School of Foreign Languages, and also teachers' and testers' attitudes towards the achievement tests used in the institution. The study will help testers and teachers evaluate the quality and adequacy of the tests in terms of course content and teaching objectives. The results will provide testers with feedback on teachers' attitudes towards the tests. The results of the study may be guidance for testers in order to improve the quality of the tests.

Key Terminology

The terms which are often mentioned in this study as follows:

Achievement Tests: The tests which are designed with particular reference to objectives of a course or a language program to measure the learners' mastery of these objectives (McNamara, 2000).

Test Specifications: "A set of guidelines as to what the test is designed to measure and what language content or skills will be covered in the test" (Brown & Hudson, 2002, p. 87).

Content Validity: The degree to which a test samples the content or objectives of the course or area being assessed adequately (Cohen, 1994).

Washback: The influence of tests in areas such as curriculum or syllabus design, testing practices, teaching methods, course materials, learning strategies, and knowledge of language to be tested (Shohamy, 2001).

Conclusion

In this chapter, a brief summary of the issues related to language tests and the test development process was given. The statement of the problem, research questions, the significance of the study, and key terms of the study were covered. The second chapter of the study is a review of literature on achievement tests, test development process, evaluation of language tests, and the local studies conducted on achievement tests. In the third chapter, participants, instruments, data collection procedures, and data analysis are presented. In the fourth chapter, the data analysis procedures and the findings are presented. In the fifth chapter, overview of the study, discussion of findings, pedagogical implications, limitations of the study, implications for further research are presented.

CHAPTER II: LITERATURE REVIEW

Introduction

This study investigates five testers' perceptions of test development process at Muğla University School of Foreign Languages. This study also explores the testers' and 30 English teachers' attitudes towards the achievement tests.

This chapter first reviews the literature on the roles of achievement tests in educational settings. This is followed by a discussion of the development process of language tests, which includes the construction of test specifications, and the item-writing process. The discussion of the evaluation of language tests in terms of such properties as validity, reliability, authenticity, practicality, interactiveness and washback will be the succeeding section in this chapter. Finally, the studies conducted on achievement tests will be discussed.

Achievement Tests

Language tests are a set of tasks attempted by learners in order to demonstrate their language abilities and knowledge of language. There are different types of language tests which generally fall into five categories: proficiency, placement, diagnostic, aptitude and achievement tests. The main focus of this study is on achievement tests because the research was conducted to investigate testers' and teachers' attitudes towards these tests as used at Muğla University School of Foreign Languages.

Achievement tests are tests which are designed with particular reference to objectives of a course or language program to measure the learners' mastery of these objectives. Such tests are considered to be criterion-referenced as the learners' scores are compared with the level of mastery achieved, rather than with the scores of other students (Bachman, as cited in Johnson, 1989; Brown, 1996; Bond, 1996; Lynch & Davidson, 1994; McNamara, 2001; Thomas, 2003). The language performance of the learners is generally measured according to an agreed-upon criterion or a standard syllabus. Achievement tests are well-suited to provide feedback on learners' achievement levels in the language, the effectiveness of teaching and the language program itself (Bachman, 1990; Bachman & Palmer, 1996; Bailey, 1998; Brown, 1996; Brown & Hudson, 2002; Cohen, 1994; Davies & Pearse, 2000; Genesee & Upshur, 1996; McNamara, 2000; Spolsky, 1995; Weir, 1995).

Achievement tests provide feedback on the learners' achievement level in the language with reference to specific learning objectives of a program (Bailey, 1998; Brown, 1996; Brown & Hudson, 2002; Brown, 2004; Spolsky, 1995; Weir, 1995). In an educational program, achievement tests are used to accumulate evidence for how much the learners have learned the content of a course and how successful they have been in achieving the objectives of that program (Bailey, 1998; Brown, 1996). The results of achievement tests provide the learners with periodic feedback on their progress in the language learning. Johnston (2003) proposes that "learners need to have a sense of how well they are doing: of their progress, of how their work measures up to expectations ..." (p. 77). Spolsky (1995) points out that learners may constantly test and examine their changing knowledge and language skills with the results of the achievement tests. Learners obtain systematic feedback on what they know and what they are able to do in the language based on the test results. Weir

(1995) states that the achievement tests should aim to indicate how successful the learning experiences had been for learners rather than to show in what respects they are deficient. In this way, learners may become motivated towards future learning experiences with a sense of accomplishment and mastery of the specified content area. Brown (1996) and Brown (2004) suggest that achievement tests may be related to diagnostic decisions as well, indicating learners' both overall strengths and weaknesses. These tests may help learners periodically monitor their weakness in the language as well as their overall strengths. According to Brown and Hudson (2002), "such diagnostic feedback affords the students the opportunity to refocus their energies on any remaining weaknesses in order to make their learning as effective and efficient as possible" (p. 31). The results of achievement tests will provide learners with information about which specific language skills they need to focus on more, thus leading the learners to establish individual learning goals both prior to and after the test. The learners may check periodically their own performance in terms of these individual goals with the aid of achievement tests.

Another use of achievement tests is to evaluate the effectiveness of teaching in a language program (Bachman, 1990; Brown, 1996; Brown & Hudson, 2002; Cohen, 1994; Genesee & Upshur, 1996; Spolsky, 1995). The purpose of language teaching should be to encourage learners to achieve a high degree of language learning. Brown (1996) states that "all language teachers are in the business of fostering achievement in the form of language learning" (p. 14). First, teachers need information about their learners' progress in the language learning in order to encourage their achievement of learning. Spolsky (1995) points out that achievement tests help the teachers continually check on their learners' progress to determine whether successful learning is taking place. The results of achievement tests indicate

how well the students have met the learning objectives, thus helping teachers evaluate the effectiveness of teaching and its methodology. According to Cohen (1994), “teachers can see how well the students achieved the course objectives with the help of these tests and check for any discrepancies between the learners’ actual performance and what they are expected to learn” (p. 7). Based on the results, teachers may make decisions regarding appropriate changes in teaching procedures and learning activities (Bachman, 1990). Teachers may plan and organize ongoing instruction according to the information gained from the achievement tests about what and how much their learners have learned. Brown and Hudson (2002) support this, stating that “teachers can use the information about the learners’ progress to decide how to tailor their teaching energies so the students will most benefit from the continuing instruction” (p. 31). According to the results of achievement tests, teachers may decide if they should do any remedial work on a certain subject (Genesee & Upshur, 1996). In this way, teachers may identify and plan remedial instruction in the areas in which a majority of the learners has demonstrated weaknesses. This may promote learning by helping the learners to master more of the objectives.

Finally, achievement tests may be used to evaluate the language program itself in an educational setting (Bachman, 1990; Brown, 1996; Weir, 1995). Since the content of achievement tests is generally based on the content of the curriculum, the results of these tests indicate what has been learned of the curriculum in terms of the learning objectives. To Bachman (1990), “the performance of students on achievement tests can provide an indication of the extent to which the expected objectives of the program are being attained and thus pinpoint areas of deficiency” (p. 62). Weir (1995) also suggests that the stakeholders may determine which

objectives have been met and where changes might have to be made based on the results of achievement tests. The stakeholders may use this information in order to determine the degree to which the learning objectives are appropriate and achievable for the learners. The results of achievement tests may lead the stakeholders to make meaningful and appropriate changes in the curriculum or syllabus design because these tests are periodic checks on the quality of language program offered (Brown, 1996). This implies that it is possible to improve a curriculum with appropriate subsequent changes based on the results of achievement tests in order to better suit the language needs of learners.

Test Development Process

In the business of language teaching, teachers are generally central to the assessment of their learners' progress in language learning. However, in educational institutions with large populations of teachers and learners, it is generally not a personal matter of choice for teachers to decide when and how to assess their learners. Test developers or testers generally prepare and administer the tests in these institutions. This makes testers responsible for providing accurate information about the learners' knowledge of language and their language abilities based on their performance in the tests. It is important for stakeholders to be able to rely on the test results in order to determine the effectiveness of the learning process, instruction and the program. Testers ideally need to demonstrate that the test development process is carefully planned to avoid any discrepancies between the tests and the stakeholders' expectations from the tests. Bachman and Palmer (1996) offer that "careful planning of test development is crucial in all language testing situations, for three reasons. First, it assures that the test will be useful for its intended purpose... Second, it tends to increase accountability... Then, it increases the amount of satisfaction test

developers experience...” (p. 86). A test development process generally involves the construction of test specifications and the item-writing process.

Construction of Test Specifications

The first and crucial step of the test development process should ideally be the construction of test specifications. Test specifications ensure careful planning of the tests, and facilitate the item-writing process, test scoring and contribute to reliability of the tests (Bachman & Palmer, 1996; Brown & Hudson, 2002; Davidson & Lynch, 2002; McNamara, 2000).

Test specifications are generally statements about what a test will measure and what the test will be like, referring to the objectives in the curriculum of a program or a standard syllabus. To Brown and Hudson (2002), “test specifications can be used for communicating the test writers’ intentions to the users of the test” (p. 87). The specifications will enable testers to demonstrate to stakeholders that the test development process is carefully planned and the test serves its intended purpose. To Bachman and Palmer (1996), having clear specifications of the test may ensure that “the performance on the test tasks will correspond as closely as possible to language use and that the test scores will be maximally useful for intended purposes” (p. 93). In the absence of test specifications, test development can potentially proceed with little clear direction (McNamara, 2000).

Test writers can benefit from test specifications both in writing the test tasks and items and also in the scoring process. The test specifications specify and limit the range of responses to be given for the test tasks with an answer key or criteria. This helps the scorers or raters to decide whether the responses are appropriate or acceptable. The test scores may vary by scorer if an answer key or criteria is not explicitly specified, which affects the reliability of test scores. Davidson and Lynch

(2002) state that “test specifications have an integral role in assuring that a test is consistent, the relationship of the specifications to a test is a major component of the reliability of the test” (p. 22). This makes the test results more dependable for making decisions about learners’ language abilities and knowledge of language.

According to Davidson and Lynch (2002), there are basically five components of test specifications: general description, prompt attributes section, the response attributes section, the sample item, the specification supplement. These sections help test developers to specify the objectives which will be measured in the tests and the scoring criteria.

According to Davidson and Lynch’s model, the first section of a test specification is the general description (GD), which expresses the test content with the test purpose, the language skills to be measured, and the reason for assessing those particular skills. In GD section, the overall test content is outlined with reference to specific learning objectives in the curriculum and the language skills to be assessed in the test. Madaus, Haney and Kreitzer (1996) indicate that “the first step in constructing a test is to define the domain, so one can readily decide whether a particular aspect of knowledge, or particular skill, task, ability or performance falls within the domain” (p. 39). McNamara (2000) supports that “establishing test content involves careful sampling from the domain of the test, that is, the set of tasks or the kinds of behaviors in the criterion setting...” (p. 25). Bachman and Palmer (1996) state that ‘the purpose of test task’ and ‘the definition of the construct to be measured’ need to be specified in any test specification. Brown and Hudson (2002) define the GD as the combination of ‘general test description’ and ‘skill area description’. Brown and Hudson further clarify that ‘general test description’ will involve “what the test is designed to do, the general test objectives and the test

format. In essence, overall test descriptor is an abstract of what the test looks like” (p. 87). The ‘skill area description’ presents the language skills to be measured with particular areas to be measured and the levels of achievement and proficiency are detailed. As Weir (1995) puts it, “a test should assess full range of appropriate skills and abilities as defined by the objectives of the syllabus...” (p. 23). Genesee and Upshur (1996) suggest that item writing process can proceed only when a clearly agreed upon set of objectives is available. When the learning objectives or specific test objectives are not explicit to the testers, they may have difficulties in deciding on which desired objectives or skills will be measured in the test. The overall purpose of the test and the language elements need to be specified by the test writers prior to the test preparation process. Alderson et al. (1995) specify that test writers need to identify which linguistic features will be tested, and whether the features will be a specific list of vocabulary, grammatical structures, speech acts or pragmatic features. In the GD section, test writers need to specify the objectives and skills to be tested in order to sample appropriately during the item-writing process.

The second component of a test specification is the prompt attributes (PA) section in Davidson and Lynch’s model of test specification. Davidson and Lynch (2002) explains the PA as “a detailed description of what the test takers will be asked to do, including the form of what they will be presented with in the test item or task, to demonstrate their knowledge or ability in relation to the criterion being tested.” (p. 23). This section allows the test writers to see the format of the tasks and the item types to be used in the test. Brown and Hudson (2002) clarify that the PA section “provides a series of statements that attempt to delimit the general class of material that the examinees will be responding to when answering the type of item involved. Here, any factors that constrain the item construction process should be defined....”

(p. 94). The PA section might include specification of the length of the test items, linguistic level and structure within the items, type of discourse genre to be used, and topical information to be used in the test items. Bachman and Palmer (1996) define this section technically as “identification and definitions of tasks in the target language use domain” (p. 173). In Bachman and Palmer’s model, this section refers to four components of communicative competence, and specifies the test tasks in terms of their grammatical, textual, functional, pragmatic, sociolinguistic and topical characteristics. In the PA section, the specifications constrain the variability of the test items and test tasks. “This section of specifications attempts to control the variability in the types of items which different test writers might generate” (Brown & Hudson, 2002, p. 94).

The third section of a test specification is the response attributes (RA) section which describes what test takers should do in the test. Brown and Hudson (2002) point out that the RA section “either defines the characteristics of the options from which the students will select their responses or presents the standards by which the students’ responses will be evaluated” (p. 94). McNamara (2000) mentions the RA section as the ‘response format’ which demonstrates the way in which the learners will be required to respond to the test materials. The RA section gives a description of the acceptable or unacceptable responses for the items or tasks, or indicates the degree to which the responses will be accepted as correct. Brown (1996) states that the RA section is “a clear description of the types of (a) options from which students will be expected to select their receptive language choices or responses, or (b) standards by which their productive language responses will be judged” (p. 77). Bachman and Palmer (1996) define the RA section as ‘expected response’ and present it as including five elements: channel (aural, written, visual, or else), form,

language (target or else), length (of response), type (item or prompt, or other), speed (of response).

According to Davies and Pearse (2000), if test writers limit how the learners will respond to a test task, this will likely increase the test reliability and the test scores can be trusted to a greater extent. In the RA section, it is possible for test writers to limit the responses and clarify the scoring criteria. Alderson et al. (1995) suggest that the response or scoring criteria given in this section need to be clear for raters or scorers, with specification of the importance of accuracy, appropriacy, spelling, length of utterance or script. The RA section is useful in order to minimize inter-rater reliability problems of tests because it specifies the expected response with brief criteria.

Another component of a test specification is the sample item (SI) section which shows the sample items which will be included in the actual tests. SI section will provide the testers with guidelines about how the items will be presented and about the instructions to be given for the test items and test tasks. McNamara (2000) mentions that test writers should sample the type of the materials with which candidates will have to engage in the test. The presentation of sample items to the test writers facilitates the item writing process because the test writers will have clear understanding of the actual items, and the instructions which will be given for each item. Brown and Hudson (2002) state that this section “serves as an example test item derived from the test specifications. Such an item should be presented along with any directions that will be given to the students” (p. 94). Bachman and Palmer (1996) propose that “the primary function of test instructions is insure that the test takers understand the exact nature of the test procedures and of the test tasks, how

they are to respond to these tasks, and how their responses will be evaluated” (p. 181).

The last section of a test specification is specification supplement (SS) part which may be optional for test writers. The SS section of a test specification provides more details about the assessment criterion included in the former sections. Brown and Hudson (2002) state that “limits for the content of a particular item may need to be delineated in the specification supplement” (p. 95). This section helps the test writers to specify to a greater extent the content of particular items; for example, “if gerunds and infinitives will be tested, this part may set out which verbs can be used with the test items” (Brown & Hudson, 2002, p. 94).

Item Writing Process

After the construction of test specifications, the next stage in the test development process is selecting and writing the actual test items which will be used in the test. Alderson et al. (1995) point out that “it is surprising how many test writers try to begin item writing by looking at past papers rather than specifications... due to the fact that many tests lack proper test specifications.” (p. 43). In any situation, test writers need to take the following into consideration: the level of the learners, the learners’ familiarity with the items, avoidance of ambiguity, avoidance of irrelevant information, independence of the items, item organization, revision of the items and time allotment (Alderson & Clapham, 1995; Brown, 1996; Brown & Hudson, 2002; Brown, 2004; Cohen, 1994; Huerta-Macias, 1995; Genesee & Upshur, 1996; Hughes, 1989; Kenyon, as cited in Kunnan, 1998; Ransom & Santa, 1999; Rudner & Schafer, 2001; Weir, 1995).

Learners’ proficiency level in the language is an important factor which needs to be taken into consideration while selecting and writing test items (Brown, 1996;

Brown & Hudson, 2002). The test items should match with the learners' level so that they will be able to demonstrate in the test what they know and what are able to do in the language. To Brown (1996), "each item should be written at approximately the level of proficiency of the students who will take the test" (p. 52). Brown and Hudson (2002) claim that "item writers should maintain the goal of creating a test that measures what examinees know, or what they can do, with regard to the domain being tested or to the particular program's objectives" (p. 99). This implies that if the learning objectives are clear for the testers, they may know the exact level of the learners and produce test items and tasks which are appropriate for the learners' level. Brown and Hudson (2002) argue that item writers should not produce items "whose language is at the levels of complexity above the examinees' level of language proficiency" (p. 61). First, the learners need to understand the items and what they are required to accomplish in a test task in order to be able to demonstrate their actual knowledge of language and language abilities. When the learners do not understand the language of the test item, the test scores may be ambiguous because answering correctly or incorrectly may be due to a lack of knowledge or skill being tested or an inability to process information in the item (Brown & Hudson, 2002). Thus, this will affect the test reliability or the consistency of the test results.

Next, the learners' familiarity with the test items needs to be considered while selecting and writing test items (Brown, 1996; Huerta-Macias, 1995; Ransom & Santa, 1999). According to Ransom and Santa (1999), the learners need to be familiar with the format of the test and the types of questions and responses required. If the learners are not familiar with the items and tasks in the test, they may be confused and may not complete the task although they know the correct answer (Huerta-Macias, 1995). Brown (1996) claims that one of potential errors which

affects the test reliability is presenting a type of item which is new to the students. This will affect the reliability of the test scores because it may not provide accurate information about the learners' actual performance in the language.

Writing unambiguous test items is another concern for test writers in the item writing process (Alderson & Clapham, 1995; Brown, 2004; Brown, 1996; Brown & Hudson, 2002; Genesee & Upshur, 1996; Kenyon, as cited in Kunnan, 1998; Rudner & Schafer, 2001). Test writers can write test items which are not ambiguous by giving clear indication of what the test takers are being asked to do (Alderson & Clapham, 1995; Genesee & Upshur, 1996). The test items and tasks should be clearly worded and easily understandable by the test takers in order to know what they have to do in each task. Brown (1996) and Kenyon (as cited in Kunnan, 1998) note that if test writers produce unclear and ambiguous test tasks, this may affect negatively the test performance of the learners. The use of ambiguous items may be a source of test unreliability (Brown, 2004; Rudner & Schafer, 2001). When learners misunderstand a test task, they may misinterpret what they are expected to do in that test task. Any misinterpretation of the test tasks and items may lead the learners to answer the questions wrong and this may also result in inaccurate interpretation of their actual performance on that given test.

Another concern for test writers is the avoidance of irrelevant information in test tasks and test items (Brown, 1996; Brown & Hudson, 2002). Brown (1996) suggests that test writers should "avoid including extra information which is irrelevant to the concept or skill to be measured" (p. 53). According to Brown and Hudson (2002), addition of irrelevant or unimportant information may also cause the test items and test tasks to be unclear and ambiguous for the learners. Any extra information will just take extra time for the students to read, which may affect their

performance in the test. The use of unnecessary and extra information may also lead the learners to process the unnecessary information in the material and complete the tasks irrelevantly, which may reduce the degree of consistency of the test results.

Independence of test items is another important factor which contributes higher quality in item selection and writing (Brown & Hudson, 2002; Weir, 1995). Brown and Hudson (2002) clarifies that being independent for the test items means being free of the previous item or the following one. Weir (1995) also mentions that test writers should avoid such interdependence of the test items. Answering one test item should not be dependent on the ability to answer another because it becomes difficult to make accurate estimate of the ability being measured. Brown and Hudson (2002) explain that the use of interdependent test items may affect the learners' performance because they may not connect one dependent test item to another, or if they answer the initial item wrong, this may make them answer the following one incorrectly.

The organization of the test items is another issue which needs consideration during the item-writing process (Brown, 1996; Brown & Hudson, 2002; Weir, 1995). Brown and Hudson (2002) argue that "test items should be clearly organized and items of like format should be grouped together within sections or subtests so examinees will not be confused by unnecessary shifts in the types of responses they are required to make" (p. 63). Brown (1996) also mentions the organization of the items stating "all the parts of each item should be on the same page" (p. 52). It is possible for the learners to answer an item incorrectly, for example, when they do not recognize or miss the correct answer which is on the following page. Weir (1995) offers another suggestion for the item organization indicating that "if there is a variety of tasks and items testing a particular ability, easier tasks and items should be

put first” (p. 23). This may encourage all the students at the initial phase of the test, and lead them to try and show their best for the rest of the test.

The revision of the test items helps test writers identify any problems with the items and decide on changes before the test administration (Alderson et al., 1995; Brown, 1996; Cohen, 1994; Hughes, 1989; Weir, 1995). One way of revising the test items is to operate a ‘pre-testing’ or ‘pilot testing’ process (Alderson et al., 1995; Cohen, 1994; Hughes, 1989). This process involves the procedures in which a test is administered first to a group as similar as possible to the one for which it is intended. The pre-testing session may help testers explore any additional problems with test items, time and instructions. This will be an opportunity for test writers revise and change any problematic item or any other problems with the test before the actual test administration.

Another way to check and revise the items is to have an outsider who has not written the test items. Weir (1995) suggests that “all tests must be carefully proofread and glaring mistakes eliminated.” (p. 24). Brown (1996) also specifies that item writers “should always have at least one or more colleagues... look over and perhaps take the test so that any problems may be spotted before the test is actually used to make decisions about students’ lives.” (p. 53). Such revision or feedback on the test items and the overall test may help test writers to identify any possible problems which may arise for the examiners and the learners during the test administration process.

The time allocated for the test tasks should be adequate for the learners to complete the tasks (Bachman, 1990; Brown, 1996; Weir, 1995). Brown (1996) mentions that the learners should have enough time to answer all the test items. A reasonable amount of time should be provided for the test tasks so that the learners

will be able to complete the tasks satisfactorily. According to Weir (1995), “if too little time is made available, stress will result and we will not be eliciting the students’ best performance.” (p. 24). The learners may find the test pointless if the allocated time is not enough, and this may cause the test to lack face validity for the learners. The learners may not be able to answer the items which they know if they are not provided with sufficient time. This may result in meaningless and inappropriate decisions about the learners’ knowledge of language and their language abilities. It is also suggested that test writers might indicate the length of time which is allocated for accomplishing the test tasks (Bachman, 1990). In this way, learners will know how much they should spend on the test and its parts, and thus they can complete the test tasks with efficient use of time.

Evaluation of Language Tests

Tests used systematically in educational programs need to be evaluated in terms of their quality by the stakeholders in order to ensure that the tests serve the goals of the educational program. Bachman and Palmer (1996) suggest that quality of tests can be evaluated “on the basis of a model of test usefulness” (p. 17). This test usefulness model is concerned with six qualities: validity, reliability, authenticity, practicality, interactiveness and impact/ washback.

Validity

Validity refers to the degree to which a test measures what it is intended to measure (Alderson et al. 1995; Bachman & Palmer, 1996; Brown, 1996; McMillan, 2000; McNamara, 2000). If a test claims to measure English writing ability, then it should measure the ability to write English. It is desirable for educators to base their decisions about the learning process on the tests which are actually testing what they aim to test. Tests should ideally provide the stakeholders with a measure which can

be interpreted as an indicator of learners' language abilities and performance (Bachman & Palmer, 1996; McNamara, 2000). Test users and test developers should investigate whether the tests are measuring the abilities and proficiencies, which are intended to measure (Brown, 1996; McMillan, 2000). This leads the stakeholders to evaluate the tests in terms of validity through what is known as test validation process.

Test Validation Process. Test validation refers to the procedures through which the interpretations made on the basis of test scores are justified (Alderson et al., 1995; Bachman & Palmer, 1996; Brown, 1996; McNamara, 2000). McNamara (2000) posits that the purpose of test validation is "to ensure the defensibility and fairness of the inferences about candidates that have been made on the basis of test performance" (p. 10). The test validation process starts with consideration of test design and continues with the gathering of evidence in order to support the interpretations made on the test scores (Alderson et al., 1995; Bachman & Palmer, 1996; McNamara, 2000). Alderson et al. (1995) emphasize that it is best to validate a test in as many ways as possible. They further suggest that "the more different types of validity that can be established, the better, and the more evidence that can be gathered for any type of validity, the better" (p. 171). The strategies for investigating the validity of a test generally fall into three types of validity: internal, external (or criterion-related) and construct validity. To Brown (1996), external validity is not often appropriate for criterion-referenced (CR) tests because it is based on correlational analysis with the assumption of normal distribution depending on test score variance. Since CR tests generally do not provide such variance in the test scores, internal and construct validity are applicable to the validation process for CR tests.

Internal Validation. Internal validation process involves the procedures through which the stakeholders investigate the internal validity, the perceived test content and its effect. Internal validity may be investigated in terms of three types of validation: face validity, content validity and response validity (Alderson et al., 1995; Cohen, 1994; Weir, 1990). Since this study does not involve the learners, response validity, a type of internal validity, will not be discussed because it requires gathering information from test takers.

Face Validation. Face validity refers to whether test looks valid in terms of measuring what it is supposed to measure (Alderson et al., 1995; Cohen, 1994; Weir, 1990). Investigation of face validity or the face validation process requires judgments of the stakeholders in a program about whether the tests look valid in terms of its intended purpose. Alderson et al. (1995) suggest that investigation of face validity is generally gathering intuitive judgments of non-expert test users such as learners, teachers, and administrators. These stakeholders may demonstrate their perceptions of tests through meetings, interviews or questionnaires after a test administration session. It is important to investigate face validity because the tests which do not appear valid to the stakeholders may not be taken seriously for their given purpose (Weir, 1990). If the tests are not considered valid, test writers need to evaluate the test content and revise the tasks that have been used in the tests. According to Cohen (1994), tests which appear to be indirect in terms of what they aim to measure may lack face validity. For example, if a test is intended to measure the learners' writing skills, multiple-choice format may not be accepted as a direct or valid measure by the learners. It is important for test writers to ensure face validity because its absence may distract and confuse the learners in the tests. This will affect the learners'

performance negatively, and thus the consistency and interpretability of the test scores.

Content Validation. Content validity refers to the extent to which a test is a representative sample of the content or skills to be measured (Alderson et al., 1995; Brown, 1996; Brown & Hudson, 2002; DeVincenzi, 1995; Sireci, 1998).

Determining the degree of content validation requires a pre-specified content of a domain, which refers to the objectives in educational programs. According to Sireci (1998), content validation typically involves judgments regarding the knowledge and skills measured by each item and comparing these judgments with the content domain. Brown and Hudson (2002) argue that content validation is concerned with whether “the content of a test is related to the content of a well-described course, that is the objectives and/ or item specifications...” (p. 213). The content relevance of tests to test specifications is generally investigated by experts, but test users and teachers may also be involved in content validation process. DeVincenzi (1995) claims that “teachers must be able to make accurate assumptions about the test content in order to influence administrative decisions about test use” (p. 181). Alderson et al. (1995) offer two procedures for content validation: the first alternative is the experts’ analysis and judgments of the test content with test specifications or curriculum/ syllabus. The second alternative is gathering teachers’ and test users’ judgments about the relevance of test content to desired skills to be measured and to the objectives the learners are required to master. Both resulting judgments will help test developers determine an estimate of the content validity of the test and its components. If the content validation procedures reveal problems, then test developers need to gather other sorts of evidence for validity such as face or construct validity. Having well-specified content or explicit objectives in the

curriculum, and the construction of test specifications can help the test writers assure the content validity of tests. Brown (1996) suggests that “clear item specifications can help to make items much more consistent and also more valid in the sense that, when specifications are used, the items will more likely to match those specifications, which in turn match the objectives of the test” (p. 234).

Construct Validation. Construct validity refers to the degree to which test scores reflect the ‘construct’ or language ability which it is claimed to measure (Bachman & Palmer, 1996; Bachman, Davidson & Milanovic, 1996; Brindley, 2001; Cohen, 1994; Coombe & Hubley, 1999; Gorsuch, 1997; McNamara, 2000). Before the construct validation process, test educators are required to have a thorough understanding of the ‘construct’ or the ability expected from learners (Gorsuch, 1997). According to Brindley (2001) ‘construct’ is definition of the nature of the abilities being assessed (p. 140). If a writing test is developed, the representatives of the program or testers need to specify first what is meant by ‘writing ability’, or which components of this ability are expected from the learners in a course or program. Such a specification provides a basis for examining the relationship between the test scores and the ability or abilities which the test has been intended to measure.

Coombe and Hubley (1999) suggest that construct validation is concerned with the fit between underlying theory or methodology of language learning approach and the testing. If the ‘construct’ or language ability is defined referring to communicative approach, the test should measure language ability adopting a communicative approach.

Bachman and Palmer (1996) state that construct validation is concerned with judgments about whether test scores reflect the ability which is measured and

whether the interpretations made on the basis of test scores are meaningful and appropriate. According to Cohen (1994), test developers need to provide evidence that the test scores are a true reflection of the areas of language ability to be measured. A test or its scores will likely to possess construct validity if the test content directly reflects the components of the 'construct' or specified language ability. McNamara (2000) puts it, tests may lack construct validity "if tests introduce factors that are irrelevant to the aspect of ability to be measured" (p. 53). In order to obtain valid test scores, both test developers and test users need to examine whether the test procedures and test items actually reflect what learners are expected to know and what language abilities they are required to acquire. Bachman et al. (1996) suggest that if test developers "carefully pay attention to both abilities to be measured and the characteristics of the tasks utilized, the interpretations made on the basis of test scores will be valid" (p. 126).

Reliability

Reliability is an essential quality for language tests and is concerned with the degree to which the test scores are consistent (Brown, 1996; Cohen, 1994; Cohen, as cited in Celce-Murcia, 2001; Dietel, Herman, & Knuth, 1991; Gorsuch, 1997; Rudner, 1994; Shohamy, 1997; Weir, 1995). Cohen (1994) explains 'reliability' as whether a test administered to the same learners a second time would yield the same results. This means that a reliable test should provide consistent scores across different characteristics of the testing situation. For CR tests, especially for achievement tests, it is generally not practical for test users to administer tests a second time to the same population. According to Rudner (1994) and Weir (1995), a test may be considered as sufficiently reliable if its results provides accurate and stable estimate of the ability levels of individuals. The test itself may affect the

learners' performance in the test, and in turn the consistency of the test scores (Shohamy, 1997). However, the reliability of an achievement test and the consistency of the scores may be controlled by test writers in terms of test factors (Brown, 1996). These factors are the extent of sampling the objectives, the degree of ambiguity of the items, the clarity and explicitness of the instructions, the restriction on freedom of response, the quality of layout, the familiarity that the respondents have with the format, and the length of the total tests (Brown, 1996; Cohen, as cited in Celce-Murcia, 2001; Gorsuch, 1997). As Dietel et al. (1991) suggest, if test writers take these factors carefully into consideration, test results will be more accurate estimates of the learners' performance, thus allowing the stakeholders make meaningful decisions.

Authenticity

Authenticity is basically concerned with the degree to which test tasks are relevant to real life language use (Bachman, 1990; Bachman, 1991; Bachman & Palmer, 1996; Brown & Hudson, 2002; Coombe & Hubley, 1999; Halleck & Moder, 1995; Hoekje & Linnel, 1994; Lewkowicz, 2000; McNamara, 2000; Purpura, 1995; Spence-Brown, 2001; Weir, 1995; Wiggins, 1990). If a test and its tasks are closely related to the features of real-life language use, the tests are generally considered to be authentic. The notion of authenticity is best considered in terms of a continuum rather than as an absolute value. Bachman (1991) suggests that the degree of authenticity is relative; there is 'low' or 'high' authenticity, rather than 'authentic' or 'inauthentic'. The property of authenticity should be considered as two types: situational and interactional authenticity (Bachman, 1990).

Bachman (1991) defines 'situational' authenticity as the extent to which "tests and test tasks are relevant to the features of specific target language use" (p.

690). In order to maximize situational authenticity, test writers do not need to sample the actual tasks a domain of non-test language use, but rather try to design tasks that have same critical features as tasks in that domain. However, learners may have different perceptions about the content relevance of the characteristics of the test tasks to their target language use situations, and the learners' perceptions may be different from the test writers (Lewkowicz, 2000). Thus, it is important for test writers to consider these different perspectives and the learners' needs while assessing situational authenticity in tests.

Another type of authenticity is 'interactional' authenticity, in which the interaction between the test tasks and the learners is concerned (Bachman, 1991). Assessing 'interactional' authenticity is more complex than situational authenticity because it requires careful consideration of "the components of the learners' language ability such as areas of language knowledge and metacognitive strategies" (Bachman & Palmer, 1996). The test tasks designed to be interactionally authentic will likely enable the learners use their metacognitive strategies, or higher language abilities and strategies. Halleck and Moder (1995) state that "emphasis on authentic tasks generally has the effect of shifting attention from lower level skills of language competence toward the higher cognitive abilities of textual competence and strategy use" (p. 754). This form of interactional authenticity, as Wiggins (1990) points out, allows learners to demonstrate their language abilities with intellectual tasks which are meaningful for them.

Since authentic tasks (either situational or interactional) allow learners to demonstrate their language abilities and estimate their language performance in non-test situations, this may mean that the results of relatively authentic tests can be generalized to overall proficiency ability (Brown & Hudson, 2002). Weir (1995)

argues that “if test tasks reflect real-life tasks in terms of important identified conditions and operations it is easier to state what a student can do through the medium of target language” (p. 28). Similarly, Bachman and Palmer (1996) claim that “authenticity provides a means for investigating the extent to which score interpretations generalize beyond performance on the test to language use in the target language use domain, or to other similar non-test language use domains” (p. 24).

According to McNamara (2000), though, there are limits to the authenticity of language tests due to the inevitable gap between the test and the criterion. It is not often possible to set out accurate standards or a certain criterion for the tasks in the target language domain. This may threaten the reliability of the scores and the construct validity of the tests. Hoekje and Linnel (1994) argue that it is not acceptable to have authentic but invalid and unreliable tests in educational contexts. Spence-Brown (2001) suggests that “for the achievement assessment where the target may include other things (such as mastery of objectives) and where preparation and practice that would not occur in the real world are expected, full authenticity may not be an appropriate test purpose” (p. 478). However, test tasks designed relatively authentically (perceived as situational or interactional, or both) may have a potential positive effect on test takers’ performance (Bachman & Palmer, 1996; Coombe & Hubley, 1999). When the test is considered to be relevant to real life situations, this may affect positively the test takers’ performance on the test. Test takers may try to do their best in the tests or tasks, which they perceive as relevant to non-test contexts (situational), and challenging for their knowledge and language abilities (interactional).

Practicality

Practicality is another quality which needs consideration in the evaluation of language tests (Bachman & Palmer, 1996; Brown, 2004; Hughes, 1989; Genesee & Upshur, 1996; Purpura, 1995; Weir, 1990). Bachman and Palmer (1996) define practicality as “the relationship between the resources that will be required in the test design, development, and use of the test and the resources that will be available for these activities” (p. 36). The practicality of a test refers to the extent to which the test requirements do not exceed the existing resources in an institution. Weir (1990) suggests that “a valid and reliable test of little use if it does not prove to be a practical one. This involves questions of economy, ease of administration, scoring, and interpretation of results” (p. 34). It is possible to ensure practicality if a test is designed, developed, administered and scored with the available resources of an institution (Hughes, 1989; Purpura, 1995). Bachman and Palmer (1996) categorize these resources as “human resources (test writers, scorers or raters, test administrators), material resources (space, equipment, materials) and time” (p. 36). The practical aspects of a test need to be considered initially in order to avoid any problems during the test design and the test administration process. To Genesee and Upshur (1996), the practical aspects of tests fall into four categories: “cost, administrative time (class time), compilation time (scoring process), administrator qualifications, acceptability (in terms of stakeholders)” (p. 57). Brown (2004) argues that “a test is practical if it is not excessively expensive, stays within appropriate time constraints, is relatively easy to administer and has a specific and time-efficient scoring procedure” (p. 19).

Interactiveness

Interactiveness is another aspect of the usefulness of language tests, and has been defined by Bachman and Palmer (1996) as the extent and type of involvement of the test takers' individual characteristics in accomplishing a test task. According to Bachman and Palmer (1996), these individual characteristics are concerned with "test takers' areas of language ability, metacognitive strategies, topical knowledge, and affective schemata" (p.25). Purpura (1995) mentions interactiveness as the degree of which "a test serves to engage a test taker's language ability, or the task will elicit test performance which replicates a genuine interaction" (p. 5). Purpura further states that in order to engage learners' language ability in a genuine interaction, test writers set tasks in which the learners have to use their knowledge of topic, their language knowledge, their strategic knowledge (metacognitive, cognitive, social, and affective strategies). The property of 'interactiveness' may be achieved by means of relatively 'interactional' authentic test tasks in which the relationship between the learners' language abilities and metacognitive strategies are focused (Bachman, 1991).

Washback

Tests which are used systematically may influence teaching and learning processes positively or negatively in educational settings. This effect of testing is generally known in the field as the 'washback' effect or 'impact' (Alderson et al., 1995; Alderson & Wall, 1993; Bachman & Palmer, 1996; Bailey, 1996; Davies, 1990; Hamp-Lyons, 1997; Hughes, 1989; Shohamy, 2001; Taylor, 2004; Wall & Alderson, 1996). Hamp-Lyons (1997) distinguishes these two terms and claims that 'washback' effect refers to the influence that tests have on the immediate educational context while the 'impact' is a broader term regarding to effects on education and society" (p. 297). Shohamy (2001) and Taylor (2004) defines washback narrowly as

the influence of tests in areas such as curriculum or syllabus design, testing practices, teaching methods, course materials, learning strategies, and knowledge of language to be tested.

Many scholars agree that the washback effect of testing has not been fully explored although it deserves much closer attention (Alderson et al., 1995; Alderson & Wall, 1993; Davies, 1990; Bailey, 1996; Hamp-Lyons, 1997; Wall & Alderson, 1996). Most researchers in language testing suggest that such aspects as test purpose, test method, test content, test organization are the factors which determine the potential washback effect of language tests (Alderson et al., 1995; Bachman, 1990; Bachman & Palmer, 1996; Bailey, 1996; Brown, 1996; Brown, 2000; Brown, 2002; Brown & Hudson, 1998; Cohen, 1994; Davies, 1990; Davies & Pearse, 2000; Genesee & Upshur, 1996; Hamp-Lyons, 1997; Hughes, 1989; McNamara, 2000; Weir, 1995).

The purpose of language tests should serve the goals of an educational program and be clear to the stakeholders in order to promote positive effect on learning and teaching (Bailey, 1996; Davies, 1990; Hughes, 1989). The purpose of a test generally depends on what an educational program intends to teach (Bailey, 1996). If the aim of an educational program is to measure the extent of student achievement of instructional goals, it will be appropriate to use criterion-referenced tests, or specifically achievement tests rather than norm-referenced tests (Davies, 1990; Hughes, 1989). In this context, the scores of norm-referenced tests may not provide stakeholders with the information needed, thus producing the potential for negative washback. This implies that stakeholders should ideally know what kind of information they want to obtain from tests in order to make meaningful decisions

about the learning process. Then it is likely for tests to generate positive washback to stakeholders.

The test method should be related to the course objectives, utilizing a variety of test formats with integrated content in order to generate the potential of positive washback in teaching and learning (Alderson et al., 1995; Bachman, 1991; Brown, 2002; Cohen, 1994; Hughes, 1989; McNamara, 2000). Bachman (1991) remarks that the selection of test method is important in designing language tests. It is important for the language tests used in an educational setting to be closely linked to the goals and objectives of that language program. Hughes (1989) argues that the test method or techniques need to be closely related to the objectives in the syllabus or curriculum. For example, “if we want to assess the skill of writing, using only multiple-choice items would be inappropriate and undesirable” (Hughes, 1989, p. 1). This situation may also threaten face validity of that test leading the learners to think that the test is not appropriate (Cohen, 1994). The learners may not take the test seriously, thus affecting their performance on the given test.

Furthermore, a variety of test formats should be used in order to measure learners’ language abilities as long as these formats are not at variance with the course objectives. Brown (2002) mentions that using a variety of examination formats, including written, oral, aural, and performance-based formats yields positive washback to stakeholders. Such test formats enable learners to demonstrate their language abilities in all four language skills. Thus, test users and teachers may obtain a large sample of the learners’ language abilities and knowledge of language. McNamara (2000) suggests that variation in test formats in terms of integrated content and skills may have more positive washback than discrete item test formats. In this way, learners’ abilities are not confined in exclusively grammatical and

structural language use. Hughes (1989) suggests that tests should sample wide range of test methods or items so that learners can demonstrate their language abilities in various contexts. From different point of view, Alderson et al. (1995) recommend for test writers that "... to promote positive washback test developers vary the test methods so that no one method predominates and becomes predictable" (p. 44). They claim that learners may increase their repertoire of test-taking strategies preparing for unpredictable and various test tasks or items.

Test content also needs to be based on goals and beliefs of the educational program in which it will be used and to include authentic tasks to promote positive washback (Bachman and Palmer, 1996; Bailey, 1996; Brown & Hudson, 1998; Brown, 2002; Cohen, 1994; Weir, 1995). As Brown (2002) puts it, "washback becomes negative washback when there is a mismatch between ...the content (e.g., the material, abilities being taught) and the test" (p. 15). Bachman and Palmer (1996) suggest one way to minimize the potential for the negative impact, which is to design test content which correspond closely to objectives of the curriculum and instruction. In this way, tests may promote beneficial 'washback' to learners, teachers and administrators, indicating accurately whether or how much of these objectives have been achieved by the learners. For example, "if a program set a series of communicative performance objectives, but assesses the students with multiple-choice structure tests, a negative washback effect will probably begin to work..." (Brown & Hudson, 1998, p. 668). Test users or teachers may not know whether the learners use the language communicatively in social contexts. This may prevent the stakeholders from making appropriate decisions about the learning and curriculum. Cohen (1994) supports that it is important for assessment instruments to affect educational practices and beliefs of that program positively. The content of test

should coincide with educational beliefs of a language program in order to generate positive washback.

According to Brown (2002) “washback is an important aspect of the testing and curriculum endeavor” (p. 14). He argues that:

“washback, whether it is positive or negative, can be a potential boon or threat to language teaching curriculum...because, through washback, a test can steer a curriculum in one direction or another (in terms of teaching, course content, course characteristics, and, or class time) either with or against the better judgment of the administrators, teachers, students” (p.13).

If a test yields a positive effect on stakeholders and a program, this will result in appropriate changes or adaptations in the curriculum, thus contributing to curriculum development. Brown and Hudson (1998) state that “if assessment procedures in a curriculum do not correspond to a curriculum’s goals and objectives, the tests are likely to create a negative ‘washback’ effect on those objectives and on the curriculum as a whole” (p. 667). Another suggestion to promote beneficial washback is that test content should utilize authentic tasks and texts (Bailey, 1996; Weir, 1995). According to Weir (1995), “the more features of real life use of language that can be built into tests, the more positive the washback effect and the easier it will be to make statements about what students can or can not do in the language” (p. 6).

Test organization procedures, to promote positive washback, concern the involvement of teachers and administrators in the testing procedures, and the professional competence of test developers (Bachman & Palmer, 1996; Davies & Pearse, 2000; Genesee & Upshur, 1996; Lynch & Davidson, 1994). As Genesee and Upshur (1996) state, “...teachers are actively continuously involved in second language evaluation...Even when the teachers are not the actual decision makers, they are affected “ (p.3). It is important not to exclude teachers from the testing procedures because they may determine whether the tests match with learning and

teaching activities. According to Lynch and Davidson (1994), “mismatch can often go unnoticed unless teachers are included in the test development process” (p. 737). When there is mismatch between the tests and teaching practices, this may result in negative washback on teaching. Teachers may not rely on the tests as accurate measure of the progress their learners have made in learning. Davies and Pearse (2000) mention that teachers need to make comments on the tests and suggest changes or modifications if necessary. The suggestions made by teachers may lead test developers to make appropriate changes considering the content relevance of tests to teaching. In addition, it may be possible for administrators to promote positive washback on teaching by encouraging the professional competence of the testers (Alderson et al., 1995; Bailey, 1996). When the test writers or test developers are well qualified on testing or assessment, test users may be able to rely on the tests being prepared.

Studies Conducted on Achievement Tests

Studies have been conducted on such aspects of achievement tests as the validity of these tests, and stakeholders’ perceptions of achievement tests (Aksan, 2001; Dalyan, 1990; Ketevu, 2001; Kuntasal, 2001; Serpil, 2000).

In his study, Aksan (2001) investigated the teachers’ perceptions of whether the content of language tests reflect the content of course book and their teaching practices at Niğde University. Another equally important purpose of the study was to find out whether the teachers follow the course book content in their teaching. The teachers’ perceptions were elicited with a questionnaire, which consisted of 40 test items chosen randomly among five teacher-made final tests administered at Niğde University in 2000-2001 academic year. The results of the study showed that the teachers agreed on that the tests were sampling the teaching practices. The findings

also showed that the teachers were following the course book content in their teaching. This study was a local study and the results of the study provided the stakeholders with empirical evidence for the content validity of teacher-made final tests.

Serpil (2000) also explored the teachers' perceptions of the content validity of achievement tests at Anadolu University School of Foreign Languages. He distributed questionnaires to the teachers to find out the teachers' perceptions of whether the tests represent the content of intermediate course material and teaching objectives. These objectives were identified through interviews which were carried out with the teachers. Then, the content of the tests were compared with the course materials and teaching objectives. The findings showed low agreement between the teaching objectives and the content of the tests, and low correlation between the course material and the tests, especially on exercise types, except for one achievement test. Serpil concludes that this conflict among the results may result from the lack of clear objectives and the lack of test specification in the test preparation process.

Ketevu (2001) explored the validity of achievement and proficiency tests at Bilkent University School of English Language (BUSEL). In her study, she investigated the concurrent and predictive validity of the tests used at BUSEL. The aim was to find out whether the tests serve their intended purposes and whether test writers prepare the tests containing actually what they aim to test. The results of Ketevu's study showed that the achievement tests have satisfactory level of predictive validity.

Dalyan (1990) investigated teachers' perceptions of preparing achievement tests at Anadolu University Institute of Social Sciences, and analyzed a sample

achievement test used in the institution. Dalyan points out that the tests used during the instructional process need to guide and direct student learning and monitor progress toward course objectives. He further suggests that the success of any formal language program is crucially linked with its testing philosophy and practice, and that teachers need training on testing. The results of Dalyan's study revealed that teachers need training on assessment and testing, especially on achievement tests.

Kuntasal (2001) investigated the testers' and teachers' perceptions of achievement tests as used at Middle East Technical University Department of Basic English. In her study, the testers and teachers were given questionnaires to provide their opinions about the testing practices in the institution. The results of the study showed that both teachers and testers were satisfied with the testing practices, the relationship between the testing and teaching, and the design of the tests used at Middle East Technical University Department of Basic English. However, the findings suggested that the testers were not totally sure about the cooperation with the teachers. Kuntasal's study was similar to this study and research questions posed for this study were inspired by those of her study.

Conclusion

First, this chapter reviewed the literature on the roles of achievement tests in educational settings. Next, test development process including the construction of test specifications, and item writing process was discussed. Then, the evaluation of language tests in terms of such properties as validity, reliability, authenticity, practicality, interactiveness, and washback was discussed. Finally, the studies conducted on achievement tests were presented.

CHAPTER III: METHODOLOGY

Introduction

This study investigated testers' perceptions of the test development process at Muğla University School of Foreign Languages and explored teachers' and testers' attitudes towards the achievement tests used at Muğla University School of Foreign Languages.

This study addressed to the following research questions:

1. What are the testers' perceptions of the test development process at Muğla University School of Foreign Languages?
2. What are the teachers' attitudes towards the resulting achievement tests used Muğla University School of Foreign Languages?
3. What are the testers' attitudes towards the resulting achievement tests used at Muğla University School of Foreign Languages?
4. Are there any discrepancies between the teachers' and testers' attitudes towards the achievement tests used at Muğla University School of Foreign Languages?

This chapter will present information about setting, participants, instruments, data collection procedures and the data analysis.

Setting

The study was conducted at Muğla University School of Foreign Languages. The students who attend Muğla University School of Foreign Languages have to take

a placement test prepared by the testing unit at the beginning of each academic year. The students are placed at appropriate levels of the program according to their test results. There are three levels at Muğla University School of Foreign Languages: C level (zero-beginner), C+ level (false-beginner), and B level (intermediate). The courses offered at Muğla University School of Foreign Languages are not skill based. The English instructors teach the course book ‘Opportunities’ as a main course which integrates all four language skills. The students do not take any level tests in order to move another level during the academic year. The students are assessed by means of progress achievement tests and final achievement tests and on their class performance throughout the year. Table 1 below shows weighting of the criteria for assessing students at Muğla University School of Foreign Languages:

Table 1

Weighting of the Students’ Assessment Criteria

Midterm I	5 %
Midterm II-Midterm III	10% - 10%
Midterm IV	15%
5 Quizzes	15%
Teacher Assessment	15%
Final Exam	30%
Total	100%

The midterms are given every two months during the year, and the final proficiency exam is administered two weeks after the fourth midterm at the end of the year. Instructors teaching the same class give their students a mark every month. This mark is based on the teachers’ evaluation of a student’s class participation, class work and assignments. The overall weighting of the teacher assessment grade in the assessment system is 15%.

Participants

There are 40 English teachers who are currently working at Muğla University School of Foreign Languages, five of whom also work in the test office. Two teachers refused to participate in this study and three others were unavailable on the days the survey was carried out. The questionnaires were distributed to 30 English teachers and five testers.

Table 2 below shows background information about the teachers and Table 3 shows background information about the testers.

Table 2

Background Information about Teachers

Degree Programs	BA	MA	PhD	Other	
Frequency	27	3	0	0	
Percentage	90 %	10 %	0 %	0 %	
Total years of teaching experience	1-4	5-8	9-12	13-16	+ 16
Frequency	5	10	5	6	4
Percentage	16.7 %	33.3 %	16.7 %	20.0 %	13.3 %
Teaching experience at Muğla University	1-4	5-8	9-12	13-16	+ 16
Frequency	8	15	7	0	0
Percentage	26.7 %	50.0 %	23.3 %	0 %	0 %
Taken courses related to ‘testing’			Yes	No	
Frequency			15	15	
Percentage			50 %	50 %	
Testing experience at Muğla University			Yes	No	
Frequency			16	14	
Percentage			53.3 %	46.7 %	
Testing experience at another institution			Yes	No	
Frequency			0	30	
Percentage			0 %	100 %	

Table 3

Background Information about Testers

Degree Programs	BA	MA	PhD	Other
Number of testers	5	0	0	0
Total years of teaching experience	1-4	5-8	9-12	13-16 + 16
Number of testers	1	3	1	0
Teaching experience at Muğla University	1-4	5-8	9-12	13-16 + 16
Number of testers	3	2	0	0
Taken courses related to 'testing'			Yes	No
Number of testers			3	2
Years of testing experience at Muğla University	1 year	2 years	3 years	3+
Number of testers	3	1	1	0
Testing experience at another institution			Yes	No
Number of testers			1	4

The number of the testers who participated in the study was 5. All the testers involved in this research are female. The testers are responsible for preparing the midterm exams, the quizzes, and the final proficiency exam for any level of the students at Muğla University School of Foreign Languages. The testers also teach at different levels. Two of the testers teach zero-beginners, other two teach false beginners, and one of them teaches intermediate-level students. As can be seen Table 3, three of the testers have been working in the testing office for a year, one of the testers has been working there for two years, and the coordinator of the test office has been working there for three years.

Instruments

The instruments employed in this study were both questionnaires and interview schedules.

Questionnaires

Two parallel questionnaires were employed in this study. Oppenheim (1992) indicates that questionnaires are often used since little time is required for their implementation, and they are easy to process. In addition, questionnaire technique is an efficient and practical way of gathering data from large population of participants. Bachman and Palmer (1996) suggest a model of test usefulness in order to help the stakeholders evaluate the quality of tests. This test usefulness model is concerned with properties of “validity, reliability, authenticity, practicality, interactiveness and impact, washback” (p. 17). These properties provided a basis for the preparation of the questionnaires in this study. The teachers’ questionnaire consisted of three sections and 45 items (see Appendix A). The testers’ questionnaire consisted of four sections and 58 items (see Appendix B).

Teachers’ Questionnaire. The teachers’ questionnaire consisted of three sections which were concerned with background information, teachers’ attitudes towards the tests and further comments on the tests and overall testing practices at Muğla University School of Foreign Languages. Table 4 presents the distribution of the questions on the teachers’ questionnaires and the focus of each section.

Table 4

Distribution of Questions on the Teachers’ Questionnaire

<u>Section no</u> Focus	<u>Section I-</u> Background information	<u>Section II-</u> Teachers’ attitudes towards the tests	<u>Section III-</u> Further comments on the tests and overall testing practices
Number of Questions	7	34	3

Section I involved 7 questions which solicited background information about the teachers: their gender, degrees completed, years of teaching experience, years of teaching experience at Muğla University School of Foreign Languages, whether they have worked in the test office at Muğla University School of Foreign Languages or elsewhere and whether they have taken any courses on testing, measurement and evaluation or assessment.

Section II of the teachers' questionnaire basically referred to research question 2, which is concerned with the teachers' attitudes towards the tests used at Muğla University School of Foreign Languages. This section was comprised of 34 Likert scale items in which the participants were asked to mark the alternative that best corresponds to their opinion. The alternatives were 'Strongly Disagree' (SD), 'Disagree' (D), 'Uncertain' (U), 'Agree' (A) and 'Strongly Agree' (SA).

Section III was designed to obtain the teachers' further comments on the testing system at Muğla University School of Foreign Languages. This section included three open-ended questions which allow the teachers to provide their comments on the midterms, the quizzes, and overall testing practices.

Testers' Questionnaire. The testers' version of the questionnaire was made up of four sections which were concerned with background information, their attitudes towards the tests used at Muğla University School of Foreign Languages, their perceptions of the test development process, and their further comments on overall testing practices at Muğla University School of Foreign Languages. Table 5 shows the distribution of the items on the testers' questionnaire and the main focus of each section.

Table 5

Distribution of Questions on the Testers' Questionnaire

<u>Section no</u> Focus	<u>Section I-</u> Background information	<u>Section II-</u> Testers' attitudes towards the tests	<u>Section III-</u> Testers' perceptions of the test development process	<u>Section IV-</u> Further comments on the tests and overall testing system
Number of Questions	7	30	18	3

In Section I, the participants were asked 7 questions in order to provide their background information: their gender, degrees completed, years of teaching experience, years of teaching experience at Muğla University School of Foreign Languages, how long they have been working at the test office, whether they have worked in the test office of any other institution and whether they have taken any courses on testing, measurement and evaluation or assessment.

Section II referred to research question 1, and contained 30 questions, which sought the testers' perceptions of the tests used at Muğla University School of Foreign Languages. In Section III, there were 18 questions which were concerned with the testers' perceptions of the development process of the tests used at Muğla University School of Foreign Languages. Both the second and third section consisted of Likert scale items in which the testers were asked to mark the alternative that best corresponds closely to their opinion. The alternatives were 'Strongly Disagree' (SD), 'Disagree' (D), 'Uncertain' (U), 'Agree' (A) and 'Strongly Agree' (SA).

Section IV was designed in order to elicit the testers' further comments on the testing system and the present testing practices at Muğla University School of Foreign Languages. This section was organized in the same way as Section III of the teachers' questionnaire with three open-ended questions which allowed the testers to

provide their further comments on the midterms, the quizzes, and overall testing practices at Muğla University School of Foreign Languages. The results of the questionnaires were the basis for the interview schedule to be prepared.

Interviews

Semi-structured interviews were employed in this study. These “are guided by a list of questions or issues to be explored, but neither the exact wording nor the order of the questions is determined ahead of time” (Merriam, 1998, p. 74). This approach enables the researcher to ask follow-up questions according to the responses received from the participants. This helps the researcher gain in-depth insights about the participants’ perceptions. The researcher carried out these interviews in this research with five randomly selected teachers and the five testers. The interview questions were designed after the data analysis of the questionnaires, and these questions were open-ended questions. Table 6 shows the issues covered in the interviews which were carried out with the teachers and the testers.

Table 6

Issues Covered in the Interviews

Interviews with Teachers	Interviews with Testers
Satisfaction with the tests	Cooperation with teachers
Content validity of the tests	Referring the tests to objectives
Involvement of the teachers in the testing procedures	Involvement of the teachers in the testing procedures
Appropriateness of tests to students’ level	Appropriateness of tests to students’ level
Informing the students about test content	Informing the students about test content
Teacher-made midterms and quizzes	Selection of items and tasks in the tests
Authenticity of test tasks	Authenticity of test tasks
Quality of test items	Quality of the test items
The lack of oral assessment	The lack of oral assessment
Effect of the tests	Difficulties of working in the test office
Additional Comments	Satisfaction with current testing practices
	Additional Comments

Separate interview schedules were drawn up for the teachers (see Appendix C), and the testers (see Appendix D). All the participants were given a consent form to sign prior to the interview sessions (see Appendix E). The interviewees were asked how and why questions addressing the results from different sections of the questionnaires in order to gather more detailed information tests used at Muğla University School of Foreign Languages. The participants' responses to the interview questions were tape-recorded and transcribed.

Data Collection Procedures

The questionnaires were designed based on the review of literature. The questionnaires were piloted at Hacettepe University School of Foreign Languages on 19 March 2004. The aim of the piloting was to check for any problems with the items of the questionnaires and any difficulties in understanding the items. 10 teachers and 5 testers who currently work at Hacettepe University School of Foreign Languages participated in the pilot study. For the pilot study, the researcher included a separate section in the questionnaire in order to allow the participants to make comments on whether there were any unclear or ambiguous items in the questionnaires. The participants reported some difficulties in understanding some of the items in the questionnaires and possible problems which would arise in the actual data collection. After making necessary changes in the questionnaires according to the results of piloting, the researcher administered the questionnaires at Muğla University School of Foreign Languages on 30 March. The completed questionnaires were collected over the following three days. After the initial data analysis of the questionnaires, the researcher prepared the schedules for interviews.

The pilot interviews were carried out with 3 current students of the MA TEFL Program. The researcher explored whether there should be any changes in the final

versions of the interview schedule. The piloting of interviews was completed on 22 April 2004.

After making a few changes and corrections in the interview schedule, the researcher carried out the interviews at Muğla University School of Foreign Languages on 27-28-29 April 2004. The researcher carried out the interviews with five randomly chosen teachers and the five testers over three days. The interviews were tape-recorded and transcribed.

Data Analysis

First, the researcher analyzed the quantitative data gained from Likert scale questions in Section II of the teachers' questionnaire, and Section II and III of the testers' questionnaire using SPSS 9.05. For the Likert scale items, the weighting was as it follows, Strongly Disagree (SD)= 1, Disagree (D)= 2, Uncertain (U)= 3, Agree (A)= 4 and Strongly Agree (SA)= 5. Frequencies for items were calculated in order to have a general view about the participants' perceptions of the tests used at Muğla University School of Foreign Languages. One-way chi square values were calculated only for the items in the teachers' questionnaires because the number of the testers was 5, which is too small for analysis using chi square tests.

Second, the researcher analyzed the qualitative data gathered from three open-ended questions in Section III of the teachers' questionnaire and Section IV of the testers' questionnaire (see Appendix H). Participants' responses were entered in a table, and key terms and key words were identified. Then, participants' responses were analyzed based on the interpretation of these patterns with reference to research questions.

Finally, in order to do data analysis for the responses given to the interview questions, the researcher used qualitative methods. After the interviews were

transcribed, the responses of each participant were transcribed in a matrix. Each response which referred to the data from the questionnaires and to the research questions was highlighted (see Appendix G).

Conclusion

In this chapter, the educational setting in which the researcher conducted this study was explained. Then, the participants were presented who were involved in this survey study. The instruments in the study, and the data collection procedures were presented. At the end of the chapter, the data analysis of the study was briefly explained.

In the next chapter, the data gained from the questionnaires and the interviews will be analyzed and presented in three sections. In the first section, the data will present the data gained from the Likert scale questions in Section III of the testers' questionnaire, and the interview transcripts. In the second section, the data gained from the Likert scale questions in Section II of both the teachers' and testers' questionnaire, and the interviews will be presented. In the last section, the data gained from three open-ended questions in Section III of the teachers' questionnaire and Section IV of the teachers' questionnaire will be analyzed qualitatively.

CHAPTER IV: DATA ANALYSIS

Introduction

This study was conducted in order to investigate the testers' perceptions of the test development process at Muğla University School of Foreign Languages, and the testers' and the teachers' attitudes towards the resulting achievement tests. The research questions posed for the study were as follows:

1. What are the testers' perceptions of the test development process at Muğla University School of Foreign Languages?
2. What are the teachers' attitudes towards the resulting achievement tests used at Muğla University School of Foreign Languages?
3. What are the testers' attitudes towards the resulting achievement tests used at Muğla University School of Foreign Languages?
4. Are there any discrepancies between the teachers' and testers' attitudes towards the achievement tests used at Muğla University School of Foreign Languages?

Data Analysis Procedures

In this study, the researcher used both quantitative and qualitative data analysis procedures. Three sets of data were used in data analysis procedures. The first set of data, which was gathered from Likert scale questions in the two questionnaires, was analyzed quantitatively. The second set of data, which was collected through three open-ended questions in two questionnaires, was analyzed

qualitatively. The third set of data, which is composed of the interview data, was analyzed qualitatively.

The first phase of the data analysis procedure was the analysis of the Likert scale questions in two questionnaires. The data from the 34 Likert scale questions in Section II of the teachers' questionnaire and the 48 Likert scale questions in Sections II and III of the testers' questionnaire was entered into SPSS 9.05. The frequencies for each item in the two questionnaires were calculated. One-way chi square values were calculated only for the items in the teachers' questionnaire because the number of the testers included in the study was too small to permit using a chi square test.

The second phase of the procedure was to analyze the data gathered from the three open-ended questions in Section III of the teachers' questionnaire and Section IV of the testers' questionnaire. Participants' responses were entered in a table, and key terms were identified. Then, participants' responses were analyzed based on the interpretation of these patterns with reference to the research questions.

The last phase in the data analysis procedure was the interpretation of the interview data. After the interviews were transcribed, the responses of each participant were entered in a matrix. Each response which referred to the data from the questionnaires and to the research questions was highlighted, and codified.

This chapter is made up of three sections which are organized with reference to four research questions. In the first section, an analysis of the questions in Section III of the testers' questionnaire provides evidence of the testers' perceptions of the test development process at Muğla University School of Foreign Languages, which will be used to answer research question 1. In the second section, an analysis of the questions in Section II of the two questionnaires, which are concerned with the teachers' and testers' attitudes towards the achievement tests, is presented and refers

to research questions 2, 3, and 4. Finally, the third section presents the qualitative data gathered from the last section of the questionnaires, which also to refer to research questions 2, 3, and 4. In each section, interview data is presented with the quantitative data gathered from the questionnaires in order to clarify the participants' responses to the questionnaires.

Testers' Perceptions of the Test Development Process

The questions in Section III of the testers' questionnaire sought to answer research question 1 which was concerned with the testers' perceptions of the test development process at Muğla University School of Foreign Languages. This section presents the data from Section III of the testers' questionnaire under 3 sub-headings: cooperation during the test development process, testers' perceptions of the test preparation process, and testers' perceptions of the item-writing process.

Cooperation during the Test Development Process

Questions 1, 2, 10 and 11 in Section III of the testers' questionnaire were concerned with the testers' perceptions of the cooperation among themselves and with the curriculum unit. Table 7 below shows the items related to the cooperation throughout the test development process.

Table 7

Items Related to Cooperation in the Test Development Process

Questions	SD	D	U	A	SA
Q1. There is active cooperation between the testing office and the curriculum unit in the test development process.	0	0	0	4	1
Q2. All the testers work together while preparing the tests.	0	0	1	2	2
Q10. The testers determine together which language skills will be measured in each test according to the course objectives.	0	0	0	3	2
Q11. The testers determine together the item types according to the purpose of the tests.	0	0	1	1	3

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree.

As can be seen in Table 7, for question 10, all five testers agreed that they determine together which language skills will be measured in the tests. For questions 1 and 2, there is a general agreement among the testers on cooperation with the curriculum unit and working cooperatively among themselves during the test preparation process. However, the data gained from the interviews showed no strong evidence of active cooperation among the testers and with the curriculum unit during the test development process.

When the testers were asked about how they design the tests during the interviews, two participants (P7 and P8) indicated that they work separately and each tester prepares different parts of the tests:

(Participant 7) ...**we assign ourselves to prepare different parts.** For example, one prepares vocabulary and the other does grammar.

(Participant 8) ...We share the parts of an exam. One prepares one part of the exam and the other prepares another part. We meet and work together. First, **we work separately**, then meet and make comments.

The cooperation between the testers and curriculum unit was not mentioned by four testers during the interviews. During the interviews, one tester (P6) perceived this cooperation as distribution of the syllabus, which may not be assumed to reflect active and continuous cooperation between the two units:

(Participant 6) ...with the curriculum [unit], yes. The curriculum unit informs us what should be tested in the tests by **distributing the syllabus.**

Testers' Perceptions of the Test Preparation Process

The following discussion is concerned with the testers' perceptions of the test preparation process at Muğla University School of Foreign Languages. Table 8 below shows the frequency of the values for the items of Section III of the testers' questionnaire which are related to the test preparation process. Questions 2 and 3

were concerned with the testers’ perceptions of the planning of the midterms and quizzes. Questions 5, 6, 7, 8, 9, 14,15, and 17 were concerned with the testers’ perceptions of more specific aspects of the test preparation process such as having a specific written plan prior to the test preparation process, referring to the course objectives, specification of the purpose for each test, consideration of students’ proficiency level, encouraging integrative language use, determining the scoring criteria, and use of existing resources in the institution.

Table 8

Items Related to the Test Preparation Process

Questions	SD	D	U	A	SA
Q3. I think the midterms are carefully planned.	0	0	0	2	3
Q4. I think the quizzes are carefully planned.	0	0	0	2	3
Q5. The testers have a specific written plan about what the test will be like prior to designing the midterms.	0	2	1	2	0
Q6. The testers have a specific written plan about what the test will be like prior to designing the quizzes.	0	2	1	2	0
Q7. The testers design the tests with reference to explicit course objectives.	0	1	2	2	0
Q8. The testers specify the purpose of each test prior to the test development process.	0	0	1	2	2
Q9. The testers take the students’ proficiency level into consideration while designing the tests.	0	0	0	3	2
Q14. The tests are designed in a way that encourages the students to use their language abilities integratively.	0	0	1	3	1
Q15. The testers determine the scoring criteria according to the priority of the language skills to be measured.	0	0	0	2	3
Q17. The existing resources of the institution correspond to the requirements of the tests.	0	0	1	2	2

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree.

The results in Table 8 show that all five testers unanimously agree that the midterms and quizzes which are used at Muğla University School of Foreign Languages are carefully planned. Although this notion is supported by all of the testers, the analysis of more specific questions about the test preparation process reveals that there are disagreements among the testers about certain aspects of the test preparation process. These aspects are namely the use of a written plan prior to the test preparation process and reference to course objectives in the test

development process. The disagreements among the testers on these aspects were clarified in the interview data.

For questions 5 and 6, as displayed in Table 8, two testers indicated that they make use of a specific written plan before preparing the midterms and quizzes whereas two of testers disagreed about having a specific written plan for the tests prior to the preparation process, and one tester ticked the option uncertain. During the interviews, the testers were asked about whether they design the midterms and quizzes according to a specific written plan. Participants' responses explain why the testers were different from each other in their responses to questions 5 and 6. Two testers (P9 and P10) indicated that they generate a specific written plan before they prepare each test. P9 and P10 further stated that this plan includes percentages of language skills, an answer key and the scoring criteria. One of the testers (P6) defined this plan as a schedule written by the testing unit coordinator:

(Participant 9) **Yes, we have a specific plan.** It includes **percentages of the language skills**. It changes for every test. We decide what we expect from the students. We use other course books, the items we used in previous years. We have **answer key** and **scoring criteria** beforehand.

(Participant 10) **Yes, we have a kind of plan.** We declare the **percentages or prior importance of the skills**.

(Participant 6) We prepare the tests according to **a kind of schedule written by our coordinator**.

Two of the participants (P7 and P8) disagreed about having a written plan, and they also indicated that they work separately first on different parts of a test or on a specific language skill to be measured in the test such as grammar, vocabulary, reading:

(Participant 7) **Not a written plan**, we assign ourselves to prepare different parts. For example, one prepares vocabulary and the other does grammar. We have such kind of schedule.

(Participant 8) **No, we do not have such a written plan...**

As can be seen from participants' responses above, three of the testers said in different words that they have a plan or a schedule, written or not, before the midterms and quizzes are being prepared. However, what they pointed out is not a fully written test specification; they mentioned only three parts of a test specification, such as the percentages of language skills, an answer key, and scoring criteria.

For question 7, two testers agreed, two were uncertain and one disagreed about operating the test preparation process with reference to explicit course objectives. In question 7, four of the testers may have interpreted 'course objectives' as 'course book objectives', because, during the interviews, all of the testers stated that they refer to the course book objectives while preparing the tests. One of the testers (P10) mentioned that the course book objectives determine the test content. P9 stated that curriculum requirements are referred to in addition to the course book objectives; however, it is not clear what these requirements are:

(Participant 10) **We use course book objectives** and the content of the units in the course book. We are dependent on the book, sometimes we can't decide. **The course book decides instead of us,** as teachers.

(Participant 6) Usually [**we use**] **the course book objectives**, ... as both teachers and testers, we can figure out what the students should be able to do at the end of each unit, for example.

(Participant 9) What we want our students to be able to do as teachers determines what will be asked in the tests. **We make use of the curriculum requirements and the course book objectives.**

During the interviews, one teacher (P1) and one tester (P7) both suggested that there are no curriculum objectives. P7 indicated that the use of objectives given in the course book is the consequence of lacking explicit objectives in the curriculum and the syllabus.

(Participant 1) ... I think we are only measuring whether the students have achieved **the course book objectives** because **our objectives are not clear in the curriculum**.

(Participant 7) **We prepare the tests according to course book objectives. Because we do not have specific or explicit objectives in the syllabus or curriculum.**

The interview transcripts above reveal that the testers agree on the use of course book objectives, not course objectives, as a reference while preparing the tests. The interview data also supported that the use of course book objectives results from the lack of explicitly specified objectives in the curriculum.

Testers’ Perceptions of the Item Writing Process

In Section III of the testers’ questionnaire, questions 12, 13, and 16 were concerned with such aspects of the item-writing process as including items from the course book, including previously used items, and proofreading of the test items.

Table 9 below displays the results for questions 12, 13, and 16.

Table 9

Items Related to the Item Writing Process

Questions	SD	D	U	A	SA
Q12. The tests include the items which are taken directly from the course book.	3	1	0	0	1
Q13. The testers make use of the items which have been used in previous tests.	1	2	1	0	1
Q16. The test items are proofread by another colleague who has not written them.	0	0	0	0	5

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree.

As is seen in Table 9, for question 16, all the testers confirmed that the test items are checked or proofread before test administration process by an outsider colleague who has not written the items. In the absence of pre-testing, having an outsider check the test items and overall test increases the reliability of the test by reducing the chance of mistakes and errors (Weir, 1995).

For questions 12, and 13, only one tester agreed that items which are taken directly from the course book and from the previous tests are used in constructing the tests. The interview data suggested that instead of taking the test tasks directly from the course book, the testers generally develop new items for the tests by making use of other resources. During the interviews, four participants agreed on the adaptation and development of the test items:

(Participant 7) ...**Sometimes we create or adapt from the other course books according to the level and needs of the students.**

(Participant 9) **We use other course books, the items we used in previous years...** We select the test tasks according to what we teach in the class and the students' needs. **We develop and create the items, use other resources...**

Among the five testers, only one of them (P6) said that the items are taken from the course book, but she further clarified that these items are not presented as they are in the course book.

(Participant 6) We take sometimes **from our course book**, workbook, **but make changes**, and generally we develop the items ourselves.

Teachers' and Testers' Attitudes towards the Achievement Tests

In this section, the results for the Likert scale items in Section II of the teachers' questionnaire and the testers' questionnaire are presented with relevant interview data. The frequencies for the Likert scale items are given in the tables. However, as was mentioned earlier, chi square values are displayed only for the items in the teachers' questionnaire because the small population of the testers did not allow similar analysis of their data.

The data is presented in this section under nine sub-headings: teachers' and testers' general opinions about the tests, teachers' opinions on preparing the tests for their classes, cooperation between the teachers and testers, teachers' and testers' perceptions of the content of the tests, of the test tasks, of the test items and

instructions, of the relationship between learning and the tests, of the effect of tests on teaching, and of working in the test office.

Teachers’ and Testers’ General Opinions about the Tests

The first discussion focuses on the teachers’ and testers’ general opinions about the tests used at Muğla University School of Foreign Languages. Table 10 below displays the results for questions 2 and 3 in Section II of two questionnaires which are related to the satisfaction of the teachers and testers with the midterms and quizzes.

Table 10

Teachers’ and Testers’ General Opinions about the Midterms and Quizzes

Questions	Position	SD	D	U	A	SA	X2
Q2. I am satisfied with the midterms.	Teachers	0	1	7	19	3	40.00**
	Testers	0	1	0	3	1	--
Q3. I am satisfied with the quizzes.	Teachers	2	2	6	18	2	32.00**
	Testers	0	1	0	3	1	--

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square

**p ≤ .01

Table 10 shows that the chi square calculations for questions 2 and 3 are significant at the level of $p \leq .01$. The responses given to questions 2 and 3 reveal that the majority of teachers (22) are generally satisfied with the midterms, and 20 teachers are satisfied with the quizzes.

The results for questions 2 and 3 shows that out of five testers, only one seems to be not satisfied with the tests they prepare whereas the other four testers reported their satisfaction with the resulting tests. Similarly, when the testers were asked about whether they are satisfied with the tests during the interviews, only one tester (P6) stated that she is not satisfied with the overall testing practices in terms of the reliability of the exams, and the qualifications of the testers:

(Participant 6) We need **more qualified testers**. It would be better then. The **reliability of the exams** can be questioned. And in-service training may be provided.

The interviews also provided the teachers with opportunity to clarify the responses they gave to questions 2 and 3 as well. Out of five teachers, two of the participants were satisfied with the tests. However, even the participants (P3 and P5) who are satisfied with the tests identified some problems, such as the lack of an oral exam and the instructions given in the tests:

(Participant 3) **Yes, I am satisfied**. But we are **not testing speaking** because of the crowded population of the students.

(Participant 5) **Yes, quite well. Sometimes there are problems with instructions** or rubrics. **Unclear and not appropriate for their levels** I think.

Those (P4 and P1) who are not satisfied with the tests mentioned additional weaknesses, such as focusing exclusively on grammar, not using authentic test tasks, and adopting a traditional approach for designing tests:

(Participant 4) **Not totally. We ignore how the students express themselves orally. We generally focus on grammar.**

(Participant 1) **I am not totally satisfied**, there are some problems but it is hard to define them exactly. **Our tests seem to me old-fashioned**, I mean **traditional**. For example, **reading tasks should be more authentic or updated. The tests still focus on grammar, not language use.**

Teachers' Opinions on Preparing the Tests for Their Classes

Questions 31 and 32 in Section II of the teachers' questionnaire were concerned with whether the teachers would like to prepare the tests individually for their classes. These two questions were asked only to teachers. Table 11 presents the results for questions 31 and 32.

Table 11

Teachers’ Opinions on Preparing the Tests for Their Classes

Questions	SD	D	U	A	SA	X2
Q31. I would like to prepare the midterms for my classes in cooperation with my partners.	8	9	2	7	3	6.69
Q32. I would like to prepare the quizzes for my classes in cooperation with my partners.	2	5	6	13	4	11.66*

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree;
X2= Chi Square
*p ≤ .05

For question 31, 17 teachers disagree about wanting to prepare the midterms for their own classes. In contrast, 17 teachers, in question 32, indicated that they would like to prepare the quizzes for their own classes in cooperation with their partners. The chi square calculations are different for these two questions with the significant result for question 32 indicating a less equal distribution of responses across the five opinions.

The interview data supported the analysis that the teachers do not necessarily feel like preparing the midterms for their own classes. During the interviews, when the teachers were asked whether they would like to prepare the midterms and quizzes for their own classes, the participants stated both disadvantages and advantages of teacher-made midterms. The lack of standards between classes, inequality of assessment, impracticality were mentioned as disadvantages whereas teachers’ involvement in the assessment process, awareness of students’ level, and the students’ paying more attention to instruction were mentioned as the advantages of teacher-made midterms:

(Participant 1) **Yes**, [I would like to prepare the midterms] but it has disadvantages and advantages. If teachers prepare their own tests then there **won’t be any standard between the classes. It is useful because my students will focus on how I teach in the class** they will know that they will be tested in that way.

(Participant 2) **Of course**, [I would like to]. I teach it and **I want to measure it by myself**. Why don't we trust the teachers? **The teachers know their students' level the best**.

(Participant 3) **No, I do not want it** because I can take the side of my students **It would be unequal and not practical to prepare our own midterms...**

(Participant 4) **No, I do not want to prepare the midterms**. It is **not practical**; we have large population of teachers and students...

(Participant 5) If I had one class, I would like to do it. **But it is not practical for now**. Now all the classes are taught the same things. **We should have a sort of standard test...**

The participants further agreed on wanting to prepare the quizzes for their classes:

(Participant 3) ...The quizzes may change; the percentages are not high, not have side affect. **Each teacher may prepare the quizzes**.

(Participant 4) ... But **the teachers may prepare the quizzes**.

(Participant 5) ... But **maybe quizzes, yes I would like to [prepare the quizzes]**.

Cooperation between the Teachers and Testers

Table 12 displays the results for questions in Section II of two questionnaires which are related to the cooperation between the teachers and testers, in particular teachers' opportunities to make comments on the tests.

Table 12

Items Related to the Cooperation between the Teachers and the Testers

Questions	Position	SD	D	U	A	SA	X2
Q1. There is active cooperation between the teachers and the testers in the test development process.	Teachers	3	7	5	10	5	4.66
	Testers	1	0	2	1	1	--
Q28. The teachers are provided with opportunities to make comments about the tests while the tests are being prepared.	Teachers	13	7	5	3	1	14.62**
	Testers	3	0	1	1	0	--
Q29. The teachers are provided with opportunities to make comments about the tests after the tests have been administered.	Teachers	5	8	5	8	4	2.33
	Testers	0	0	0	3	2	--
Q30. I think the teachers' comments about the tests are taken into consideration in the test development process.	Teachers	3	8	4	11	4	7.66
	Testers	0	0	1	2	2	--

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square ** p ≤ .01

As can be seen in Table 12 above, the chi square result for question 28 is highly significant at the level of $p \leq .01$, a result supported by the fact that most of the teachers (20) disagreed on not having the opportunity to make comments on the tests in the test preparation process. The non-significant results for questions 1, 29, 30 indicate that the teachers were divided otherwise on the extent to which they participate with the testers in the test development process. The testers agreed that teachers do not participate in the process of test preparation but generally claim to consider the teachers' comments while the tests are being made.

For question 28, the majority of the teachers (20) disagreed that they are given opportunities to make comments on the tests while the tests are being prepared. In order to gain deeper understanding for question 28, during the interviews, the testers were asked if the teachers are involved and whether they should be involved in the process of test preparation. Four of the testers (P6, P7, P8 and P9) disagreed on the teachers' active involvement during the test preparation process. Two participants (P6 and P7) suggested that the teachers might see the tests beforehand and comment on the test items:

(Participant 6) **No, no need** [for teachers to be involved in the test preparation process]. **The teachers may comment on the tests before the tests**; it may be useful.

(Participant 7) **No, I do not think so**. It may cause chaos. It is difficult to compile all different things they think. **Teachers may comment before the tests and after the tests**.

(Participant 8) **No**, if they are involved in the test preparation process then why is the testing office formed? **No need, but maybe few of them may be involved**.

(Participant 9) **No need for the others**. One of them already does, she is a proofreader of the tests. She checks the questions and makes comments if we make mistakes in the tests. We are five and six with a proofreader is enough... **The teachers should make comments about what is important for the students, they should report**.

Their ideas are important because we assess not only our students but also their students.

Only one tester (P10) strongly indicated that she would like to work with other teachers:

(Participant 10) **Yes, they should I think.** We need other teachers...I would like to work with a crowded group. **I want the other teachers to be involved in the test development process.**

The results for question 1 in Table 12 show that half of the teachers (15) agreed that there is cooperation with the testers in the test development process although they disagreed that they are allowed to see or make comments on the test during the process of test preparation. This suggests that there is not active cooperation between two groups during the whole process of test development. The interview data supported that there is no active cooperation during the test preparation process. During the interviews, the testers were asked if there is cooperation with the teachers, and all five testers indicated that other teachers are not involved in the test preparation process except after the test administration. Their responses below show, though, that they perceive cooperation as taking place because they are also teachers:

(Participant 7) **Of course there is cooperation if we call it so. We are both teachers and testers** and we know what is taught in the classes. Is it a kind of cooperation?

(Participant 8) **No, there is not cooperation with the other teachers, only among us as teachers.**

(Participant 9) **In fact, no.** Before the tests, they do not know the questions. After the tests we take their comments and take into account these comments.

(Participant 10) Yes, **only among us.** I want to have such cooperation in choosing important points for the students. They may check the items or questions.

For question 28, the majority of teachers (20) agreed that they are not involved in the test preparation process, and, during the interviews, they reported that they would like to be involved as teachers while the tests are being prepared. They also specified why their involvement as teachers in the test preparation process is useful. The teachers pointed to issues which were concerned with the accountability of the tests, and giving pre-feedback to the testers:

(Participant 1) **Yes, I would like [to be involved] as a teacher.** Better for teachers and testers. **The testers will guarantee themselves and they will see the problems beforehand.**

(Participant 4) **Absolutely, I would like to be involved as a teacher to make comments.**

(Participant 5) **Yes, I would like to be involved as a teacher.** There may be irrelevant or ambiguous questions. **I can give feedback on the items or tasks.** And only one proofreader sometimes doesn't work.

Interestingly, one participant (P2) insisted on preparing her own midterms and not including others:

(Participant 2) **Actually, I would like to prepare and evaluate the tests by myself.** Why should the other teachers be involved my class evaluation?

The responses given to question 29 show that 12 teachers agreed and 13 teachers disagreed on having the opportunity to make comments on the tests after the test administration. During the interviews, the teachers were asked whether they are provided with opportunities to make comments, and when they make these comments. The participants stated that they make comments after the tests are administered, but they further said that it is pointless to make comments on specific items after the tests:

(Participant 1) **Yes, after the tests.** In the past, we used to come together and make more comments than we do now about the specific test items generally not holistically.

(Participant 4) **Yes, after the tests.** The teachers know the best what to test and how to test. But it is useful to take the teachers' comments before the tests because of reliability of the tests. There may be ten or twenty teachers, for example. **However, generally the teachers' comments are concerned with mistakes or typos, items. These comments are not effective.**

(Participant 5) Yes. But we do not have frequent meetings. We can't see the exams and items. **We make comments after the tests. It is too late...**

The results for question 30 in Table 12 show that half of the teachers (15) agreed that the teachers' comments are taken into consideration while the tests are being prepared by the testers. Four teachers ticked uncertain for question 30, and 11 teachers disagreed about the testers' consideration of their comments. The interview transcripts supported that the teachers are divided as to whether these comments are considered. Two participants (P2 and P3) stated that the testers are trying to do their best by considering these comments:

(Participant 2) **Yes, I think so.**

(Participant 3) **Yes, I think that the testers are trying to do their best.** The next [test] comes better.

Two of the teachers (P1 and P5) were not totally sure about whether these comments are considered by the testers.

(Participant 1) **Sometimes** [they take our comments into consideration], but it is too late after the tests.

(Participant 5) ... **they say they take into consideration [our comments].** But again **I don't know.**

One of the teachers (P4) thinks that these comments are not considered by the testers:

(Participant 4) **I do not think so.** The comments focus on specific items, they do not need to care about these items, they won't use these items.

The testers were also asked about whether they take consideration of the teachers' comments during the interviews. They all agreed that they take these

comments into consideration for forthcoming tests, and one participant (P10) mentioned specific comments such as the appropriateness of the tests for the students' level, the test organization and rubrics:

(Participant 10) **Yes, we consider the comments.** Especially the way we ask questions, the level of students, and rubrics, whether the students may understand the rubrics **for [the] next tests.**

Teachers' and Testers' Perceptions of the Content of the Tests

Questions 6 and 7 in the Section II of two questionnaires were related to the content validity of the midterms and the quizzes, and questions 14 and 15 were concerned with the students' familiarity with the test content for the midterms and the quizzes. Table 13 below shows the results for questions 6, 7, 14 and 15.

Table 13

Items Related to the Content of the Tests

Questions	P	SD	D	U	A	SA	X2
Q6. The midterms reflect an appropriate sample of language skills relative to what has been taught in the class.	Tc	0	2	6	18	4	33.33**
	Ts	0	0	1	0	4	--
Q7. The quizzes reflect an appropriate sample of language skills relative to what has been taught in the class.	Tc	0	3	5	20	2	43.00**
	Ts	0	0	2	0	3	--
Q14. The students are aware of the content of the midterms.	Tc	0	1	3	15	11	29.33**
	Ts	0	1	0	1	3	--
Q15. The students are aware of the content of the quizzes.	Tc	2	8	6	9	5	5.00
	Ts	0	1	1	3	0	--

Note: P= Position; Tc= Teachers; Ts= Testers; SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square

** $p \leq .01$

Except for question 15, the chi square calculations for the questions presented in Table 13 above are significant at the level of $p \leq .01$. The non-significant result for question 15 indicates that the teachers are divided as to whether the students are familiar with the content of the quizzes beforehand. The significant results for questions 6, 7 and 14 show that most of the teachers agreed on appropriate sampling,

or content relevance, of the tests to teaching, and the students' familiarity with the test content for the midterms.

For questions 6 and 7, the majority of the teachers (22) think that the tests coincide with their teaching. This implies that most teachers feel that the tests possess content validity with reference to what is being taught in class.

The interview data also supports that the teachers perceive the tests as representative of their teaching and teaching practices in the classroom. One of the participants (P1) stated that the tests match with what she teaches but not with what she wants to teach. P1 further claimed that measuring only the course book objectives results from lacking clear objectives in the curriculum. In addition, one participant (P3) also agreed on the content relevance of the tests to his teaching objectives and practices. P5 indicated that the testers know what is being taught as teachers. During the interviews, the testers also stated that the tests appropriately sample what is being taught in the classes, one participant (P7) stating that the tests are in relation to teaching objectives and the syllabus:

(Participant 1) **Yes, they reflect what I teach but not what I want to teach** actually... I think **we are only measuring whether the students have achieved the course book objectives** because **our objectives are not clear in the curriculum.**

(Participant 3) **They do measure the objectives. The tests test what we teach, teaching objectives... Yes, the tests are matching with the teaching techniques in the class.** The same examples or questions appear in the tests as in the class.

(Participant 5) **Yes, they do** [match with what I teach]. **We have got curriculum and course book. Testers prepare according to them and they are teachers and know what is taught in the classes.**

(Participant 7) **The test questions are parallel with what we teach in the classes and the syllabus we follow.**

However, during the interviews, when the teachers were asked whether the tests reflect how they teach or learning activities which take place in the classrooms,

the responses were interesting. Two respondents (P1 and P4) stated that there are discrepancies between the tests and how they teach the language. P1 indicated that the tests do not correspond to teaching language communicatively, and P4 mentioned his focus on interaction in his teaching and the mismatch between those interactive learning activities and the tests:

(Participant 1) **No, there is not match between how I teach and the test. I believe language is communication not having good command of grammar or vocabulary.** I want match between how I teach or believe and the tests.

(Participant 4) **I focus on pair work and group work, talking to each other. But it is not tested using language in interaction with the others.**

In Table 13, the findings of question 14 show that most of the teachers (26) agree that the students know the content of the midterms. During the interviews, the teachers were asked whether it is useful for the students to have information about the test content beforehand. Most of the participants agreed that the students' familiarity with the test content is useful. P2 and P4 feel that helping students focus that way was positive. P3 suggested that it is useful for the students to know about the item types or the way the tests are organized. The general impression is that the teachers think that it is advantageous for the students to know about the test content, and one participant (P5) mentioned that doing this would help helping the students feel more comfortable:

(Participant 2) **Yes, why not? We can do it there is no harm in it. They know what to study according to this information.**

(Participant 3) **Yes, it is [useful]. If they know, they know how to study. You direct them into the way the questions are asked, advantageous.** But if you give details, they will study only that point. **Not only the content but also the way the tests are organized or prepared** should be informed the students. And it is useful also for the quizzes.

(Participant 4) **It is necessary and useful.** I think we should apply mock exams. But there is much burden for the testers. **They can prepare themselves or study for the tests and what they are going to be asked.** They need to know; otherwise it would be difficult to study lots of pages or points.

(Participant 5) **Certainly, it must be. Even the students should see some samples of the tests.** The students will see what they are going to be asked and we are doing this by giving the percentages of grammar, vocabulary, or writing part etc. not the content. It is the same for the quizzes. They may know the item types; **it would make the students feel more comfortable during the exam.**

Only one participant (P1) disagreed about informing the students about the test content. P1 suggested that the students would then only focus on the specific points which will be asked in the tests:

(Participant 1) ... **it is not useful, because the students will only focus on what will be tested in the exams.** It makes the students' job easier... My students only focus on grammar and vocabulary not for example, speaking. They are not interested in other things that won't be asked in the tests.

The testers were also asked about their opinions about informing the students about the test content during the interviews. Out of five testers, three of them (P6, P7 and P9) suggested that it is advantageous for the students to know the test content because it helps the students feel more relaxed, reducing anxiety:

(Participant 6) **Yes, it can be useful.** They should know the content and **it is helpful. They can be more motivated.** [It is] the same for the quizzes, **a kind of advantage for the students.**

(Participant 7) **Yes, of course. It is useful because we do not want to ask them what they do not know. This is motivating; they won't be surprised.** To know what they will be asked in the exams **makes them relaxed, and kills anxiety.** We inform the students about content of the tests on the notice board.

(Participant 9) **Yes, it is useful for the students. It reduces anxiety and stress. They know beforehand what to study and what they should know or be able to do in the language at that level.** The quizzes are progress quizzes and we inform them about the content. We have tried and it worked. We inform them orally and in written form. We informed them how they will be assessed for writing in the last quiz. It has really worked.

Two participants (P8 and P10) were against informing the students about the test content, and they claimed that directing the students' attention to specific points may prevent the students from considering the language as a whole:

(Participant 8) **No, I don't think so. The students should be ready for everything.** Unfortunately, we inform them. The quizzes should be pop quizzes. They don't need to know the content of these tests.

(Participant 10) **No, it is not useful. The students just focus on specific things according to this information. They do not understand that English is whole not separate points.** I think the same for the quizzes as well.

Teachers' and Testers' Perceptions of the Test Tasks

Questions 10, 11, 12, 13, 24 and 25 in Section II of the two questionnaires concerned the test tasks in terms of encouraging strategy use, reflecting real-life language use and the adequacy of the time allotted for the tests. Table 14 below displays the results for these questions.

Table 14

Items Related to the Test Tasks

Questions	Position	SD	D	U	A	SA	X2
Q10. The students are required to make use of different strategies to complete the tasks in the midterms.	Teachers	0	5	6	15	4	20.33**
	Testers	0	0	0	2	3	--
Q11. The students are required to make use of different strategies to complete the tasks in the quizzes.	Teachers	1	4	9	13	3	16.00**
	Testers	0	2	1	1	1	--
Q12. The tasks in the midterms reflect the features of real-life language use.	Teachers	3	5	12	10	0	16.33**
	Testers	1	0	0	3	1	--
Q13. The tasks in the quizzes reflect the features of real-life language use.	Teachers	3	7	14	6	0	18.33**
	Testers	1	0	2	1	1	--
Q24. A reasonable amount of time is given to complete the tasks in the midterms.	Teachers	0	0	1	17	12	42.33**
	Testers	0	0	0	1	4	--
Q25. A reasonable amount of time is given to complete the tasks in the quizzes.	Teachers	0	0	1	17	12	42.33**
	Testers	0	0	0	1	4	--

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square

** $p \leq .01$

Table 14 shows that the chi square calculations for questions 10, 11, 12, 13, 24 and 25 are highly significant at the level of $p \leq .01$. A majority of the teachers agreed that adequate time is allotted for the test tasks in the midterms and the quizzes, and that the midterms and the quizzes require the students to use different strategies. However, for questions 12 and 13, a plurality of the teachers was uncertain as to whether the tasks in the midterms and the quizzes reflect real-life language use. The interview data suggests that the teachers think that the authentic tasks are not often used in the tests, and that the testers have different understandings of what real-life language use means.

Both the teachers and the testers were asked the extent to which the test tasks reflect real-life language use during the interviews. According to the interview transcripts, two teachers (P2 and P3) agreed that the test tasks reflect the real life language use to certain extent and P3 clarified that these tasks are limited to functional language use in the tests.

(Participant 2) **Yes, they reflect [real-life language use].** They are useful because our students need English in real life.

(Participant 3) **Yes, but only in functional use. Not all the items. It should reflect in all skills, for example, in writing, academic texts. It depends on the department and the students' needs.**

Two of the participants (P1 and P5) do not find these tasks authentic at all.

(Participant 1) **Not too much. What we teach and test is only useful to pass future exams, not for their future or real life.**

(Participant 5) **No, but the authentic tasks or real life language use is very important. Such tasks should be used by the testers.** Because the language is same everywhere. They need in different contexts.

One of the participants (P4) suggested that such authentic tasks may not be appropriate for the students at lower levels:

(Participant 4) **They do not often reflect the real life usage of language. Also it is not appropriate for beginners. Maybe at higher levels.** It is not possible to use real life usage or authentic tasks in every test. I believe if the students are ready, they may be asked. When the testers were asked, during the interviews, whether they use

authentic tasks in the tests, all five participants stated that they make use of authentic tasks in their responses. Participants also elaborated how they select these tasks and what kind of tasks they utilize in the tests. However, what they pointed out was concerned thematic language use, not necessarily the reflection of genuine language use in the task types:

(Participant 7) **Yes, of course. Especially in functional use part of the tasks, such as how to use right language while shopping, or in restaurant...**

(Participant 8) **Yes, we should teach real life language use and test it.** For example, **how to take order, or make phone calls.** We take from the course book, I mean the subjects they have studied in the class.

(Participant 10) **Yes, in functional language use. We just test the real life language in,** for example, **writing dialogues or filling the gaps. We should assess orally and in reading tasks.**

Teachers' and Testers' Perceptions of the Test Items and Instructions

The following discussion is about the teachers' and testers' perceptions of the test items and instructions. Questions 16,17, 18, 19, 20, 21, 22 and 23 in Section II in the questionnaires were concerned with the properties of test items. Questions 16 and 17 were concerned with the students' familiarity with the item types used in the midterms and quizzes. Questions 18, 19 were concerned with the clarity of the instructions and questions 20 and 21 were about giving more than one correct answer in the test items. And questions 22 and 23 were about giving clues in the test items and test tasks. Table 15 below presents the results for these questions.

Table 15

Items Related to the Test Items and Instructions

Questions	Position	SD	D	U	A	SA	X2
Q16. The students are familiar with the item types used in the midterms.	Teachers	0	2	1	11	15	31.51**
	Testers	0	0	0	0	5	--
Q17. The students are familiar with the item types used in the quizzes.	Teachers	0	3	3	12	12	21.00**
	Testers	0	1	0	0	4	--
Q18. The instructions given in the midterms are clear.	Teachers	0	1	5	15	9	25.33**
	Testers	0	0	0	1	4	--
Q19. The instructions given in the quizzes are clear.	Teachers	0	2	3	16	9	28.33**
	Testers	0	0	0	1	4	--
Q20. There is more than one correct answer in some of the items used in the midterms.	Teachers	4	16	6	4	0	24.00**
	Testers	2	2	0	0	1	--
Q21. There is more than one correct answer in some of the items used in the quizzes.	Teachers	4	15	5	5	0	21.17**
	Testers	2	2	0	0	1	--
Q22. Some of the items in the midterms give clues to the answer of the other tasks.	Teachers	5	5	8	9	1	7.00
	Testers	3	1	1	0	0	--
Q23. Some of the items in the quizzes give clues to the answer of the other tasks.	Teachers	4	7	11	8	0	11.66*
	Testers	4	1	0	0	0	--

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square

** $p \leq .01$ * $p \leq .05$

The chi square calculations for questions 16, 17, 18, 19, 20 and 21 are significant at the level of $p \leq .01$. For questions 16, 17, 18 and 19, the majority of testers and teachers agreed that the students are familiar with the item types used in the tests and the instructions are clear for the students. For questions 20 and 21, the majority of the teachers disagreed that the test items have more than one correct answer. The chi square calculation for question 23 is only significant at the level of $p \leq .05$. The results for question 23 show that a plurality of the teachers were uncertain about whether the items in the quizzes provide clues for other items. The non-significant result for question 22 and the findings of question 23 indicate that the teachers are somewhat divided as to whether the test tasks give clues to the other tasks in the midterms. Similar to the findings of questions 22 and 23, the interview

data suggests that the teachers agree that some of the items give clues, but they are not totally sure about whether these clues are positive or negative.

The interview transcripts show that four teachers think that there are some clues given in the test items. However, the teachers mostly find these clues useful for the students' performance in the test. One participant (P1) stated that the clues sometimes given in reading tasks affect the students' performance in the test positively, and another participant (P2) said that the students are required to use some test strategies with these clues:

(Participant 1) **Sometimes, especially in reading tasks.** I think if they see the clues it means that they have understood the tasks, then **it affects positively.**

(Participant 2) **Sometimes, but this is useful, I think, we make the students use strategies.**

(Participant 4) **I haven't analyzed in this sense. But I do not think that it affects negatively.**

(Participant 5) **Sometimes.** I don't think that the students even recognize these clues.

Among five teachers, only one participant (P5) stated that these clues should not be given:

(Participant 3) **Sometimes,** but the items should not give clues.

During the interviews, the testers were also asked whether they include clues in the test items in the midterms and quizzes. All the testers stated that they avoid including clues in the test tasks in general. The data gathered from the interview transcripts show that the responses given during the interviews generally support the data gained from the testers' questionnaire. During the interviews, two testers (P6 and P9) admitted that they sometimes unintentionally include the clues in the test tasks. Additionally, P6 found these clues useful for the students whereas P9 stated

that these clues help the teachers understand if the students have understood the tasks:

(Participant 8) **I hope, no.** We all check the tests before we administer the test. Also proofreader does this.

(Participant 6) **Sometimes, not intentionally.** But mostly they do not understand the clues; **we recognize these clues sometimes afterwards. But sometimes they are useful in reading tasks.**

(Participant 9) **Sometimes unintentionally** for the midterms not quizzes. Because the quizzes are short. But I think **these clues sometimes help us [understand], whether the student has understood the tasks.**

Two of the participants (P7 and P10) were against using the clues in the test tasks because they believe that the clues may affect the students' performance negatively in the tests, which contradicts the point of view of the teachers:

(Participant 7) **No, it makes the tasks confusing. It may affect their performance negatively.** If they get the clue in a wrong way, they can answer the other tasks wrong again.

(Participant 10) **No, I don't think so.** The clues may affect their performance negatively. These are kind of copies. **We can't test their abilities appropriately**, if they answer one task wrong they will answer another wrong too.

Relationship between Learning and the Tests

The following discussion is concerned with teachers' and testers' perceptions of the relationship between the learning process and the tests. Questions 8, 9, 26, and 27 were related to whether the tests promote learning in terms of appropriateness for the students' proficiency level, and are given on an appropriate time schedule to promote learning. Table 16 below displays the results for questions 8, 9, 26 and 27.

Table 16

Items Related to the Relationship between Learning and the Tests

Questions	Position	SD	D	U	A	SA	X2
Q8. The midterms are appropriate for the proficiency level of the students.	Teachers	0	0	6	21	3	51.00**
	Testers	0	1	0	1	3	--
Q9. The quizzes are appropriate for the proficiency level of the students.	Teachers	1	1	8	17	3	30.66**
	Testers	0	1	1	1	2	--
Q26. The midterms are scheduled in a way that promotes learning.	Teachers	1	1	7	19	2	39.33**
	Testers	0	0	0	4	1	--
Q27. The quizzes are scheduled in a way that promotes learning.	Teachers	1	1	6	20	2	43.66**
	Testers	0	0	1	3	1	--

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square

** $p \leq .01$

The chi square calculation for each item is significant at the level of $p \leq .01$.

The results for questions presented Table 16 above indicate that most of the teachers agreed on the appropriateness of the tests for the students' proficiency level and appropriate time schedule for the tests.

The interview data supported the findings of the data taken from the questions presented in Table 16. Out of five teachers, two participants (P2 and P4) were uncertain about the appropriateness of the tests for the students' level, claiming that the tests should be more challenging. One participant (P5) mentioned some instructions are difficult for the students, and one of them (P1) pointed out that it is the program which overestimates the students' proficiency level:

(Participant 2) **Generally, yes. But they should be more difficult.**

(Participant 4) **Yes, it is appropriate. Sometimes, I find the exams not challenging or easier according to the students' level.**

(Participant 3) **Yes, appropriate.**

(Participant 5) **Yes, there is no problem. But sometimes I find some rubrics and instructions difficult for the students, but not too difficult.** And examiners or we do not explain, they need to understand.

(Participant 1) **The tests are appropriate but I think we teach and test intermediate level; indeed the students’ level is pre-intermediate.**

When the testers were asked, during the interviews, about whether the tests are appropriate for the students’ level, all participants agreed on the appropriateness of the tests. One tester (P9) also mentioned overestimating the students’ proficiency level. P10 indicated that the testers avoid asking questions higher than the students’ level, a situation which may lead the testers to produce unchallenging or easier tests:

(Participant 9) **In our institution, the level we teach is higher than their actual level. The tests match with what we teach but.**

(Participant 10) **Yes, it is appropriate. But we try not to ask the questions that are higher for their levels. What we ask is parallel with what they learn, I think.**

Teachers’ and Testers’ Perceptions of Effect of Tests on Teaching

The following discussion is about the teachers’ and testers’ perceptions of effect of the tests on teaching. Questions 4 and 5 concerned the effectiveness of the feedback obtained from the test results, and question 33, which was asked only to the teachers, explored the washback effect of the tests on teaching. Table 17 below shows the data gathered from questions 4, 5 and 33.

Table 17

Items Related to Effect of the Tests on Teaching

Questions	Position	SD	D	U	A	SA	X2
Q4. The results of the midterms provide me with useful information about how much my students have learned the course content.	Teachers	0	3	9	12	6	15.00**
	Testers	0	2	0	0	3	--
Q5. The results of the quizzes provide me with useful information about how much my students have learned the course content.	Teachers	2	2	8	15	2	23.00**
	Testers	0	1	1	2	1	--
Q33. I think the tests affect my teaching practices positively.	Teachers	1	2	7	16	4	24.33**

Note: SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square

** p ≤ .01

Table 17 shows that the chi square calculation for each item is highly significant at the level of $p \leq .01$. The results presented in Table 17 indicate most of the teachers agreed that the tests provide them with useful feedback on their students' progress in the course. In addition, the majority of the teachers agreed on the positive washback of the tests on teaching. The results of the questions in Table 17, suggest that there is more division among the testers; however, the interview data shows division among the teachers about the washback effect of the tests as well.

The interview transcripts showed that the teachers perceive the washback effect of the tests as both negative and positive on teaching practices. During the interviews, the teachers were asked if the tests affect their teaching practices, and if so, the teachers were asked whether this effect is positive or negative. Four of the participants claimed that the tests have negative effects on their teaching, which are test-driven instruction, dedicating class hours to the preparation for tests, and that the students pay more attention to the skills which will be measured in the tests:

(Participant 3) **Yes, negatively. I try to teach how the students or what they are asked in the tests.** It affects favorably.

(Participant 4) **No, it does not affect my teaching negatively. Sometimes I make revision before the exams,** however; I only waste few hours on this. **I do not teach for the tests. My beliefs are important; I do not care about the tests.**

(Participant 5) **Yes, to some extent negatively. Before the tests the students ask me to prepare them to the exam and we focus only on the exam at that time.** We can say negatively.

(Participant 1) **Yes, but negatively. Because my students only focus on grammar and vocabulary,** for example, **not speaking. They are not interested in other things which won't be asked in the tests.**

Out of five teachers, only one participant (P2) mentioned that the tests have positive effect on teaching and learning, which is to encourage the students to study harder:

(Participant 2) **Yes, the tests affect the teaching and learning positively. If there are no tests I don't think the students will study hard.**

Teachers’ and Testers’ Perceptions of Working in the Test Office

The last discussion of this section is concerned with the teachers’ and testers’ perceptions of working in the test office. Question 18 in Section III of the testers’ questionnaire and question 34 in Section II of the teachers’ questionnaire were designed as to elicit the testers’ and teachers’ views on working in the test office. Table18 below displays the data gained from questions 18, and 34 in the questionnaires.

Table 18

Teachers’ and Testers’ Perceptions of Working in the Test Office

S	Questions	Position	SD	D	U	A	SA	X2
3	Q18. I would like to work in the test office in the next academic year.	Testers	1	0	3	0	1	--
2	Q34. I would like to work in the test office in the next academic year.	Teachers	11	10	7	2	0	15.66**

Note: S= Section number in the questionnaire, SD= Strongly Disagree; D= Disagree; U= Uncertain; A= Agree; SA= Strongly Agree; X2= Chi Square
** $p \leq .01$

As is seen in Table 18, the chi square calculation for the question 34 is highly significant at the level of $p \leq .01$. This shows that the overwhelming majority of the teachers (28) do not want to work in the test office. Among the five testers, three of them ticked the option “uncertain” for question 18 whereas one agreed and one disagreed on working in the test office. The interview data clarified the testers’ perceptions of working in the test office.

During the interviews, the testers were asked what it is like to work in the test office, and whether it is difficult to be a member of the test office. All five testers agreed that it is difficult to work in the test office. Three participants (P9, P7 and P8)

stated that they are not paid whereas two of them (P7 and P8) mentioned workload of the testers both as testers and teachers:

(Participant 7) **We are not only testers but also teachers. We are not paid. It is not satisfactory.**

(Participant 8) **Yes, of course [difficult]. We are doing a voluntary work. It is too much work for us as a teacher and a tester.**

(Participant 10) **Yes, [it is] difficult. It takes a lot of time to work there. There is not one level and too much work. You have to decide and prepare quickly the tests.** It is heavy load.

(Participant 6) **Yes, very difficult.** Great responsibility, **we work too much.**

Only one of participants (P9) was positive about working in the test office stating that it contributes to self-development:

(Participant 9) **It is difficult, but it is delightful as much as it is difficult. I learn new things, and improve myself.** I love working even if **we are not paid.**

Teachers’ and Testers’ Additional Comments on Testing Practices

This last section presents the data collected through Section III of the teachers’ questionnaire and Section IV of the testers’ questionnaire. Table 19 below shows the data gained from three open-ended questions in Section III of the teachers’ questionnaire which were “Do you have any additional comments on the midterms, quizzes and overall testing practices?” regarding the aim of the study.

Table 19

Teachers' Questionnaire Section III

	<u>Strengths</u>		<u>Weaknesses</u>	
<i>Midterms</i>	Student friendly	1	Too long	2
	Periodic checks on the students' progress	1	Reliability and validity should be questioned	3
	Motivating	1	Limited to grammar only	5
	Effective assessment forms	2	Inappropriate to the students' proficiency level	2
	Contextual tasks	1		
<i>Quizzes</i>	<u>Strengths</u>		<u>Weaknesses</u>	
	Periodic checks on the students' progress	2	Hard to decide the students' overall performance	2
	Motivating	2	Inappropriate to the students' proficiency level	1
			Few of them	2
			Too long	2
<i>Overall Testing Practices</i>	<u>Strengths</u>		<u>Weaknesses</u>	
	Strong	1	The lack of oral exam	6
	Organized system	1	Not qualified testers	1
			Training on testing	1

Similar to the findings of interviews, the themes which emerged from Table 19 revealed that the lack of an oral exam, the focus on grammar in the tests and the quality of the tests in terms of reliability and validity received greater mention by the teachers. Six of teachers mentioned the lack of an oral exam, five respondents indicated that the tests are exclusively limited to grammar, and three of them suggested that the issues of reliability and validity should be explored for the tests. The interview data suggests that some teachers think that the lack of oral exam affects their teaching.

During the interviews, the teachers were asked whether the lack of oral exam affects their teaching and the students' performance in the class. Three participants (P1, P3 and P5) mentioned negative washback, indicating that their students do not pay attention to speaking because they are not assessed in terms of their performance in speaking. One participant (P3) stated that the lack of oral assessment affects not

only the students' performance in the classroom but also his teaching. P3 specified this negative washback as his focus on the things that will be asked in the tests:

(Participant 1) **Yes, now they do not give importance to speaking because it is not tested.** But, it does not affect my teaching style, I still focus on all skills.

(Participant 3) **Exactly, it affects the students' performance in the class. They sometimes become frustrated because there is no speaking test.** It is useful to focus on the class I think. **It affects my teaching, I feel like that I focus on what will be asked in the tests.** I individually assess my students' speaking.

(Participant 5) **Yes, the students do not pay attention the things that will not be assessed.** But **the oral assessment should be implemented.** But I don't know again whether this will affect positively, we haven't applied. But [it is] impossible with this large population of students.

P4 and P2 stated that the lack of formal oral exam has no effect on their teaching, and P2 justified that grammar and vocabulary is more important than speaking:

(Participant 4) I have never felt such a thing. We haven't seen whether it is more effective for their speaking skills because we haven't implemented such an exam before. **I focus on all the skills in my classes.**

(Participant 2) **No, it does not affect.** I think speaking is necessary, but not that important. **Grammar and vocabulary is more important.** Speaking is easy and if they know vocabulary and grammar, they can do it themselves.

In the testers' questionnaire, Section IV was provided to obtain the testers' additional opinions about the tests and the overall testing practices. Table 20 below shows the results for three open-ended questions in Section IV of the testers' questionnaire, which were "Do you have any additional comments on the midterms, quizzes and overall testing practices?"

Table 20

Testers' Questionnaire Section IV

	<u>Strengths</u>	<u>Weaknesses</u>	
	Measure integrated skills	1 The students know the content	1
		Too many of them	1
<i>Midterms</i>		Too long time provided for exams	2
		Physical settings of rooms	1
<i>Quizzes</i>	<u>Strengths</u>	<u>Weaknesses</u>	
	---	Unnecessary	1
<i>Overall</i>	<u>Strengths</u>	<u>Weaknesses</u>	
<i>Testing</i>	---	No make up for the final exam	2
<i>Practices</i>		Too much work	1
		Not assessing speaking	2

Similar to the point of view of the teachers, two testers mentioned the lack of oral exam as weakness of overall testing practices at Muğla University School of Foreign Languages. During the interviews both the testers were asked about their comments about the lack of oral assessment. Three participants (P6, P7 and P8) mentioned the negative washback of the lack of oral exam on teaching and learning, such as focusing more on the things to be asked in the tests, and students' paying more attention to other skills:

(Participant 6) Since we don't evaluate the students orally in exams, they think that we do not have right to ask them to speak in the class. But as teachers, we do assess with classroom participation as teacher assessment. Sometimes I find myself focusing on what will be asked in the tests, not speaking.

(Participant 7) It is a big problem for our institution. The students do not focus on speaking. Few students in my classes are willing to speak, not the others. We should do oral assessment to make them motivated. I assess their speaking skills in the class.

(Participant 8) Yes, they do not want to speak. The students want me to focus on writing and reading skills. The students are not interested in speaking. I focus on all skills in my classes, speaking as well.

One of the testers (P10) indicated that the teachers may assess their students within oral classroom exams:

(Participant 10) **We can test speaking in our classes. It should not be only one formal exam.** They will prepare only for that exam. Then they will forget it. **I want all the teachers assess their students' speaking skills in classroom exams.** We should stretch to the year.

One of the participants (P9) indicated that lack of oral exam does not affect her teaching in the class, and she linked the students' avoidance of speaking to their characteristics:

(Participant 9) Well, I think **in fact we assess their speaking as teachers in class performance not in the exams.** We give marks according to mostly on their speaking performance. **We somehow assess the speaking.** We observe and grade them. **It does not affect their performance,** they need to speak in the class. **Sometimes the performance is based on the students' characteristics.** We try to focus on all skills in our classes.

According to the interview transcripts above, all testers agreed that they assess their students on speaking with the classroom activities. During the interviews, the teachers were asked if they have any additional comments on the test or the testing system at Muğla University School of Foreign Languages. One teacher (P1) pointed out that members of the testing unit should be permanent whereas another (P3) suggested that the teachers need to be trained on how to implement oral assessment. One participant (P5) indicated that the criteria for writing assessment in the tests should be more clear, and she further stated that the teachers need training on assessment in order to evaluate their students individually in the classroom.

(Participant 1) There should be a good testing unit. There should be **permanent members** not changing the members every year.

(Participant 3) **We need some training on oral assessment.** Testers should not teach only work for the testing unit.

(Participant 5) Assessment of writing is problem sometimes. The criteria should be more certain. And **we need training on assessment as teachers, we also need to assess our students individually.**

When the testers were asked about their further comments about the test and the testing system in the institution, three testers (P6, P7 and P8) also agreed that they need in-service training about testing and assessment. P6 also stated that the reliability of the exams should be questioned whereas one participant (P9) demanded assistance from the administrators about their teaching hours. Another respondent (P10) emphasized having clear and explicit curriculum objectives, which is closely related to the discussions in earlier sections:

(Participant 6) **We need more qualified testers.** It would be better then. The **reliability of the exams** can be questioned. And **in-service training** may be provided.

(Participant 7) **We are not qualified enough. We need training and trainer specialized in assessment.**

(Participant 8) We **need seminars and workshops** maybe on testing and assessment.

(Participant 9) Sometimes even as five people we can not cooperate with each other and we should have more time to work. It is too much work; **the administrators should be more helpful about the course schedule for testers.**

(Participant 10) I am satisfied, but there may be some changes. **We need a strong syllabus and explicit objectives. We can see what our students should be able to do in the course and facilitates test preparation.**

Conclusion

In this chapter, the data gained from the questionnaires and the interviews were analyzed and presented in three sections. In the first section, the data was consisted of the Likert scale questions in Section III of the testers' questionnaire and the interview transcripts which answered research question 1. In the second section, the data composed of Likert scale questions in Section II of both the teachers' and

testers' questionnaire and the interview data which sought to answer research questions 2, 3 and 4. In the last section, the data gained from three open-ended questions in Section III of the teachers' questionnaire and Section IV of the teachers' questionnaire was analyzed qualitatively.

The next chapter will present an overview of the study, the discussion of findings, pedagogical implications, limitations of the study, implications for further research, and conclusion.



CHAPTER V: CONCLUSION

Introduction

This study investigated testers' perceptions of the test development process at Muğla University School of Foreign Languages and also explored the teachers' and testers' attitudes towards the resulting achievement tests. The research questions posed for this study were as follows:

1. What are the testers' perceptions of the test development process at Muğla University School of Foreign Languages?
2. What are the teachers' attitudes towards the resulting achievement tests used at Muğla University School of Foreign Languages?
3. What are the testers' attitudes towards the resulting achievement tests used at Muğla University School of Foreign Languages?
4. Are there any discrepancies between the teachers' and testers' attitudes towards the achievement tests used at Muğla University School of Foreign Languages?

In this chapter, overview of the study, discussion of findings, pedagogical implications, limitations of the study, implications for further research, and conclusion will be presented.

Overview of the Study

Thirty English teachers and five testers who are currently working for Muğla University School of Foreign Languages participated in this survey study. The

researcher used two data collection instruments in this study. First, the teachers and testers were given a questionnaire. Second, semi-structured interviews with five randomly chosen teachers and five testers were conducted one month after the distribution of the questionnaires.

The teachers' questionnaire was composed of three sections. In Section I, there were 7 questions which aimed to obtain background information about the teachers. Section II consisted of five-point Likert scale ranging from strongly disagree (SD) to strongly agree (SA). 35 Likert-scale items in Section II of the questionnaire aimed to investigate the teachers' attitudes towards the achievement tests. Section III consisted of three-open ended questions which were designed in order to obtain the teachers' additional comments on the midterms, quizzes and overall testing practices.

The testers' questionnaire was organized similarly to the teachers' questionnaire. In Section I, 7 questions were designed in order to obtain background information about the testers. Section II and Section III consisted of five-point Likert scale ranging from Strongly Disagree (SD) to Strongly Agree (SA). There were 30 Likert-scale items in Section II of the testers' questionnaire which aimed to investigate the testers' attitudes towards the achievement tests. Section III was also composed of 18 Likert-scale items which were designed to investigate the testers' perceptions of the test development process at Muğla University School of Foreign Languages. In Section IV, there were three open-ended questions which aimed to explore the testers' additional comments about the midterms, quizzes and overall testing practices.

Two different but related schedules of interview questions were designed to be asked to teachers and testers following the analysis of data gathered from the

questionnaires. The interview questions were asked to the participants in order to allow them to clarify the information they provided in the questionnaires.

Discussion of Findings

The results of the data analysis suggested that both the teachers and testers are generally satisfied with the progress achievement tests used at Muğla University School of Foreign Languages. The results of the questionnaires and the participants' responses during the interviews revealed that both teachers and testers agreed that the tests closely match with teaching practices and provide effective feedback on the product of learning. However, the data gained from both the questionnaires and the interviews pointed to such weaknesses of the current testing practices as the lack of explicit objectives in the curriculum, the negative effect of the tests on teaching and learning, the lack of test specifications for the test preparation process, and the limited number of authentic tasks used in the tests. In addition to these weaknesses, the interview data indicated that there is no active cooperation between the teachers and testers in the test preparation process. Finally, the results from the questionnaires indicated that the majority of the teachers do not want to work in the test office, and the testers have complaints about having excessive working hours and not being paid for their work in the test office.

The analysis of the data gathered from the questionnaires demonstrated that the majority of the teachers (22) consider the tests an appropriate sample of what is being taught in classes, and four testers also agree on the appropriate sampling of the tests. The interview data supported that both the teachers and testers feel that the tests possess content validity. During the interviews, all five teachers agreed that the tests reflect what they teach in the class. When tests coincide with the teaching that precedes them, it is possible for teachers to obtain accurate information about how

much their students have learned (Bachman, 1990). Two teachers indicated that there are discrepancies between the tests and how they teach in the class. This results from these two teachers' beliefs about communicative and interactive approaches to language teaching and learning because the tests used at Muğla University School of Foreign Languages do not adopt a communicative or interactive approach.

In data analysis of the questionnaires, it was found that four testers agreed on the use of course objectives as reference while preparing the tests. However, the interview data revealed that four testers perceived the course book objectives as course objectives. Out of five testers, only one of them disagreed about referring to the course objectives while preparing the tests. During the interviews, all five testers stated that they make exclusive use of the course book objectives because there are no explicit objectives in the curriculum and syllabus. Similar to the testers, during the interviews, two teachers also stated that the tests are limited to measure only whether the students have mastered the course book objectives due to the lack of clear objectives in the curriculum. Brown (1995) mentions that there is no such thing as a perfect course book. Ur (1996) points out that "course books have their own rationale and chosen teaching and learning approach. They do not usually cater for the variety of levels of ability and knowledge" (p. 185). The objectives of a course book may not necessarily serve the goals of a language program. The nature of curriculum should be clear to teachers and tester in terms of its goals and objectives (Brown & Hudson, 2002). Then, it will be possible for testers to specify the test content appropriately and to inform other stakeholders in an educational program about the expected content of each test.

The results of the study demonstrated that the tests sometimes have a negative effect on teaching practices and learning. The analysis of the data gathered

from the teachers' questionnaire showed that 20 teachers agreed on the positive effect of the tests on their teaching practices. However, during the interviews, out of five teachers, four of them mentioned the negative effects of the tests, such as dedicating class hours to the test preparation, focusing on specific points to be asked in the tests, exclusion of oral skills, and the students' attention to certain points which will be covered in the tests, which can be seen the negative washback of the tests (Brown, 1999). Similarly, Kohonen (as cited in Arnold, 1998) also points out that "anticipation of the testing procedures has a washback effect on learning because students prepare for the tests and organize their knowledge in memory in the light of how they are going to be tested" (p. 279). During the interviews, teachers talked about informing the students about the test content both negatively and positively. The students' focusing on the specific skills was mentioned as the negative effect, while reducing anxiety and making the students more comfortable in the test were mentioned as the positive aspects.

In initial analysis of the questionnaires in the first section of Chapter 4, all 5 testers stated that the tests are carefully planned. However, further data analysis revealed that they do not construct test specifications prior to the test preparation process. All 5 testers indicated that they only prepare some parts of a complete test specification such as listing the language skills to be measured and the scoring criteria. However, the first step of the test development process should ideally be the construction of test specifications, that are generally the statements about what a test will measure and what the test will be like (Alderson et al., 1995; Bachman & Palmer, 1996; Brown & Hudson, 2002; Davidson & Lynch, 2002). According to Bachman and Palmer (1996) such statements help testers have clear understanding of the purpose and content of the test, test methods, organization of the test. It should be

noted that having test specifications for the tests help the testers evaluate test usefulness in the test development process. Additionally, test specifications are not only useful for testers but also for teachers to evaluate the validity and reliability of tests. The teachers may determine whether the tests serve the intended purpose, coinciding with teaching and the learning process, and whether the decisions made based on the test results are consistent with the students' language abilities and knowledge of language.

The findings of the study demonstrated that the testers do not often use the tasks which correspond to real life language use for the tests. During the interviews, four of teachers indicated that the authentic tasks are not often used in the tests, and that these tasks are limited to the 'functional-language use' part of the midterms. Testers were also asked about the extent to which they use authentic tasks in the tests during the interviews. Participants' responses revealed that the real life language use was perceived by the testers as thematic functional language use. The results showed that all 5 testers agreed that they utilize authentic tasks; however, what they have pointed out was not the use of the tasks which closely correspond to real life language use. However, authenticity is the degree to which test performance replicates specified non-test language performance (Bachman, 1990). It may not be possible for all testers and teachers to specify appropriately non-test or real life situations, and to sample the tasks relevant to all students' needs in the target language (Bachman, 1991). According to Purpura (1995), in authentic tasks, test takers or students are encouraged to produce the target language which might be seen in genuine language use. However, it is not always easy to sample genuine language use in testing situations and to determine the scoring criteria on which the students will be assessed. This may be the case for testing practices at Muğla University

School of Foreign Languages because the students' perceptions of real life language use may diverge according to their academic interests and their disciplines.

Data analysis of the questionnaires indicated that both teachers and testers agree that there is cooperation between them in the test development process. The results of the data gathered from the questionnaire revealed that half of the teachers (15) agreed on the cooperation with the testers, and four testers agreed on the cooperation with the teachers in this process. However, the interview data revealed that this cooperation takes place only after the test administration process, and that there is no active cooperation between the teachers and testers in the test preparation process. During the interviews, 4 teachers stated that they would like to be involved in the test preparation process by making comments on the overall organization of the test, the items, and instructions. They also indicated that such cooperation is helpful for the testers in terms of obtaining pre-feedback before the test administration. Teachers may also make judgments about the relevance of the tests to the course content and to the tasks their students are required to complete during the course (Alderson et al., 1995). This may contribute to the quality of the tests because the testers will ensure that the test have relatively higher the degree of content validity. More than half of teachers (17) indicated in the questionnaire that they would like to prepare quizzes while the same number disagreed about preparing the midterms because of potential problems with the lack of standards across classes and inequality of assessment. All five testers also stated, during the interviews, that it would be better to take the teachers' comments about the tests before the test administration process. However, four testers do not agree that the teachers should be involved in the test preparation. Three of them indicated that they are able to ensure the content validity of the test because they are also teachers at different levels.

Finally, most of the teachers stated in the questionnaire that they do not want to work in the test office. When the testers are being interviewed, they mentioned the difficulties of working in the test office such as loaded working hours working as both teachers and testers; and not being paid for working as testers.

Pedagogical Implications

Lacking explicit objectives in the curriculum has received mention both by the teachers and the testers in the study. As results of the study suggested, the lack of clear curriculum objectives leads the teachers and testers to feel that the students are expected to master only the course book objectives throughout the year. If the teachers are trained about writing learning objectives according to what they expect from the students when they leave the School of Foreign Languages at Muğla University, they may articulate objectives for the curriculum and the syllabus in cooperation with the curriculum unit.

Based on the data gained from both the questionnaires and interviews, testers at Muğla University School of Foreign Languages do not generate test specifications prior to the test design process. Testers might construct test specifications which may not be detailed but include a general framework for each test. This framework may involve the objectives to be measured and the item types for the language skills to be measured.

Teachers may also be provided with opportunities to make comments about the tests during the test development process, before the tests are administered, in the meetings with the testers. Additionally, in these meetings, the testers may give feedback to the teachers about how the teachers' comments are taken into consideration for forthcoming tests. The majority of the teachers would like to prepare the quizzes for their own classes in cooperation with their teaching partners.

The teachers may be assigned to prepare their own quizzes for their own classes. Another alternative might be adding the weighting of the quizzes to the weighting of teacher assessment in overall students' assessment in order to encourage teacher control of their students' assessment.

According to the data analysis, the teachers believe that a negative effect of the tests on teaching and learning results from informing the students about the content of the tests and the lack of oral exam. Teachers and testers stated that the students pay attention during the course mostly to the specific points which will be assessed in the tests, such as knowledge of grammar. The students may be informed about the learning objectives they are expected to achieve in the tests instead of the priority of language skills to be measured in the tests. As for the lack of formal oral assessment in the institution, the teachers may be trained about how oral assessment can be embedded into instruction through in-service training. Another alternative may be for the teachers to reach consensus about how to implement oral assessment within class activities and the criteria on which the students will be assessed for these activities.

Finally, the results of the questionnaires revealed that most of the teachers are not interested in working in test office, and the interview data suggested that the testers have complaints about the difficulties of working in the test office. It may be possible for the administrators to make the testing work more attractive giving financial support to the testers and/or reducing the teaching hours of the testers.

Limitations of the Study

This survey study investigated the testers' perceptions of the test development process at Muğla University School of Foreign Languages, and further explored the teachers' and testers' attitudes towards the resulting achievement tests.

The results of this study may not be generalized to other educational contexts because it was a local study which involved 30 teachers and 5 testers working at Muğla University School of Foreign Languages. In addition, this study relied on the testers' responses in the questionnaires and the interviews for their perceptions of the test development process. As another data collection instrument, observations would have been made during the development process of one or more than one tests in order to obtain genuine data about the procedures in the test development process.

Implications for Further Research

This study explored the teachers' and testers' attitudes towards the tests used at Muğla University School of Foreign Languages. The first alternative for further study may be to add observations to support the data gained for the questionnaires and the interviews. Another study may involve the students as well, and aim to explore the students' perceptions of the tests in terms of the content relevance of the tests to the instruction and learning activities which take place in the class.

Since this survey study was at the local level, it may be replicated by other researchers from other universities where the conditions are similar to those of Muğla University School of Foreign Languages. This would allow representatives of language programs to evaluate the tests which are systematically used in their institutions in terms of relevance to the course content and to the learning objectives in the curriculum.

Another alternative for further research would be exploring the teachers' and testers' points of view about alternative assessment forms. Finally, exploratory research may be conducted in order to explore the impact of classroom assessment activities, particularly oral assessment, on teaching and the language learning process.

Conclusion

The research investigated testers' perceptions of the test development process at Muğla University School of Foreign Languages, and explored the teachers' and testers' attitudes towards the resulting achievement tests. The data was collected through two parallel questionnaires, and interviews. The findings of the questionnaires and the interviews revealed that both teachers and testers are generally content with the tests in terms of the content validity and providing useful feedback on the learning. The results also suggested that there are weaknesses in the current testing practices due to the lack of explicit objectives in the curriculum, the negative effect of the tests to certain extent on teaching and learning, the lack of test specifications for the tests, and the limited number of authentic tasks used in the tests. The study also pointed to the lack of active cooperation between the teachers and testers in the test preparation process. Finally, the results showed that the testers have some complaints about workload, and not being supported financially. It is possible for the stakeholders at Muğla University School of Foreign Languages to achieve better testing practices with consideration of the recommendations given in this chapter.

REFERENCE LIST

- Aksan, M. M. (2001). *Instructors' Perceptions of the content validity of the English language exams at Niğde University*. Unpublished master's thesis, Bilkent University, Ankara.
- Alderson, J. C., & Clapham, C. (1995). Assessing student performance in the ESL classroom. *TESOL Quarterly*, 29, 184-187.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14, 115-129.
- Bachman, L. F. (1989). The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 242-258). Cambridge: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, 25, 671-704.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *TESOL Quarterly*, 13, 125-149.
- Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257-279.
- Bailey, K. (1998). *Learning about language assessment*. London: Heinle & Heinle.
- Brindley, G. (2001). Assessment. In R. Carter & D. Nunan (Eds.), *Teaching English to speakers of other languages* (pp. 137-143). Cambridge: Cambridge University Press.
- Brown H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.

- Brown J. D. (1995). *The elements of language curriculum*. Boston: Heinle & Heinle.
- Brown J. D., (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Brown, J. D. (1999). Using testing to unify language curriculum. In M. Harvey & S. Kanburoğlu (Eds.), *Proceedings of the 4th International ELT Conference, Achieving a coherent curriculum: Key elements, methods and principles*, Bilkent University, Ankara, Turkey.
- Brown, J. D. (2000). University entrance examinations: Strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 3, (2), 4-8.
- Brown, J. D. (2002). Extraneous variables and the washback effect. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 6, (2), 12-15.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 653-675.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press
- Bond, A. L. (1996). Norm- and criterion-referenced testing. *Practical Assessment, Research & Evaluation*, 5, (2). Retrieved January 14, 2004 from the World Wide Web: [http:// www.PAREonline.net/](http://www.PAREonline.net/)
- Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd Ed.). Boston: Heinle & Heinle.
- Cohen, A. D. (2001). Second language assessment. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd Ed.) (pp. 515-535). London: Heinle & Heinle.
- Coombe, C., & Hubley, N. (1999). *Creating effective classroom tests*. Retrieved on April 24, 2004 from the World Wide Web: <http://taesig.8m.com/createii.html>
- Dalyan, A. (1990). *The comments of EFL teachers on preparing achievement tests and an analysis of a sample test*. Unpublished master's thesis, Anadolu University, Eskişehir.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.
- Davies, A. (1990). *Principles in language testing*. Oxford: Basil Blackwell Ltd.
- Davies, P., & Pearse, E. (2000). *Success in language teaching*. New York: Oxford University Press.

- DeVincenzi, F. (1995). Examining standardized test content: Some advice for the teachers. *TESOL Quarterly*, 29, 180-184.
- Dietel, R. J., Herman, J. L., & Knuth, R. A. (1991). *What does research say about assessment?* Retrieved February 12, 2004 from World Wide Web: <http://www.dundee.ac.uk/languagestudies>
- Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.
- Gorsuch, G. J. (1997). *Test purposes*. Retrieved February 12, 2004 from the World Wide Web: <http://www.jalt-publications.org/tlt/files/97/jan/gorsuch.html>
- Halleck, G. B., & Moder, C. L. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly*, 29, 733-758.
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14, 3, 295-303.
- Hoekje, B., & Linnel, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, 28, 103-123.
- Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal*, (5), 8-11.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Johnston, B. (2003). *Values in English language teaching*. London: Lawrence Erlbaum Associates.
- Kenyon, D. M. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 19-40). Mahwah, NJ: Lawrence Erlbaum.
- Ketevu, E. (2001). *Investigating the validity of achievement and proficiency tests at Bilkent University School of English Language*. Middle East Technical University, Ankara.
- Kohonen, V. (1999). Authentic assessment in affective foreign language education. In Arnold, J. (Ed.), *Affect in language learning*, (pp. 279-293). Cambridge: Cambridge University Press.
- Kuntasal, İ. (2001). *Perceptions of teachers and testers of achievement tests prepared by testers in the Department of Basic English at Middle East Technical University*. Unpublished master's thesis, Bilkent University, Ankara.

- Lewkowicz, J. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, 17, 43-64.
- Lynch, B. K. (1990). A context-adaptive model for program evaluation. *TESOL Quarterly*, 24, 24-43.
- Lynch, B. K., & Davidson, F. (1994). Criterion-referenced language test development: Linking curricula, teachers, and tests. *TESOL Quarterly*, 28, 727-743.
- Madaus, G. F., Haney, W., & Kreitzer, A. (1996). *Testing and evaluation primer*. Retrieved March 21, 2004 from the World Wide Web:
<http://www.dundee.ac.uk/languagestudies/ltest/ltrfile/ltrframe.html>
- McMillan, J. C. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation*, 7, (8). Retrieved January 14, 2004 from the World Wide Web:
[http:// www.PAREonline.net/](http://www.PAREonline.net/)
- McNamara, T. (2000). *Language Testing*. New York: Oxford University Press.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18, 329-332.
- Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass.
- Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Printer Publishers.
- Purpura, J. E. (1995). *Fundamental considerations in the design of CB language tests*. Paper presented at the 1st INGED Conference, Middle East Technical University, Ankara, Turkey.
- Ransom, K. A., & Santa, C. M. (1999). High-stakes assessments in reading: a position statement. *The Reading Teacher*, 53, (3), 257-264.
- Rudner, (1994). *Questions to ask when evaluating tests*. Retrieved January 14, 2004 from the ERIC Database.
- Rudner, L. M., & Schafer, W. D. (2001). *Reliability*. Retrieved March 21, 2004 from the ERIC Database.
- Serpil, H. (2000). *An assessment of the content validity of the midterm achievement tests administered at Anadolu University Foreign Languages Department*. Unpublished master's thesis, Bilkent University, Ankara.
- Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, 14, 341-349.

- Shohamy, E. (2001). *The power of tests*. Harlow, UK: Longman.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299-321.
- Spence-Brown, R. (2001). The eye of beholder: Authenticity in an embedded assessment task. *Language Testing*, 18, 463-481.
- Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.
- Taylor, L. B. (2004). Testing times: Research directions and issues for Cambridge ESOL examinations, *TESOL Quarterly*, 38, 141-147.
- Thomas, M. (2003). Assessment and learner performance in problem based learning (PBL). *Journal of Language and Learning*, 1, (1), 1-13.
- Ur, P. (1996). *A course in language teaching*. Cambridge: Cambridge University Press.
- Wall, D., & Alderson, J. C. (1996). Examining washback.: The Sri Lankan Impact Study. In A. Cumming, & R. Berwick (Eds.), *Validation in language testing*, (194-221). Clevedon, UK: Multilingual Matters.
- Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, UK: Prentice Hall International (UK) Ltd.
- Weir, C. J. (1995). *Understanding and developing language tests*. New York: Phoenix ELT.
- Wiggins, G (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, 2, (2). Retrieved January 14, 2004 from the World Wide Web: [http:// www.PAREonline.net/](http://www.PAREonline.net/)

APPENDIX A
TEACHERS' QUESTIONNAIRE

Dear Colleagues,

I am currently enrolled in the 2004 MA TEFL Program at Bilkent University. I am conducting research on teachers' and testers' perceptions of the tests used at Muğla University School of Foreign Languages. The study also aims at exploring the testers' perceptions of the test development process.

This questionnaire is the first phase of this study. The second phase will be interviews to be held with the testers and with randomly selected teachers following the initial analysis of this data. Your answers will make an invaluable contribution to my research, providing important information about the adequacy and effectiveness of the tests and the testing practices. Your completion of the questionnaire will be regarded as consent in order to use the data supplied for the purposes of this research. The personal background information you provide will be used only for this research and under no circumstances will be shared.

If you have any questions please do not hesitate to contact with me or my thesis advisor. I thank you very much in advance for your help and cooperation.

Elif Aydın
MA TEFL Program
Bilkent University, ANKARA
Tel: (090) 312 290 61 45
aelif@bilkent.edu.tr

Dr. Bill Snyder
MA TEFL Program
Bilkent University, ANKARA
Tel: (090) 312 290 15 59
wsnyder@bilkent.edu.tr

SECTION I.

Please circle the appropriate answer for you and provide the necessary information below.

1. Gender a. Male b. Female
2. Degree Program(s) completed:
a. BA Degree b. MA Degree c. PhD Degree
3. Years of teaching experience:
a. 1-4 years b. 5-8 years c. 9-12 years d. 13-16 years e. more than 16
4. Years of teaching experience at Muğla University School of Foreign Languages:
a. 1-4 years b. 5-8 years c. 9-12 years d. 13-16 years e. more than 16
5. Have you taken any courses on 'testing', 'measurement and evaluation', or 'assessment' at the university from which you graduated? If yes, please specify which courses you have taken?
a. Yes b. No

University _____ Course (s) _____

6. Have you ever worked in the testing office at Muğla University School of Foreign Languages? If yes, please specify the dates.

- a. Yes b. No

Dates _____

7. Have you ever worked in the testing office at another institution? If yes, please specify the institution and the dates.

- a. Yes b. No

Institution _____ Dates _____

SECTION II.

The questions below are concerned with the tests prepared by the testing office at Muğla University School of Foreign Languages. Please put a cross (X) in the box the most closely corresponds to your opinion.

No	Items	Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree
		SD	D	U	A	SA
1.	There is active cooperation between the teachers and the testers in the test development process.					
2.	I am satisfied with the midterms.					
3.	I am satisfied with the quizzes.					
4.	The results of the midterms provide me with useful information about how much my students have learned the course content.					
5.	The results of the quizzes provide me with useful information about how much my students have learned the course content.					
6.	The midterms reflect an appropriate sample of language skills relative to what has been taught in the class.					
7.	The quizzes reflect an appropriate sample of language skills relative to what has been taught in the class.					
8.	The midterms are appropriate to the proficiency level of the students.					
9.	The quizzes are appropriate to the proficiency level of the students.					
10.	The students are required to make use of different strategies to complete the tasks in the midterms.					
11.	The students are required to make use of different strategies to complete the tasks in the quizzes.					

No	Items	SD	D	U	A	SA
12.	The tasks in the midterms reflect the features of real-life language use.					
13.	The tasks in the quizzes reflect the features of real-life language use.					
14.	The students are aware of what the content will be in the midterms.					
15.	The students are aware of what the content will be in the quizzes.					
16.	The students are familiar with the item types used in the midterms.					
17.	The students are familiar with the item types used in the quizzes.					
18.	The instructions given in the midterms are clear.					
19.	The instructions given in the quizzes are clear.					
20.	There is more than one correct answer in some of the items used in the midterms.					
21.	There is more than one correct answer in some of the items used in the quizzes.					
22.	Some of the items in the midterms give clues to the answers of the other tasks.					
23.	Some of the items in the quizzes give clues to the answers of the other tasks.					
24.	A reasonable amount of time is given to complete the tasks in the midterms.					
25.	A reasonable amount of time is given to complete the tasks in the quizzes.					
26.	The midterms are scheduled in a way that promotes learning (e.g. before holidays).					
27.	The quizzes are scheduled in a way that promotes learning (e.g. within appropriate class time).					
28.	The teachers are provided with opportunities to make comments about the tests while the test are being prepared.					
29.	The teachers are provided with opportunities to make comments about the tests after the tests have been administered.					
30.	I think the teachers' comments about the tests are taken into consideration in the test development process.					
31.	I would like to prepare the midterms for my classes in cooperation with my partners.					
32.	I would like to prepare the quizzes for my classes in cooperation with my partners.					
33.	I think the tests affect my teaching practices positively.					
34.	I would like to work in the test office in the next academic year.					

SECTION III.

If you have additional comments or opinions, please specify in the boxes below.

The Midterms	Strengths	
	Weaknesses	
The Quizzes	Strengths	
	Weaknesses	
Overall Testing Practices at Muğla University School of Foreign Languages	Strengths	
	Weaknesses	

APPENDIX B

TESTERS' QUESTIONNAIRE

Dear Colleagues,

I am currently enrolled in the 2004 MA TEFL Program at Bilkent University. I am conducting research on teachers' and testers' perceptions of the tests used at Muğla University School of Foreign Languages. The study also aims at exploring the testers' perceptions of the test development process.

This questionnaire is the first phase of this study. The second phase will be interviews to be held with the testers and with randomly selected teachers following the initial analysis of this data. Your answers will make an invaluable contribution to my research, providing important information about the adequacy and effectiveness of the tests and the testing practices. Your completion of the questionnaire will be regarded as consent in order to use the data supplied for the purposes of this research. The personal background information you provide will be used only for this research and under no circumstances will be shared.

If you have any questions please do not hesitate to contact with me or my thesis advisor. I thank you very much in advance for your help and cooperation.

Elif Aydın
MA TEFL Program
Bilkent University, ANKARA
Tel: (090) 312 290 61 45
aelif@bilkent.edu.tr

Dr. Bill Snyder
MA TEFL Program
Bilkent University, ANKARA
Tel: (090) 312 290 15 59
wsnyder@bilkent.edu.tr

SECTION I.

Please circle the appropriate answer for you and provide the necessary information below.

1. Gender a. Male b. Female
2. Degree Program(s) completed:
a. BA Degree b. MA Degree c. PhD Degree
3. Years of teaching experience:
a. 1-4 years b. 5-8 years c. 9-12 years d. 13-16 years e. more than 16
4. Years of teaching experience at Muğla University School of Foreign Languages:
a. 1-4 years b. 5-8 years c. 9-12 years d. 13-16 years e. more than 16

5. Have you taken any courses on 'testing', 'measurement and evaluation', or 'assessment' at the university from which you graduated? If yes, please specify which courses you have taken?

a. Yes b. No

University _____ Course (s) _____

6. How long have you been working in the testing office at Muğla University School of Foreign Languages? For _____ year (s)

7. Have you ever worked in the testing office at another institution? If yes, please specify the institution and the dates.

a. Yes b. No

Institution _____ Dates _____

SECTION II.

The questions below are concerned with the tests prepared by the testing office at Muğla University School of Foreign Languages. Please put a cross (X) in the box the most closely corresponds to your opinion.

No	Items	Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree
		SD	D	U	A	SA
1.	There is active cooperation between the teachers and the testers in the test development process.					
2.	I am satisfied with the midterms.					
3.	I am satisfied with the quizzes.					
4.	The results of the midterms provide me with useful information about how much my students have learned the course content.					
5.	The results of the quizzes provide me with useful information about how much my students have learned the course content.					
6.	The midterms reflect an appropriate sample of language skills relative to what has been taught in the class.					
7.	The quizzes reflect an appropriate sample of language skills relative to what has been taught in the class.					
8.	The midterms are appropriate to the proficiency level of the students.					

No	Items	SD	D	U	A	SA
9.	The quizzes are appropriate to the proficiency level of the students.					
10.	The students are required to make use of different strategies to complete the tasks in the midterms.					
11.	The students are required to make use of different strategies to complete the tasks in the quizzes.					
12.	The tasks in the midterms reflect the features of real-life language use.					
13.	The tasks in the quizzes reflect the features of real-life language use.					
14.	The students are aware of what the content will be in the midterms.					
15.	The students are aware of what the content will be in the quizzes.					
16.	The students are familiar with the item types used in the midterms.					
17.	The students are familiar with the item types used in the quizzes.					
18.	The instructions given in the midterms are clear.					
19.	The instructions given in the quizzes are clear.					
20.	There is more than one correct answer in some of the items used in the midterms.					
21.	There is more than one correct answer in some of the items used in the quizzes.					
22.	Some of the items in the midterms give clues to the answers of the other tasks.					
23.	Some of the items in the quizzes give clues to the answers of the other tasks.					
24.	A reasonable amount of time is given to complete the tasks in the midterms.					
25.	A reasonable amount of time is given to complete the tasks in the quizzes.					
26.	The midterms are scheduled in a way that promotes learning (e.g. before holidays).					
27.	The quizzes are scheduled in a way that promotes learning (e.g. within appropriate class time).					
28.	The teachers are provided with opportunities to make comments about the tests while the test are being prepared.					
29.	The teachers are provided with opportunities to make comments about the tests after the tests have been administered.					
30.	I think the teachers' comments about the tests are taken into consideration in the test development process.					

SECTION III.

The questions below are concerned with the development process of the tests used at Muğla University School of Foreign Languages. Please put a cross (X) in the box which corresponds the most closely to your opinion.

No	Items	Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree
		SD	D	U	A	SA
1.	There is active cooperation between the testing office and the curriculum unit in the test development process.					
2.	All the testers work together while preparing all the tests.					
3.	I think the midterms are carefully planned.					
4.	I think the quizzes are carefully planned.					
5.	The testers have a specific written plan about what the test will be like prior to designing the midterms.					
6.	The testers have a specific written plan about what the test will be like prior to designing the quizzes.					
7.	The testers design the tests with reference to explicit course objectives.					
8.	The testers specify the purpose of each test prior to the test development process.					
9.	The testers take the students' proficiency level into consideration while designing the tests.					
10.	The testers determine together which language skills will be measured in each test according to the course objectives.					
11.	The testers determine the item types according to the test purpose.					
12.	The tests include items taken directly from the course book.					
13.	The testers make use of the items which have been used in previous tests.					
14.	The tests are designed in a way that encourages students to use their language abilities integratively.					
15.	The testers determine the scoring criteria according to the priority of the language skills to be measured.					
16.	The test items are proofread by another colleague who has not written them.					
17.	The existing resources of the institution correspond to the requirements of the tests.					
18.	I would like to work in the test office in the next academic year.					

SECTION IV.

If you have additional comments or opinions, please specify in the boxes below.

The Midterms	Strengths	
	Weaknesses	
The Quizzes	Strengths	
	Weaknesses	
Overall Testing Practices at Muğla University School of Foreign Languages	Strengths	
	Weaknesses	

APPENDIX C

INTERVIEW SCHEDULE (FOR TEACHERS)

1. Are you satisfied with the midterms and quizzes? Why?
2. Do the midterms and quizzes measure the objectives or the abilities necessary for this program? Do the tests reflect or match with what you teach in the class and how you teach, learning activities?
3. Are the teachers given opportunities to make comments about the midterms and quizzes? When do they have chance to make comments? Are these comments taken into consideration by the testers?
4. Would you like to prepare the midterms and quizzes for your classes? If yes, why? If no, do you feel like that you want to be involved in the test preparation process? If yes, why?
5. How do the tests affect your teaching practices? Negatively or positively?
6. Do you think the tests are appropriate to the level of the students? If no, why do you think so?
7. Do the tasks in the midterms and the quizzes reflect the real life language use? If no, do you think that these tasks should reflect the real life language use?
8. Do some of the items in the midterms give clues to the other tasks? What about the quizzes? If yes, what kind of clues? Do you think that these clues affect the performance of the students in the tests?
9. Is it useful to inform the students about the content of the midterms? Do you think that the students should know the content of the quizzes as well? Why?
10. Do you think that lack of oral assessment affect the students' performance in oral skills?
11. Do you have any additional comments?

APPENDIX D

INTERVIEW SCHEDULE FOR TESTERS

1. Is there cooperation between the teachers and the testers?
2. Are the teachers provided with opportunities to make comments on the tests?
When? Do you consider these comments into consideration while preparing the tests? If yes, how?
3. Do you think that the teachers should be involved in the test preparation process?
Why?
4. How do you prepare the tests? With a specific written plan? If there is such a plan, what does it include?
5. Do you prepare the tests with reference to the specific objectives? If yes, are these objectives your teaching objectives or those in the syllabus or course book?
6. Are the tests appropriate for the level of the students? If no, why do you think so?
7. Is it useful to inform the students about the content of the midterms? Do you think that the students should know the content of the quizzes as well? Why?
8. Do some of the items in the midterms and quizzes give clues to the other tasks? If yes, what kind of clues? Do these clues affect the students' performance in the test?
9. How do you select test tasks? Do you use the tasks which reflect real life language use in the midterms and quizzes? If yes, what kind of tasks are these?
10. Do you think lack of oral assessment affect the students' performance in oral skills?
11. Is it difficult to work in the test office? Why?
12. Are you satisfied with the current testing practices in the institution?
13. Do you have any additional comments?

APPENDIX E
INFORMED CONSENT FORM

Dear Interviewee,

You have been asked to participate in a survey study which is intended to investigate the teachers' and the testers' attitudes towards the achievement tests used at Muğla University in School of Foreign Languages. The study also aims at exploring the testers' perceptions of the test development process.

In order to achieve the goals of the study, first you answered a questionnaire, which investigated your attitudes towards the tests. This interview will be the second phase of the study. You are going to be interviewed in order to have deeper insights of your attitudes towards the tests and the overall testing practices at Muğla University in School of Foreign Languages.

Your participation in the interview will bring valuable contribution to the findings of the study. Any information received will be kept confidential and your name will not be released. This study involves no risk to you.

I would like to thank you for your participation and cooperation.

Elif Aydın
MA TEFL Program
Bilkent University

I have read and understood the information given above. I hereby agree to my participation in the study.

Name: _____

Signature: _____

Date: _____

APPENDIX F

SAMPLE INTERVIEW

I: Good morning Banu...

P 10: Good morning...

I: I need to ask you some questions about the tests we use and you prepare in our institution. Would you like to answer them for me?

P 10: Yes, of course.

I: I want to start with cooperation between the testers and the teachers. Do we have or is there cooperation between the teachers and the testers while the tests are being prepared?

P 10: Yes, of course, only among us. I want to have such cooperation in choosing important points for the students. They may check the items or questions...

I: You want to consult them, then? Should the teachers be involved in designing the tests?

P 10: Yes, they should I think. We need other teachers... as I said in choosing the items..

I: When do the teachers make comments on the tests?

P 10: Well, of course after the tests... not in preparing the questions...

I: And... Do you usually consider the teachers' comments about the tests while preparing the tests, designing the tests?

P 10: Yes, of course.

I: Then, how?

P 10: Yes, especially in the way of asking questions... If they say the students do not understand some questions or rubrics... We change in the other tests.

I: Ok... And... This is about designing the tests... How do you prepare the tests? The midterms and quizzes? Do you have a specific written plan?

P 10: Yes we usually make plan... But it is not in details.

I: What does it include? What do you have in that plan?

P 10: Well... Yes, we have a kind of plan. It tells how many points will be given to vocabulary or grammar, for example... And the other parts...

I: You mean the percentages?

P 10: Yes, exactly, we declare the percentages or prior importance of the skills different parts of the tests...

I: Do you use specific objectives, Banu? I mean, do you prepare the tests referring to specific objectives?

P 10: Yes, of course.

I: Are they your own teaching objectives for the students or else?

P 10.. The objectives... Yes... We use course book objectives and the content of the units in the course book. We are dependent on the book, sometimes we can't decide. The course book decides instead of us as teachers... The course book... And we sometimes find it difficult to decide what to ask... It is hard to decide what we ask or what the students should be asked... Sometimes...

P10: Ok, Banu... And are the tests appropriate for the level of the students?

P 10: Yes, it is appropriate. But we try not to ask the questions that are higher for their levels. What we ask is parallel with what they learn, I think.

I: Is it useful to inform the students about the content of the midterms?

P 10: No, it is not useful. The students just focus on specific things according to this information. They do not understand that English is whole not separate points. I think the same for the quizzes as well.

I: Ok... And do some of the items in the midterms give clues to the other tasks? What do you think?

P 10: No, I don't think so. The clues may affect their performance negatively. These are kind of copies. We can't test their abilities appropriately, if they answer one task wrong they will answer another wrong too.

I: What about the quizzes?

P 10: Yes, the same too.

I: How do you select the test items?

P 10: We write and create...

I: About real life language use now... Do you use the tasks which reflect real life language use in the midterms and quizzes?

P 10: Yes, in functional language use. We just test the real life language in, for example, writing dialogues or filling the gaps. We should assess orally and in reading tasks.

I: Ok, we don't assess the students' speaking skills with a formal exam... And do you think lacking speaking exam or lack of oral assessment affect the students' performance in oral skills?

P 10: I think, we can test speaking in our classes. It should not be only one formal exam. They will prepare only for that exam. Then they will forget it. I want all the teachers assess their students' speaking skills in classroom exams. We should stretch to the year.

I: Hum... The last question... Is it difficult to work in the test office?

P 10: Yes, difficult. It takes a lot of time to work there. There is not one level and too much work. You have to decide and prepare quickly the tests. It is heavy load.

I: Are you satisfied with the current testing practices in the institution?

P 10: I am satisfied, but there may be some changes. We need a strong syllabus and explicit objectives. We can see what our students should be able to do in the course and facilitates test preparation process. I would like to work with crowded group. I want the other teachers to be involved in the test development process.

I: Do you have any further comments, opinions?

P 10: No...

I: Thank you for your participation.

P 10: You are welcome always.

I: Thanks...

APPENDIX G

ANALYSIS OF SAMPLE INTERVIEW

FOCUS	P9	P10
<i>Cooperation between the teachers and the testers</i>	In fact, no . Before the tests, they do not know the questions. After the tests we take their comments and take into account these comments.	Yes, only among us. I want to have such cooperation in choosing important points for the students. They may check the items or questions.
<i>Teachers' Involvement in the test preparation process</i>	No need for the others. One of them already does, she is a proofreader of the tests. She checks the questions and makes comments if we make mistakes in the tests. We are five and six with a proofreader is enough.	Yes, they should I think. We need other teachers.
<i>Teachers' comments</i>	Yes, of course.	Yes, we consider the comments. Especially the way we ask questions, the level of students, and rubrics, whether the students may understand the rubrics for next tests.
<i>A specific written plan, and its components</i>	Yes, we have a specific plan. It includes percentages of the language skills. It changes for every test. We decide what we expect from the students. We use other course books, the items we used in previous years. We have answer key and scoring criteria beforehand.	Yes, we have a kind of plan. We declare the percentages or prior importance of the skills.
<i>Reference to specific objectives</i>	Yes , what we want our students to be able to do as teachers determines what will be asked. We make use of the curriculum requirements and the course book objectives.	We use course book objectives and the content of the units in the course book. We are dependent on the book, sometimes we can't decide. The course book decides instead of us as teachers.
<i>Appropriacy of the tests for the students' level</i>	In our institution, the level we teach is higher than their actual level. The tests match with what we teach but.	Yes, it is appropriate. But we try not to ask the questions that are higher for their levels. What we ask is parallel with what they learn.
<i>To inform the students about the content of the tests</i>	Yes, it is useful for the students. It reduces anxiety and stress. They know beforehand what to study and what they should know or able to do in the language at that level....	No, it is not useful. The students just focus on specific things according to this information. They do not understand that English is whole not separate points. I think the same for the quizzes

APPENDIX H

ANALYSIS OF OPEN-ENDED QUESTIONS IN QUESTIONNAIRES

SECTION III TEACHERS' QUESTIONNAIRE

<i>Midterms</i>	<i>Strengths</i>		<i>Weaknesses</i>	
	Student friendly	1	Too long	2
	Periodic checks on the students' progress	1	Reliability and validity should be questioned	3
	Motivating	1	Limited to grammar only	5
	Effective assessment forms	2	Inappropriate to the students' proficiency level	2
	Contextual tasks	1		
<i>Quizzes</i>	<i>Strengths</i>		<i>Weaknesses</i>	
	Periodic checks on the students' progress	2	Hard to decide the students' overall performance	2
	Motivating	2	Inappropriate to the students' proficiency level	1
			Few of them	2
			Too long	2
<i>Overall Testing Practices</i>	<i>Strengths</i>		<i>Weaknesses</i>	
	Strong	1	The lack of oral exam	6
	Organized system	1	Not qualified testers	1
			Training on testing	1

SECTION IV TESTERS' QUESTIONNAIRE

<i>Midterms</i>	<i>Strengths</i>		<i>Weaknesses</i>	
	Integrated	1	The students know the content	1
			Too many of them	1
			Too long time provided for exams	2
			Physical settings of rooms	1
<i>Quizzes</i>	<i>Strengths</i>		<i>Weaknesses</i>	
	---		Unnecessary	1
<i>Overall Testing Practices</i>	<i>Strengths</i>		<i>Weaknesses</i>	
	---		No make up for the final exam	2
			Too much work	1
			Not assessing speaking	2