AN ASSESSMENT OF THE CONTENT VALIDITY OF THE MIDTERM ACHIEVEMENT TESTS ADMINISTERED AT ANADOLU UNIVERSITY FOREIGN LANGUAGES DEPARTMENT

A THESIS PRESENTED BY

HARUN SERPİL

TO THE INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF ARTS
IN TEACHING ENGLISH AS A FOREIGN LANGUAGE

BİLKENT UNIVERSITY

JULY, 2000

PE
1068
.T8
S46
2000

AN ASSESSMENT OF THE CONTENT VALIDITY OF THE MIDTERM ACHIEVEMENT TESTS ADMINISTERED AT ANADOLU UNIVERSITY FOREIGN LANGUAGES DEPARTMENT

A THESIS PRESENTED BY

HARUN SERPİL

TO THE INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF ARTS

IN TEACHING ENGLISH AS A FOREIGN LANGUAGE

BILKENT UNIVERSITY

JULY, 2000

PE
1068
.T8
S46
2000
B053074

ABSTRACT

Title :	An Assessment of the Content Validity of the Midterm Achievement Tests Administered at Anadolu University Foreign Languages Department
Author :	Harun Serpil
Thesis Chairperson :	Dr. Hossein Nassaji Bilkent University, MA TEFL Program
Committee Members :	Dr. Bill Snyder Dr. James C. Stalker Mr. John Hitz Bilkent University, MA TEFL Program

The purpose of this study was to analyze the content validity of the first semester midterm tests in the Foreign Languages Department (FLD) at Anadolu University.

Content validation involves the systematic analysis of test content to determine whether it covers a representative sample of what has been taught in a specific course. This analysis can be done through using multiple sources. Teachers' opinions about the validity of the tests can be elicited, the syllabus content can be compared to the test content and the course objectives can be compared to the test content. Use of these multiple sources ensures the soundness of the analysis and strengthens any arguments to be made about validity.

To investigate the content validity of the midterm tests in the FLD, teachers' perceptions of the tests' representation of the classroom material content were elicited by using questionnaires. All of the course materials at the intermediate level and the two midterm tests from each course were analysed for their content and task

frequencies and the correlations of these were computed using Pearson Product Moment Correlation. To learn the objectives of the courses, the coordinators responsible for each course were interviewed. The data from each of these three sources were then compared in order to assess the content validity of the tests.

The results were conflicting. The instructors of the listening, core course and grammar courses thought their representation of the course content on the test was moderate to high on the whole. Though they were able to justify their exclusion of some items on the tests, which indicated that they actually had some objectives, these were not clearly stated in a written form, a very important point for establishing test validity. The material-test frequency results for these courses indicated that; except for the first grammar midterm test and the strategy areas of the second listening midterm test, the material and tests had low correlations, especially in terms of exercise types. The analysis of the course objectives also showed that they were not specific enough and their overall agreement with the tests' content was low.

If the relevant literature is taken into account, the conflict among the results can be said to have resulted from teachers' not having clear objectives and test specifications in their test design.

Finally, the following suggestions were made to improve the testing procedures in the FLD at Anadolu University: First, course objectives should be clarified and refined. Second, explicit test specifications should be prepared. For further improvement of the validity of the tests, establishing a testing office, which would focus only on testing issues, can also be considered by the FLD administration. Implications for further study were discussed.

BILKENT UNIVERSITY

INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES

MA THESIS EXAMINATION RESULT FORM

July 7, 2000

The examining committee appointed by the Institute of Economics and Social Sciences

for the thesis examination of the MA TEFL student

Harun Serpil

has read the thesis of the student.

The committee has decided that the thesis of the student is satisfactory.

Thesis Title	An Assessment of the Content Validity of the Midterm Achievement Tests at Anadolu University Foreign Languages Department
Thesis Advisor	Dr. Bill Snyder Bilkent University, MA TEFL Program
Committee Members :	Dr. James C. Stalker Bilkent University, MA TEFL Program
	Dr. Hossein Nassaji Bilkent University, MA TEFL Program
	Mr. John Hitz Bilkent University, MA TEFL Program

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

Dr. Bill Snyder
(Advisor)

Dr. James C. Stalker
(Committee Member)

Dr. Hossein Nassaji
(Committee Member)

Mr. John Hitz
(Committee Member)

Approved for the
Institute of Economics and Social Sciences

Ali Karaosmanoğlu
Director
Institute of Economics and Social Sciences

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor Bill Snyder, without whose invaluable guidance and insights this thesis would never come true.

I am also indebted to Dr. James Stalker for bearing with the pita and for his directing us through all the hardships with his affectionate father voice, to John Hitz for his useful help in solving my computer and composition problems throughout the writing process, and Dr. Hossein Nassaji for his guidance in statistical problems in the analysis of the data.

I would like to thank Prof. Dr. Gül Durmuşoğlu, for sowing the seed of this thesis by suggesting a research on testing, and permitting me to attend the MA TEFL program.

I would also like to thank my classmates, without whose continuous help, insights and encouragement I would never be able to survive this program.

For their moral support and endless love, I also want to express my gratitude to my family.

My special thanks go to my fiancée, the ever-loving, ever-caring one...Thank you for your patience!

Not everything that counts can be counted, and not everything that is countable counts.

(A. Einstein)

To Şebnem, who really counts for me.

TABLE OF CONTENTS

LIST OF TABLES x

CHAPTER 1 INTRODUCTION 1
- Introduction to the Study 1
- Background of the Study 5
- Statement of the Problem 8
- Purpose of the Study 8
- Significance of the Study 9
- Research Questions 10

CHAPTER 2 REVIEW OF THE LITERATURE 11
- Movements in Language Testing 11
- Types and Purposes of Tests 17
  - Norm-referenced vs. Criterion-referenced Tests 17
- Reliability 24
- Validity 24
- Reliability vs. Validity 31
- Difficulties in Achievement Test Content Validation 32
- Test Specifications 33

CHAPTER 3 METHODOLOGY 38
- Introduction 38
- Participants 39
- Materials 39
  - Questionnaires 40
  - Course books 41
  - Syllabi 42
  - Tests 42
  - Tape Recordings 42
- Procedures 43
- Data Analysis 45

CHAPTER 4 DATA ANALYSIS 46
- Data Analysis Procedure 46
- Results of the Study 47
  - Analysis of the Questionnaires 47
  - Analysis of the Course Material Content vs. Test Content 63
  - Analysis of the Course Objectives 85

CHAPTER 5 CONCLUSION 88
Overview of the Study 88
Results 88
Discussion 96
Implications 99
Limitations of the Study 101
Further Research 102

REFERENCES 103

APPENDICES 109
Appendix A:
The Questionnaire for the Core-course Teachers 109
Appendix B:
The Questionnaire for the Grammar Teachers 115
Appendix C:
The Questionnaire for the Listening Teachers 119
Appendix D:
Core-course Objectives Transcript 124
Appendix E:
Grammar Objectives Transcript 125
Appendix F:
Listening Objectives Transcript 126

LIST OF TABLES

TABLE		PAGE
1	Core-course Instructors' Answers to the Questions 3, 4, 5, 10, 11, 12	48
2	Core-course Instructors' Answers to the Questions 13, 15, 20, 21, 22	51
3	Grammar Instructors' Answers to the Questions 3, 4, 6, 8, 9, 10, 11	53
4	Grammar Instructors' Answers to the Questions 12, 13, 15, 17, 18, 19, 20	56
5	Listening Instructors' Answers to the Questions 9, 10, 11	58
6	Listening Instructors' Answers to the Questions 12, 13, 14, 15, 20, 21, 22	60
7	Listening Instructors' Answers to the Question 16	62
8	(a) The Correlation Between 1st Midterm Core-Course Material Content with Test Content: Function Areas	66
8	(b) The Correlation Between 1st Midterm Core-Course Material Content with Test Content: Form Areas	67
9	The Correlation Between the Material and the Test Task Formats	68

10 (a) The Correlation Between 2nd Midterm Core-Course Material Content with Test Content: Function Areas..... 70

10 (b) The Correlation Between 2nd Midterm Core-Course Material Content with Test Content: Form Areas 71

11 The Correlation Between the Material and the Test Task Formats 72

12 The Correlation Between 1st Midterm Grammar Material Content with Test Content 74

13 The Correlation Between the Material and the Test Task Formats 75

14 The Correlation Between 2nd Midterm Grammar Material Content with Test Content 77

15 The Correlation Between the Material and the Test Task Formats 78

16 (a) The Correlation Between 1st Midterm Listening Material Content with Test Content 79

16 (b) The Correlation Between 1st Midterm Listening Material Content with Test Content 81

17 (a) The Correlation Between 2nd Midterm Listening Material Content with Test Content 82

17 (b) The Correlation Between 2nd Midterm Listening Material Content with Test Content 83

17 (c) The Correlation Between 2nd Midterm Listening Material Content with Test Content 84

CHAPTER ONE: INTRODUCTION

Introduction to the Study

Testing is an essential part of every educational program and inescapable for language teachers to be involved at one time or another. It is important in every teaching and learning experience, since it is used as a major tool to shape and reshape teaching/learning contexts (Brown, 1996; Hughes, 1989; Weir, 1990). Bachman (1991a) defines a test as, "one type of measurement designed to obtain a specific sample of behavior" (p. 20). So, for example, if you want to measure a student's ability in speaking in a second or foreign language, you must first decide on the specific areas of speaking (e.g. complaining about a problem, making explanations, giving directions, etc), then design your test to measure those specific areas.

The general purpose of language tests is to get information about the learners' language abilities in order to make sound educational decisions (Bachman, 1991; Brown, 1996). Those decisions may be from very formal and crucial ones at the institutional level, like accepting or not accepting a student to a program to not so formal, classroom-level decisions such as making changes on a specific course syllabus. Thus, it is very important to have good tests to be able to make good decisions. "Good language tests help students learn the language by requiring them to study hard, emphasizing course objectives and showing them where they need to improve" (Madsen, 1983, p. 5).

But, testing is a problematic area, because what and how to test requires experience in both teaching and testing, and not many teachers have a firm background in testing. Testing can also be as frustrating for the teachers as it is for

the students, because testing reflects the effectiveness of their teaching (Basanta, 1995). While many teachers may not be comfortable with the idea of testing, testing is central to language teaching (Brown, 1996; Davies, 1990). Teaching and testing are two inseparable parts of a language teacher's task. Properly used, tests can help students improve their learning, and teachers improve their teaching. "[Testing] provides goals for language teaching, and it monitors, for both teachers and learners, success in reaching those goals" (Davies, 1990, p. 1). But what makes a good test so that language teachers can depend on the information it provides for making decisions about their classes? First of all, a good test has to be valid and reliable.

Validity and reliability are the two essential qualities that a test must have to serve the purposes it is intended for. "A test is said to be valid if it measures accurately what it is intended to measure" (Hughes, 1989, p. 22). According to Murphy & Davidshofer (1991), a reliable test is "...one which yields consistent scores when a person takes two alternate forms of the test, or when he or she takes the same test on two or more different occasions" (p. 39). Although to be valid, a test has to be reliable, the reliability of a test does not necessarily indicate that a test is also valid; a test can give the same results time after time, but may not be measuring what it was intended to measure (Brown, 1996; Hughes, 1989). This indicates that validity is central to test construction. If test designers try to increase the reliability of their tests, and have a very reliable but invalid test, then the test can be said to be a reliable measure of something other than what they wish to measure. In order to have more valid test results, some degree of reliability can be sacrificed. Having meaningful test results with some degree of inconsistency is much better than having consistent but meaningless results (Brown, 1996; Davies, 1990; Hughes,

1989; Weir 1990). Testers might, for instance, wish to measure students' ability to use the target language conjunctions in writing, and prepare a test of writing by asking ten multiple-choice questions, including conjunctions and get reliable results, but then the results would not give them any idea of the test-takers' actual writing ability. They might just be testing grammar knowledge of the students, not the actual use of the items in writing. That writing subskill would be more appropriately measured by actually having the students write something requiring to use the specified conjunctions.

To have content validity, the content of a test needs to be a representative sample of the course content. To see how much this is achieved, the test content must be compared against the course content (Davies, 1990; Hughes, 1989). However, just analyzing the course and the test content is not enough. The test must also be closely related to the objectives of the course (Brown, 1996; Heaton, 1988; Hughes, 1989). So, in order to specify the objectives, the course syllabus should be analyzed. Then, upon determining the specific objectives and the content of a particular course, and thus clarifying the target domain, tests can be evaluated for their content validity (Bachman, 1991; Brown, 1996). For achievement tests, the domain is the content of the course and the methodology that has been used in the classroom (Weir, 1995).

Clarifying the target domain and using that domain to prepare questions requires teachers to produce their own test specifications first. That means that they need to have clear statements about " who the test is aimed at, what methods are to be used, how many papers or sections there are, how long the test takes and so on" (Alderson, Clapham & Wall, 1995, p. 10). Having test specifications is also

beneficial for the learners. "If test specifications make clear what students have to be able to do, and with what degree of success, then students will have a clear picture of what they have to achieve" (Hughes, 1989, p. 45). One other purpose for designing specifications is to avoid "hit-or-miss testing" that is usually practised by the teachers who go through the course material to write test items (Hills, 1976, p. 9).

The other two vital qualities a good test must have are authenticity and interactiveness. They are very important and enhancing factors for validity, which provide justification for the use of language tests (Bachman & Palmer, 1996). If a test has been shown to have a high degree of content validity, we can say that it is authentic, or at least has more authenticity than a less valid test. Authenticity is realized when a test requires "...performance...[which] corresponds to language use in specific domains other than the language itself" (Bachman & Palmer, 1996, p. 23). So, if our tests are authentic, we can claim that our students will actually be able to use the language they have learned in real-life situations. It has a positive effect on students too; if they think that the test is authentic, they will try their best. But the fact that tasks should be based on real-life contexts may present difficulties. As Picket (1984, p. 7) puts it; " By being a test, it is a special and formalised event distanced from real life and structured for a particular purpose. By definition, it cannot be real life that it is probing." But even if we admit that real life situations are not fully attainable, we should aim to test our students' ability to perform in specified situations with specified degrees of success (Spolsky, 1968).

Another important quality for a test to have is interactiveness. Bachman and Palmer (1996) define it as: " the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (p. 25). When testers try to

design their tests to have content validity, they have much better chance that "they will involve the test takers' linguistic, topical, affective and metacognitive characteristics in the test tasks" (Bachman and Palmer, 1996, p. 29). If the students are familiar with the test tasks from previous instruction, and if their expectations of the test about the weighting of the areas are met, they will not be hindered by trying to handle totally unfamiliar tasks, but they are more likely to focus on the content of the task and its requirements. As Brown (1994) says: "A classroom test is not the time to introduce brand new tasks because you will not know if student difficulty is a factor of the task itself or of the language you are testing" (p. 271). For example, if a student knows how to order food on the phone from previous instruction, he or she can activate that knowledge on the test to perform on a similar test task. So, as teachers and testers our aim should be trying to mirror our teaching in our testing in such a way that our students can make the most of their knowledge and skills while dealing with the tasks on the test.

Background of the Study

In Turkey, special emphasis is given to the teaching of English at the university level. Since English is the lingua franca in today's world, an ever-increasing number of universities in Turkey are adopting it as their language of instruction. But the majority of the entrants are not proficient enough in English to be able to follow the undergraduate courses, and need English instruction. To that end, most of the universities in Turkey have one or two-year English preparatory programs for their students. Anadolu University is one of these English-medium universities where students have to meet the required degree of proficiency in English to be able to go on to study their undergraduate courses. For the students

who have not been able to pass the initial English proficiency test, Anadolu University requires that those students study English for academic purposes in the Foreign Languages Department (FLD). The FLD is a one-year preparatory program at Anadolu University's Yunusemre Campus. It has an academic staff of about 60 and an annual intake of about 1200 students. The students are from the faculties of Civil Aviation, Humanities and Letters, Communication Sciences, Tourism, Fine Arts, Education, Natural Sciences, and Architecture & Engineering. According to the scores that they receive on the placement test, the students at FLD are grouped into six levels of proficiency: beginner, elementary, low-intermediate, intermediate, upper-intermediate and advanced.

Currently, there is no single, standard syllabus in FLD at the program level, comprehensive of all the courses. Instead, instructors of each separate course (i.e. grammar, core-course, listening, reading and writing), are relatively free to design their own syllabi and the achievement of the students is tested separately for each course (The core-course is an integrated-skills course). The goal of the English preparatory program is " to have students attain a high degree of English proficiency so that they will be able to follow their English-medium undergraduate courses" (Personal Communication, Director of the FLD).

All students in the FLD are given 4 midterm tests in each course (two each semester), an end-of-term oral test, and one final proficiency test over an academic year. The midterm tests, which are administered every 8 weeks, are supposed to be constructed according to the objectives of the course and the items covered up to the tests. Midterm tests are not single tests. Actually each skill is separately tested through its own midterm test. There are a total of five midterm tests: for writing,

reading, listening, grammar and core-course classes. They are not administered on the same day; midterm testing lasts a week. Although speaking has a part in the core-course objectives, it is not tested until the end of the term, as an additional component of the final proficiency test.

To be fair, the students are not asked any questions from the extra materials if those materials are not shared by all the instructors of the same level. The instructors' individual teaching differences in class are not supposed to make a difference because they have to teach the same content and the midterm test is prepared according to that shared content by negotiation among the same course instructors.

At the end of the first semester, if a student has managed to get a mean score of 70 out of 100 on all the midterm tests, s/he is allowed to continue to the next level. So, if the student was at the intermediate level, he or she goes on to the upper-intermediate level. But a student may also skip the next level to go on to the second higher level than his or her level in the first semester, if all six teachers from the first semester agree that the student has the capacity to handle the requirements of a higher level.

At the end of the second semester, whatever the the total mean scores of the students from the midterm tests are, they can take the final proficiency test. They have to get 70 out of 100 from the final proficiency test; otherwise they fail. They also have to score a minimum of 70 out of 100 as the mean score of their total midterm test scores and the final proficiency test score.

Statement of the Problem

Last year, when I was teaching at the FLD, my colleagues and I had problems in designing good test items. We tended to view testing mainly as a tool to give grades to our students. We prepared our tests in a hurry, and usually we included some items just because they were easily available and/or easy to score. Although we had emphasized some items in our teaching, sometimes we had to exclude these items from the test, because we were not able to find suitable testing materials. And occasionally, we introduced some new (and usually difficult) items or task types on the test because we believed that these kinds of items would give us a better idea of our students' achievement (e.g. Teaching all the tenses separately, but requiring the use of all of them together in a given paragraph.) My colleagues I and were dissatisfied with the tests we had prepared, mainly because we felt that there was a loose link between what we taught and tested. Also some students complained that on the midterm tests some areas were emphasized too much, while some other areas were not asked at all despite the attention they were paid in the classroom. Thus, these above mentioned problems have led me to make an analysis of the content validity of the midterm tests in the FLD, with the hope that this would be the first step towards improving our testing.

Purpose of the Study

The purpose of this study is to investigate the extent of the content validity of the first semester midterm tests at the FLD. For better teaching, it is important to learn about a test's content validity because tests which have content validity give better information about the particular study items in which students have difficulty (Davies, 1990; Hughes, 1989; Weir, 1995).

This study might help show ways towards designing more valid tests in terms of content in the FLD. Since the FLD does not have any testing professionals to prepare and/or validate tests, it is necessary for the FLD instructors to look at the validity of the midterm tests they are administering. So, as one of the instructors in the FLD myself, I want to see whether there is a positive relationship between the instructional content and the content of the midterm tests or not. Previously there has been no attempts to investigate the validity of our tests, or any kind of testing-related research in my department. This study will be the first one to probe into testing in the FLD. The instructors' perceptions of their tests representativeness of the course content will be collected. The syllabuses and the material content of each course will be analysed to see to what degree they are represented on the midterm tests. If there is any, additional materials will also be analysed and correlated with the areas covered on the tests, to see how and to what degree they are sampled. The course objectives will be elicited and their match to the tests will be assessed. Through using all these three different sources of data, problems with content validity will be specified and some solutions will be offered.

Significance of the Study

This study is intended to be beneficial for the test designers at Anadolu University Foreign Languages Department (FLD) and for the testers of other university preparatory schools. The results of the study could help the teachers in becoming aware of the issues involved in preparing achievement tests in the FLD. It is also hoped that this study will urge testers to keep the requirements of validity in mind while designing their tests, so that the tests will have better content validity.

Research Questions

This study will address the following research questions:

a) What are the instructors' opinions about the midterm tests' representativeness of the courses' content?

b) To what extent do the sampled areas on the midterm tests and the course content areas correlate?

c) To what degree do the midterm tests' content and the course objectives correlate?

CHAPTER TWO: REVIEW OF LITERATURE

According to Brown's (1995) model of analyzing the elements of language curriculum, testing is central in any language program, because from the placement of the students in the program through their graduation, all kinds of program-related decisions are made using the tests as basis. Testing directly affects teaching, because teachers use the test results to refine their course objectives, use of the materials and the activities used in the class, and if they are the testers, to prepare better tests (Brown, 1995; Davies, 1990). But, though it is so important, most teachers do not know what makes a good test, or the important qualities a test must have (Basanta, 1995; Hughes, 1989).

In this chapter, first I will introduce the basic approaches to testing in the history of language testing, by outlining how the concept of validity developed and gained more importance over reliability. Second, I will try to explain the types and purposes of tests and the essential qualities a language test should have, paying a close attention to validity, and content validity in particular. Third, I will discuss the relationship between validity and reliability. Fourth, I will talk about the difficulties involved in achievement test validation. Finally, I will explain what the test specifications are and why they are important for having valid tests.

Movements in Language Testing

An overall analysis of the history of language testing provided by Spolsky (1995) shows us that there were four main approaches up the present with regards to language testing. First, until the 1950s, there was a *prescientific movement* having its roots in the grammar-translation approach to language teaching. This movement aimed to test the language abilities through translation and free composition tests.

But there were no testing specialists yet, and these tests were far from being objectively scored (Spolsky, 1978). Brown (1996) counts the deficiencies in this early movement as follows:

> There was little concern with the application of statistical techniques such as descriptive statistics, reliability coefficients, validity studies, and so forth...But along with the lack of concern with statistics, came an attendant lack of concern with making fair, consistent, and correct decisions about the lives of the students involved (p. 24).

Language teachers were mostly intuitive about testing. They weighted knowing facts about language as heavily as skill in using the language (Madsen, 1983). For example, students' knowledge about the differences between grammatical structures like "some/any" was emphasized, rather than their actual ability to use them appropriately. The tests were also quite subjective in their task types. Translation, essay, dictation, summary and open-ended answers based on reading comprehension were the main test tasks required from the learners (Madsen, 1983; Spolsky, 1978).

The prescientific movement was replaced by the *psychometric-structuralist* movement in language testing from the early 1950s through the late 1960s (Brown, 1994; Spolsky, 1978). This new movement went from rather subjective to highly objective testing, and it was oriented towards behavioral psychology (Spolsky, 1995). Language was seen as a combination of many separate patterns learned by stimulus-response habit formation, and tests were trying to measure the discrete points taught in the Audio-lingual method (Brown, 1996; Heaton,1988). Mastery of

the language was thought to be assessed in small pieces of language. Separate phonological, grammatical and lexical elements were tested through objective tests (Heaton, 1988; Madsen, 1983). These tests used long lists of unrelated sentences that were incomplete or that contained errors in grammar or usage. Students completed or corrected those sentences by selecting appropriate multiple-choice items (Madsen, 1983). There were new improvements on the testing techniques of prescientific movement; statistical analyses began being used. "Largely because of an interaction between linguists and specialists in psychological and educational measurement, language tests became increasingly scientific, reliable, and precise" (Brown, 1996, p. 24). But there were still problems and improvements to come with respect to validity, for example, authenticity of the test tasks was not a consideration of this movement at all (Spolsky, 1985).

The following movement was *integrative-sociolinguistic*, which was emphasized in the 1970s (Spolsky, 1978). Seeing language as creative, and something more than the sum of its parts, testers started to devise cloze and dictation tests to try to assess students' language ability with extended, multi-skill tests, rather than discrete-point tests (Heaton, 1988). But, while questioning the focus put on the part-by-part grammatical analysis and testing of language by the structuralist movement, its psychometric tools continued to be used by this movement (Brown, 1996; Spolsky, 1978). The testers of this movement tried to respond to the complex and redundant nature of language by trying to test language proficiency through designing questions which involved students' use of different skills like listening and reading together (Ingram, 1985). In real life, for example, we never hear somebody talking in a way that if we don't understand some of the words that are being said,

we cannot understand his or her overall meaning. We simply complete the rest of the speaker's message by using the clues in the immediate context. There are lots of repetitions and synonyms in the usage. With this movement, actual language use outside of the classroom was begun to be considered.

There was a great deal of progress and improvement in the language testing field with these three different movements up to the 1980s. But still a crucial element was missing from the tests of those movements: "the communicative nature of the language" (Brown, 1994, p. 265). So, the next and the last movement to the present is called the *communicative movement* (Spolsky, 1995). This movement is in line with functional-notional and communicative approaches to language learning (Brown, 1994). It emphasizes the total outcome of language learning and learners' total communicative skills. It focuses on what learners can do with the language, the tasks they can carry out, and their productive capacity. "It is an authentic approach to testing in the sense that it focuses directly on total language behavior rather than on its component parts; it assesses that behavior by observing it in real or, at least, realistic language use situations which are as "authentic" as possible" (Ingram, 1985, p. 217).

Quite different from the tests in the past, "such a test might be oriented toward unpredictable data in the same way that real-life interactions between speakers are unpredictable" (Brown, 1996, p. 25). For example, in multiple choice tests, the answers are predetermined, and what the test takers will do is totally predictable. But, in a communicative test, there may be a lot of appropriate answers. On a speaking test consisting of a simulation of a repair problem context, the test-takers can give different responses to the same question asked by a repair shop

assistant, but all of them can be right so long as their answers fulfill the necessary communicative function.

Bachman (1991b) considers the characteristics of the communicative tests:

> Communicative tests can be characterized by their integration of test tasks and content within a given domain of discourse. Finally, communicative tests attempt to measure a much broader range of language abilities – including knowledge of cohesion, functions, and sociolinguistic appropriateness – than did earlier tests, which tended to focus on the formal aspects of language (p. 678).

Thus, communicative tests are contextualized and geared to the needs of the test takers. Here, notice that what a student is required to do in a test situation covers the specified content area that he or she will have to use for his or her language needs in real life situations. For example, if the student is expected to take notes in an academic situation from an English-speaking instructor, this type of task is taught and then asked to be performed on the test.

The sampling of communicative language ability in our tests should be as representative as possible. These samplings should emphasize the skills necessary for successful participation in the specified communicative situations. A test may be claimed to test communicative skills by employing real life tasks, or simulations of those tasks like role-plays, but if those tasks are not what the students will use outside of the classroom, that test may lack validity (Bachman & Palmer, 1996; Weir, 1990).

The recent approaches in language testing value validity more than reliability (Brown, 1996; Ingram, 1985; Madsen, 1983). The former type of tests from the psychometric-structuralist movement emphasized reliability by using discrete-type items like asking the meaning of single words. The greater number of smaller and more refined test items meant more reliability. But, while those tests were more reliable they were not necessarily valid in the communicative sense, because they were structure-based (Brown, 1996; Heaton, 1988).

Today, testers are almost solely concerned with measuring the language skills required to cope with communicative tasks. The evaluation of language use, rather than its form, is emphasized (Madsen, 1983). Translation tests are not often used these days. Objective tests have been used to measure all of the language skills, but now, they are mainly used to evaluate progress only in vocabulary and grammar areas (Brown, 1996; Madsen, 1983).

The traditional testing approach was only interested in the learner's output, but now the testing experts are also interested in the learning process to ensure validity of the tests (Brown, 1996; Madsen, 1983). "Validity is increased by making the test truer to life, in this case more like language in use." (Davies, 1990, p. 34). Also, the content of a communicative test should be totally relevant for a particular group of test-takers, and the test tasks should relate to real-life situations (Bachman & Palmer, 1996; Heaton, 1988). Language testing has shifted its focus from the quantitative tests to qualitative ones to include evalution (e.g. self-assessment, observation, evaluation of courses, materials, projects) (Heaton, 1988). This was due to the developing understanding of the need to value validity more than reliability (Bachman, 1991b; Davies, 1990; Hughes, 1989).

Heaton (1988) suggests a criterion which is in fact very important for a test to have validity: "Perhaps the most important criterion for communicative tests is that they should be based on precise and detailed specifications of the needs of the learners for whom they are constructed" (p. 20). That implies that if testers have well-defined student needs in mind while preparing their tests, they can address those needs by adjusting their test items according to the needs; for example testers can try to weigh a test item like note-taking the most heavily, if note-taking is the greatest need of the students.

A test's validity can only be questioned on the basis of its purpose (Bachman, 1991a; Brown, 1996). There are many purposes tests can be used for in a language program, but all these purposes can be classified under two main test types: norm-referenced or criterion-referenced. It would be beneficial to clarify these two basic test types here before elaborating on validity.

Types and Purposes of Tests

Norm-Referenced vs. Criterion-Referenced Tests

Tests can be categorized into two major groups: norm-referenced tests (NRTs) and criterion-referenced tests (CRTs). These two tests differ in their purposes, their content selection, and the interpretation we make of their results (Bachman, 1991a; Bond, 1996; Brown, 1996; Douglas, 2000; Hughes, 1989).

The main reason for using an NRT is to classify students. "The purpose of a NRT is to spread students out along a continuum of scores so that those with "low" abilities are at one end of the normal distribution, whereas those with "high" abilities are found at the other" (Brown, 1989, p. 68). While NRTs put the students into rank orders, CRTs determine "what test takers can do and what they know, not how they

compare to others" (Anastasi, 1988, p.102). CRTs tell the testers how well students are doing in terms of the predetermined, specific course objectives (Bond, 1996; Brown, 1989; Henning, 1987).

Test content is an important factor for the testers and administrators of a program in choosing between an NRT test and a CRT test. The CRT content is selected according to its importance in the curriculum, while an NRT content is chosen by how well it discriminates among students (Bond, 1996; Brown, 1996). In the process of developing an NRT, easy items are usually excluded, because more difficult items discriminate between high and low proficiency groups better. Thus, NRTs tend to eliminate those items whose content is very important for a content – valid evaluation of students' learning (Bachman, 1989; Brown, 1996). Students should know beforehand about the content and format of a CRT. However, an NRT's specific content is not known by the students until they take it. They can only learn about its general question design (e.g. multiple-choice, true-false, etc.) and some technical aspects like time duration (Brown, 1996).

While NRTs measure general language proficiency, CRTs aim to measure the degree to which students have developed skills on specific instructional objectives (Hudson & Lynch, 1984). "Often these objectives are unique to a particular program and serve as the basis for the curriculum. Hence, it is important that teachers and students know exactly what those objectives are so that appropriate time and attention can be focused on teaching and learning them" (Brown, 1989, p. 68). Bachman (1989) advises trying to have information about learners' achievement of instructional objectives as "precise" as possible, so that the necessary revision and improvement of the language program can be carried out (p. 247).

While NRT scores give little specific information about what a student actually knows or can do, CRTs give detailed information about how well a student has performed on each of the objectives included on the test (Bond, 1996; Carey, 1988; Henning, 1987). The CRT gives both the student and the teacher much more information about the accomplishment of goals than a NRT "as long as the content of the test matches the content that is considered important to learn" (Bond, 1996, p. 2).

In discussing the use of both criterion- and norm-referenced language measurement techniques, Brown (1989), and Lynch and Davidson (1994) advise strengthening the relationship between testing and the curriculum. When teachers use and develop their own CRTs, they get twofold benefits, because they can develop not only more informative tests, but also clearer course objectives (Henning, 1987). Having clear objectives is very important. "The usefulness of CRT for teaching rests in the degree to which the behavior or ability being tested is clearly defined" (Lynch & Davidson, 1994, p. 729). Basing the content of the test on specific objectives also avoids the mismatch between teaching and test content that is often found with norm-referenced tests (Bachman, 1989).

However, specifying clear objectives on one occasion is not enough. Henning (1987) calls also for constantly revising the objectives if language teachers want their tests to be useful. When teachers continuously revise their curricular objectives and test objectives, and keep the relationship between the two as close as possible, they ensure the content validity of their tests to a great extent (Bachman, 1991a; Hughes, 1989).

Since the validity of a test hinges on its purpose as well as its type and nothing can be said about a test's validity before knowing them, it might be useful to

expand on these essential features of tests here. Language tests are used for the purposes of determining proficiency, placement, diagnosis, aptitude, and achievement (Bachman,1991a; Brown, 1996; Davies, 1990; Henning, 1987; Hughes 1989).

The content of a proficiency test is not related to the content or objectives of a language program (Brown, 1996). Thus, these tests are usually designed as NRTs. Hughes (1989) explains proficiency tests as: "... designed to measure people's ability in a language regardless of any training they may have had in that language... it is based on a specification of what candidates have to be able to do in language in order to be considered proficient" (p. 9).

Placement tests are administered to place students in a program or at a certain level. Placement tests must be constructed according to the particular language skills at different levels in a program (Brown, 1996; Heaton, 1990; Hughes, 1989). Sometimes a test called a proficiency test is applied for both determining the proficiency of students and placing them into proper level of the course. A general proficiency test may be used as a placement instrument to place learners at appropriate proficiency levels, that is, from beginners through advanced (Brown, 1996).

A diagnostic test is administered to determine students' strengths and weaknesses in the specific target domains. They are developed to find out what kind of further teaching is necessary, and also can be used for syllabus or curriculum revision (Bachman, 1991a, Brown, 1996, Heaton, 1988, Hughes, 1989). Any language test has some potential for providing diagnostic data. A placement test can be regarded as a general diagnostic test in that it distinguishes relatively weak from

relatively strong learners (Brown, 1996). Achievement and proficiency tests are also often used for diagnostic purposes (Bachman, 1991a; Heaton, 1988).

Aptitude tests are given to a person before he or she starts to be taught a second language. A foreign language aptitude test measures a person's capacity or general ability in learning a foreign language. "Aptitude tests are independent of a particular language, they aim to predict success in the acquisition of any foreign language." (Brown, 1994, p. 259). Since there is no syllabus for teaching aptitude, aptitude tests have no specific content to draw on (Davies, 1990).

Achievement tests "will typically be administered at the end of a course, to determine how effectively students have mastered the instructional objectives." (Brown, 1996, p.14). Achievement tests are criterion-referenced; they are used to measure specific language points based on course objectives, and their purpose is to assess the amount of material learned by each student (Brown, 1995; Jordan, 1997). So, an achievement test's purpose and content is very different from that of a proficiency test. While a proficiency test is very general about what it expects from the test-takers, an achievement test has very specific content and specified expectations from students according to that content. A proficiency test will focus on the global performance, but an achievement test is employed by teachers to see how successful students are on specific skills, say, writing an argumentative essay.

According to Hughes (1989), there are two types of achievement tests: final and progress. "Final achievement tests are administered at the end of a course of study and their content must represent the content of the courses which they are based upon. Progress achievement tests are administered to measure the progress learners are making" (p.10-11). Progress achievement tests can also be used as

diagnostic tests to pinpoint students' weak and strong points in learning. Achievement tests are a very important part of a language program. "If we assume that a well-planned course should measure the extent to which students have fulfilled course objectives, then achievement tests are a central part of the learning process" (Basanta, 1995, p. 56).

Midterm tests in the FLD can be regarded as progress achievement tests, since their aim is to find out and assign marks to how well the students have learned up to the point at the test. To be able to follow the progress in the learners' language clearly, Hughes (1989) suggests determining clear and detailed short-term objectives and administering tests based on those short-term objectives. He recommends testers to base their test content on the objectives of the course rather than on the detailed course syllabus, since the objectives provide much more accurate information on achievement. If a test is based on the content of a poor or inappropriate course, the students taking it may be misled about their achievement. But if the test is based on objectives, the information it gives will be more useful, and the course will not stay in its unsatisfactory form. The long-term interests of the students are best served by achievement tests whose content is based on course objectives (Heaton, 1988; Hughes, 1989). Basing the test content directly on the objectives of the course requires course designers to be "explicit about objectives" (Hughes, 1989, p.11). Thus, having specific objectives available as their criteria, evaluators are able to find out how well students have achieved those objectives (Hughes, 1989). In addition to following clearly defined course objectives, achivement tests should also closely follow the teaching that occurred up to the point of their administration (Heaton, 1990).

Heaton (1990) claims that achievement tests need to reflect the classroom teaching that takes place before they are administered: "Achievement tests should attempt to cover as much of the syllabus as possible" (p. 14). Alderson and Clapham (1995) also argue that unlike exercises, since the student works alone in the test, the test items on achievement tests should provide all the necessary information that the student needs to answer the question:

> For course-related, short-term achievement tests, as with exercises, teachers need to write items that correspond with their teaching methods, reflect their teaching objectives, and mirror in some way their teaching materials. They also need to make sure that such items are not ambiguous and that they produce the expected type of response (p. 185).

Achievement tests are useful in that they provide accurate information about students' learning and help teachers make decisions about the necessary changes to their syllabi (Childs, 1989). But doing that requires the specification and clarification of the instructional objectives first. Then, teachers can learn about their students' abilities, needs, and achievement of the course objectives, using the test results (Brown, 1996; Weir, 1995). Achievement tests help determine which objectives have been met and where changes might have to be made, "an important contribution to curriculum improvement" (Heaton, 1988, p. 172).

Whatever the type of the test is, it must have two vital qualities for its results to be usable by its designers: reliability and validity. These test features will be explained below.

Reliability

A test is said to be reliable if its results are "consistent across different times, test forms, raters, and other characteristics of the measurement context" (Bachman, 1991a, p. 24). This consistency may be in terms of multiple test administrations over time (stability), between two or more forms of the same test (equivalence), or within one test (internal consistency) (Hudson & Lynch, 1984; Murphy & Davidshofer, 1991). For example, if a student scores 60% on a reading test, and one week later different raters grade the same student's reading proficiency as 62% on a similar reading test, these reading tests can be said to be highly reliable. Language testing experts agree that the longer a test is, the more likely that it is reliable (Bachman, 1991a; Brown, 1996; Henning, 1987).

Providing reliability in language tests is difficult because language learning is not a static process, and learning and proficiency involves many ever-changing variables, it is not as simple as measuring the heights or weights of people (Brown, 1996). Some possible sources causing a lower test reliability are measurement errors like fatigue, nervousness, inadequate content sampling, answering mistakes, misinterpreting instructions, unclear or ambiguous questions and guessing (Hughes, 1989; Rudner, 1994). A test must be reliable to be valid, but being reliable in itself does not mean that the test is also valid (Bachman, 1991; Brown, 1996; Hughes, 1989). "There is nothing more important to be demanded of a test than validity; if a test is not valid, it is worthless" (Hills, 1976, p. 11). But what is validity?

Validity

In The Encyclopedic Dictionary of Applied Linguistics (Johnson & Johnson, 1998), the validity of language tests is defined as: " the extent to which the results

truly represent the quality being measured" (p. 363). Questions related to a test's validity are not to be answered simply by *yes* or *no*. Rather, validity is a matter of degree (Stevenson, 1981). Douglas (2000) likens the validation process to a "mosaic":

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

Validity is central to language test construction and there are five types of validity: face, concurrent, predictive, construct, and content. The most important types of validity are considered to be construct and content (Bachman, 1991; Davies, 1990; Hills, 1976). The combination of construct and content validity gives very strong evidence for the validity of the test (Murphy & Davidshofer, 1991). Davies (1990) says: " In the end no empirical study can improve the test's validity – that is a matter for the content and construct validities" (p. 36). He asserts that valid language tests depend on testers' criteria of the language proficiency and their language knowledge.

The most superficial type of validity is face validity. "A test is said to have face validity if it looks as if it measures what it is supposed to measure" (Hughes, 1989, p. 27). For example, a test which was intended to measure pronunciation

ability but which did not require the test-taker to speak might be thought to lack face validity by the test-takers. If a test is not viewed as having validity by the students taking the test, the biggest disadvantage is that the students' negative reaction towards the test might mean that they may not perform in a way which would truly represent their ability (Hughes, 1989; Weir, 1990).

Davies (1990) suggests that if possible, a test should contain face validity, but, "construct and content validities should not be sacrificed for the sake of an increased lay acceptance of the test" (p. 23). Instead, he advises testers to sacrifice face validity, it being the least important type of validity.

Concurrent validity is established when the scores of a given test are found to be consistent with those of some other test, which have the same criteria, administered at about the same time (Alderson et al, 1995; Hughes, 1989).

Predictive validity concerns the degree to which a test can predict candidates' future performance (Bachman, 1991a; Hughes, 1989). Both concurrent and predictive validities are criterion-related (Brown, 1996). For predictive validity, criteria are the objectives of a course, and if a placement test is successful in predicting a student's success on the final test at the end of term with the same objectives, it can be said to have predictive validity (Brown, 1996; Hughes, 1989). If a test is successful in predicting learners' future language performance, that does not mean that it is also valid in measuring the learners' language abilities, because, like concurrent validity, it only deals with the relationships among test scores (Hamp-Lyons, 1996). The scores alone do not provide enough information to decide about students' competence to perform in given language use situations. It requires another type of validation, that is, construct validation (Bachman, 1991a).

Construct validity, which is considered to be the most important type of validity by researchers like Bachman and Palmer (1996), and McNamara (1996), "concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs" (Bachman, 1991a, p. 225). The term construct refers to "the particular kind of knowledge or ability that a test is designed to measure" (Read, 2000, p. 95). If we are trying to measure a particular ability through a particular test, then the test will have construct validity only if we are able to show that we are actually measuring *just* that ability (Bachman,1991a; Hughes, 1989). For example, a reading test may be claimed to test *only* the reading ability or construct, but in fact, it is very hard to justify that; that type of test would also involve other abilities like recognizing vocabulary or inferencing.

When teachers define their objectives and thus their theoretical standpoints and justifications for choosing a particular type of teaching, in a way, they define their language "constructs", so they know what to expect from their students' performance (Bachman, 1991a). And when they teach and measure those constructs, sticking to their objectives, their tests have a good degree of construct validity. "The closer the relationship between the test and the teaching that precedes it, the more the test is likely to have construct validity" (Weir, 1990, p. 27). But it is not so easy to clearly define and apply construct validity in classroom based evaluation.

Since language teachers need specific data to interpret their students' progress, and accordingly to refine their teaching, it is not practical for them to rely on such an elusive concept as construct. So, what language teachers come down to is what they have as concrete classroom materials they use in their teaching, on the

basis of which they try to measure and observe student behaviors. Having concrete answers from the students taking an achievement test may be the only way available for the teachers to evaluate student ability.

Teachers use their teaching content as the source of their tests and basis of the reflection of learning, to measure the observable language development of their students, instead of attempting to measure the unobservable constructs in their students' brains. In fact, the distinction between construct and content validity in language testing is not very clear, and many researchers agree that most content categories are constructs (Landy, 1987; Messick, 1988; Teasdale, 1996; Weir, 1990). Weir (1990) explains why these two concepts are hard to distinguish and how they overlap:

> Because we lack an adequate theory of language in use, apriori attempts to determine the construct validity of tests involve us in matters which relate more evidently to content validity...We can often only talk about the communicative construct in descriptive terms and, as a result, we become involved in questions of content relevance and content coverage (p. 24).

Every test has a particular content, a reflection of what has been taught or what is supposed to have been learned. To determine the content validity of a language test, the test's content should be examined to see if it covers a representative and proportional sample of what has been taught in a particular course and if the test content is in line with the predetermined course objectives and test specifications (Anastasi, 1988; Bachman, 1991a; Brown, 1996; Heaton, 1990; Hughes, 1989).

Bachman (1991a) identifies two aspects of content validity: content relevance and content coverage. Content relevance involves the specification of both the ability domain (defining constructs) and test method facets (measurement procedure). Content coverage is the extent to which the tasks required in the test adequately represent the behavioral domain, or language use tasks, in question. The specifications should be clear and detailed about what it is the test measures, what will be presented to the test taker, and the answers expected from the test-takers. Content coverage of a test is harder to establish than the content relevance, because it requires an "adequate analysis" of students' language needs and a "descriptive framework" of language use (Bachman, 1991a, p. 311). For example, in order to cover the content area for, say, listening skill, the listening teacher should know about all the listening content areas used in language situations in real life, and sample those areas in her teaching accordingly. But, in its ideal form, this is an almost impossible task to accomplish.

The problem of not having a domain definition that clearly identifies the language use tasks from which possible test tasks can be sampled is addressed by Bachman (1991a). He states that "if this is the case, demonstrating either content relevance or content coverage becomes difficult" (p. 245). Suppose, for example, that a language teacher is teaching a specific skill like reading. If the teacher does not have a clearly defined set of reading tasks for the students to be able to perform as an outcome of his or her course, teaching becomes haphazard, and there is no specified domain of reading to talk about. Obviously, he or she will not have any predetermined tasks to sample on the reading test, so having students only to answer true-false questions would do. If the course requires the students only to read, the

content would just be "reading." There would be no content to relate or cover on the test. So, anything related to reading would be possible to use as a measurement of the reading ability of the students. Of course this would be an extreme example, and the least desired type of teaching and testing (Bachman, 1991a; Bachman & Palmer, 1996).

Hughes (1989) stresses that content validity is important because the greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure. In classroom achievement testing the starting point must be to make our tests reflect as closely as possible what and how our students have been taught (Basanta, 1995, Bejar, 1983; Weir, 1995).

In order to investigate the content validity of language tests, "the specification of the skills and structures that is meant to cover should be examined" (Weir, 1990, p. 25). Alderson (1988) asserts that the validity of an achievement test depends on the extent to which it is an adequate sample of the syllabus, and if it covers the syllabus "adequately and accurately", it is likely to be a valid test (p. 16).

Alderson et al (1995) suggest some ways to investigate content validity:

- Comparing test content with specifications/syllabus
- Questionnaires to, and interviews with 'experts' such as teachers, subject specialists, applied linguists (p. 193).

In this study, I will use these methods mentioned above to analyze the validity of the midterm tests in the FLD. First, I will compare the test content with the syllabus content for each course, which is also suggested by Hughes (1989), then I will use questionnaires to elicit the teachers' opinions of validity, and interview the coordinators about the course objectives.

Reliability vs. Validity

Though the ideal situation for a test is to have both reliability and validity, it may not be an easy task for language teachers to achieve this, especially when they want to emphasize communicative abilities of their students. Underhill (1982) points out the inevitable conflict between reliability and validity in language tests: "The main problem...may be stated simply: high reliability and high validity are seemingly incompatible...If you believe that real language occurs in creative communication between two or more parties with genuine reasons for communicating, then you may accept that the trade-off between reliability and validity is unavoidable" (p. 17).

Along a similar line, Davies (1988) brings up the same problem in terms of sampling of the skills: " If we accept the more common sense position, that sampling can increase (or decrease) reliability, then we must recognize that reliability and validity are often at odds with one another" (p. 30). For example, essay writing, which is a direct testing of the writing skill, can be said to measure what it is intended to, so it is a valid way of testing writing ability, but it may be very difficult to establish its reliability because it is difficult to score objectively. Some testing experts assert that reliability may not be as important as validity for achievement tests: "Because an end-of-unit or end-of-week test is usually much less important for students than is an admissions test, its consequences being less dire, test reliability is less important" (Alderson & Clapham, 1995, p. 185). Teachers are more interested in having rich information from their tests about their students' progress than having consistent results.

Difficulties in Achievement Test Content Validation

Validating classroom tests is not an easy task. It presents many problems for validators that may be difficult to overcome. Major problems include absence of test specifications, practical constraints in covering and tracking all of the course content, unavailability of clear course objectives, and some affective factors like unwillingness to cooperate with the validator (Alderson & Clapham, 1995; Alderson et al, 1995; Bachman & Palmer, 1996; Hughes, 1989).

The biggest problem for classroom tests is that it is usually not possible to find carefully prepared test specifications and well-defined course objectives to compare the tests against (Brown, 1996). If specifications are not available to the test validator, he or she cannot know about the criteria of the testers in sampling what has been taught in the class, their testing objectives and expectations of performance from the students (Alderson et al, 1995; Hughes, 1989). Of course it may not be possible to include everything taught in class on the test, but the validator needs to know the criteria for the selection of a specific content area to test over another area. These criteria must be made explicit through preparing test specifications (Alderson et al, 1995; Hudson & Lynch, 1984).

Having clearly stated objectives behind the test is also very important, because the same test content can be used for a variety of different purposes (Bachman, 1991a; Brown, 1996). If the objectives are not available in a written form, as an outsider, the validator may not exactly understand the class teacher's rationale for preferring to ask one thing on the test over another. Unfortunately, most teachers lack such clearly specified objectives to refer to, and they show a tendency to test what is easiest to test (Davies, 1990; Hills, 1976; Hughes, 1989). Validation

requires comparison, and if the validator is not able to see both sides of the coin, that is, the specifications and the test, comparison is of course not possible.

A classroom context consists of very dynamic processes and its testing does not easily lend itself to an assessment of validity. To do a content validation of a classroom test, the validator has to take into account everything that has happened in that particular class as input to the students (Hughes, 1989). Students may learn some additional language items out of school and this is very hard to include in the analysis of the course content.

And sometimes – as in my case – several classrooms take the same achievement test. That brings up another problem: "How can the validator be sure that different teachers closely followed the same syllabus in their teaching, when comparing the classroom content with the test content?" If they did not, there would also be the problem of fairness, in addition to a lack of validity (Brown, 1996; Hudson & Lynch, 1984; Hughes, 1989). It is also difficult to be sure if there were some extra points that were taught as they came up during class period, like unfamiliar vocabulary, and how frequent they were. Perhaps some might suggest video-recording classes for later analysis, but it would just be impractical. In addition, teachers may feel like their testing or teaching abilities are being questioned by the validation process, and they may not be so willing to cooperate with the validator (Basanta, 1995).

Test Specifications

Previously I mentioned the importance of linking our testing to what we teach in the classroom. A classroom test should be closely related to the content covered in the class teaching (Davies, 1990; Hughes, 1989). Heaton (1988) suggests

that this relationship can be built by relating the different areas covered in the test to the length of time spent on teaching those areas in class. Students need to have been taught the necessary skills or knowledge up to the test to be able to answer all the items on it. "The insistence that every item be answerable as a result of instruction has naturally led to the notion of a "universe" of items. To define such a universe, it is necessary first to identify all its constituent elements" (Bejar, 1983, p. 9). Identifying these elements obliges teachers to define their teaching content and objectives. In some course books, there are guidelines or even some specifications for testing. But, "Even if the specifications were done by the textbook writer, the teacher will have to select what s/he considers most important, and not what is easiest to test, in order to draw up a set of specifications which reflects the emphasis of the teaching" (Basanta, 1995, p. 56). Though designing specifications is very important, in practice teachers seem not to weigh this part of teaching heavily enough. And when they do, they need to be careful about the danger of "concentrating too much on testing those areas and skills which most easily lend themselves to being tested" (Heaton, 1988, p. 13). Basanta (1995) points out the current situation in that respect, which seems to be the case also in my institution:

> In most classrooms today, English is taught through communicative textbooks that provide neither accompanying tests nor any guidance for test construction. Teachers are on their own in constructing tests to measure student progress and performance. The result is they write traditional grammar-based items in a discrete-point format that does not fit the

communicative orientation of the textbook or the underlying teaching principles (p. 56).

Alderson et al (1995) provide a list of guideline questions to ask when designing test specifications :

1. What is the purpose of the test?
2. What sort of learner will be taking the test?
3. How many sections/papers should the test have, how long should they be and how will they be differentiated?
4. What target language situation is envisaged for the test, and is this to be simulated in some way in the test content and method?
5. What text types should be chosen-writen and/or spoken? What should be the sources of these, the supposed audience, the topics, the degree of authenticity? How difficult or long should they be? What functions should be embodied in the texts-persuasion, definition, summarising, etc? How complex should the language be?
6. What language skills should be tested? Are enabling/micro skills specified, and should items be designed to test these individually or some integrated fashion? Are distinctions made between items testing main idea, specific detail, inference?
7. What language elements should be tested? Is there a list of grammatical structures/features to be included? Is the lexis specified in some way-frequency lists etc.? Are notions and functions, speech acts or pragmatic features specified?

8. What sort of tasks are required – discrete point, integrative, simulated 'authentic', objectively assessable?
9. How many items are required for each section? What is the relative weight for each item – equal weighting, extra weighting for more difficult items?
10. What test methods are to be used – multiple choice, gap filling, matching, transformation, short answer question, picture description, role play with cue cards, essay, structured writing?
11. What rubrics are to be used as instructions for the candidates? Will examples be required to help candidates know what is expected? Should the criteria by which candidates will be assessed be included in the rubric?
12. Which criteria will be used for assessment by markers? How important is accuracy, appropriacy, spelling, length of utterance/script, etc.? (p. 11-13).

These questions above are just guidelines and teachers, who make their own tests, can adapt them according to their needs. They can add more details or skip some details which they may find unnecessary to address in their own teaching/testing contexts. It is not to be expected that everything in the specification will always appear in the test, there may be too many things to be included on a single test. But having these specifications will "provide the test constructor with the basis for making a principled selection of elements for inclusion in the test" (Hughes, 1989, p. 22). As can be easily noted, test specifications need to be as clear and precise as possible, like the questions above require.

In this chapter, the history of language testing was briefly outlined, by presenting how it came to emphasize qualitative measurements, and thus validity more than objectivity and reliability. Then, types and purposes of tests were introduced and explained to clarify the place of criterion-referenced achievement testing in the overall picture of language testing. After that, two vital qualities of any language test, i.e. reliability and validity were explained with special focus on content validity. Other types of validity were also introduced in order to clarify the reason for only focusing on content validity in this study. The problems involved in achievement test content validation were discussed, where the major problems seem to lie in the absence of clear course objectives and test specifications. Finally, test specifications have been elaborated and some guidelines on preparing them were given.

The next chapter will give information about how the data were collected and analyzed for this study.

CHAPTER THREE: METHODOLOGY

Introduction

The concern of this study is to examine the degree of content validity of the first semester midterm achievement tests administered at Anadolu University Foreign Languages Department (AUFLD). In this study, the main research questions were:

1) What are the instructors' opinions about the midterm tests' representativeness of the courses' content?
2) To what extent do the sampled areas on the midterm tests and the course content areas correlate?
3) To what degree do the midterm tests' content and the course objectives correlate?

To answer the first question, a questionnaire related to the content validity of first and second midterm tests was given to the intermediate level English language instructors of the FLD. There are four classes each having about 25 students at the intermediate level. The intermediate level was chosen for the study because the material content was thought to reflect an average representation of all other levels.

To answer the second question, the frequency of the items in the content of the materials was compared with the frequency of the test content items in the midterm tests. An *item* is the smallest unit in the content of a test, requiring a specific task to be performed (Brown, 1996). For the grammar course, for instance, items refer to the appropriate use of grammar components such as pronouns, articles and passive structures. For the purposes of this study, *content* is based on the course books and materials used in each class. In the following sections, participants are

introduced, then the materials are explained, followed by procedures and data analysis.

To answer the third question, three course coordinators were interviewed to learn about the objectives of each course and these specified objectives were then related to other data and evaluated by taking the test contents of the respective courses into account.

Participants

10 intermediate level instructors participated in the study. 3 of these instructors were also course coordinators whom I interviewed for their perceptions of the first semester course objectives. Four of the instructors taught core-course (an integrated-skills course), three of them were grammar instructors, and three taught listening. The core-course, grammar and listening instructors' experience in teaching varies between 1 to 15 years.

There were 2 more instructors at this level, one for the reading course and the other for the writing, and they were also given questionnaires, but their questionnaire results had to be discarded when I realized that the data to analyze their courses were too scarce to justify an analysis.

Materials

For this study, questionnaires, the syllabuses for the core-course, grammar and listening courses, the course books for the core-course and listening, handouts for the grammar course, and six midterm tests were used. The course book for the grammar was not analyzed, because it was optional for the students to work on and not used in the classroom. I checked with the instructors to ensure that there was no extra material used in addition to the material analyzed in this study. Tests were

obtained from the Materials Development Office at the university, and the course materials were obtained from the instructors and the course-coordinators. Additionally, tape recordings of the interviews with the coordinators about the course objectives were transcribed.

Questionnaires

There were a total of five different questionnaires for five courses. The questionnaires were developed for the intermediate-level English course instructors at the FLD at Anadolu University. The questionnaires for the reading and writing courses had to be eliminated from the study because of the inadequacy of the available data. Thus, only the questionnaires for the core-course, grammar, and listening courses were analyzed in the study (see Appendices A, B, C).

The purpose of the questionnaires was to seek information about what the instructors thought about the representation of their teaching on their tests. To that end, the instructors were asked questions about the inclusion and weighting of the specific content of their teaching on their midterm tests. The basic outline of the questionnaires was the same: They contained mixed question types. There were three sections: In the first section, there were two questions about the instructors' teaching experience. The second section included questions about the first midterm test content, and the questions of the third section were about the second midterm test content. These questions were intended to elicit opinions of the instructors about the test representation of the given list of detailed course areas.

The questions about the specific content areas were designed by reading through the material and syllabus of each course. Also, the content areas

pinpointed from the materials were checked with the respective coordinators to see if those areas had actually been covered in class. Some of the areas were found not to have been covered because of time limitations, and were not included in the questionnaires.

In addition, there were three questions for each midterm test asking the instructors' assessments of their tests' overall representation of the course objectives, content, and exercise types they used in their classes. To get richer data, the participants were also encouraged to write their additional comments or opinions about the subject, if they had any.

Course Books

Two course books were analysed in terms of their specific content units: *Spectrum 4 - Student's Book* (Warshawsky & Byrd, 1994) was the course book for the core-course and was analysed for the language functions and forms covered during the first semester. *Spectrum 4 - Student's Book* is an integrated skills course book addressing six levels from beginners to upper-intermediate. It is accompanied by a teacher's book, a workbook, and audio cassettes. The intermediate level core-course book, *Spectrum 4,* was taken as the basis for determining the intermediate level core-course content items and item frequencies.

Listening 2 (Doff & Becket, 1991), which was used as the course book after the first listening midterm test was administered, was analysed for the listening course. It has 20 units, all aiming towards developing students' speaking abilities as well as listening. It addresses intermediate-level learners, and has a variety of listening topics based on a notional-functional approach. *Focus on Grammar* (Fuchs & Westheimer & Bonner, 1994) was originally intended to be analyzed for the

content coverage, but upon personal communication with the grammar course instructors, it was found that this book was not used in teaching at all. It was only used for home study as a supplementary tool, and since I was not sure about the content of those homework assignments, I did not include this book in my study. Instead, the material which consisted of the handouts used in the classroom to teach grammar was obtained from the grammar coordinator, and it was used to pinpoint the grammar topics covered during the first semester grammar course instruction. The items under each grammar category were listed and their frequencies were calculated.

Syllabi

Syllabi of the three courses; the core-course, grammar and listening were obtained from the coordinators. They included an outline of what teaching content was to be covered when. In all of the syllabi, the time allotted for the different units covered in each course was found out to be equally distributed.

Tests

There were a total of three first midterm tests and three second midterm tests, two total midterm tests for each course. The frequency of the item content on these tests were compared with the frequency of the teaching items of the courses' contents. The frequencies of the specified exercise types in the course material was also compared with the frequencies of the task formats of the test items.

Tape-recordings

To triangulate what I had found by analysing the contents of the course materials to be the course content that was aimed to be taught, I consulted the opinions of the course coordinators about the objectives of the courses that they were

responsible for (Coordinators preside over the weekly course meetings and try to control and balance the quality and quantity of instruction among the different classes at the same level). Since the objectives were not documented or specified in any written form, they asked for some time to first think about it. They said ten or fifteeen minutes would be enough to make up their list of objectives, so I allowed for fifteen minutes for them to think (During that period, they all were observed to refer to their course books and syllabuses in order to organize their thoughts). Then, when they said they were ready, I started the recording of their accounts of the objectives. All of the five coordinators were recorded. The recordings of these coordinators' talks were then transcribed and analysed to specify objectives (see Appendices D, E, and F). Objectives expressed for each course were then commented on by taking the course material content and the test content into account.

The reading and writing transcriptions were not included in the study, because for these courses, there was no enough course content information to legitimately compare the transcribed objectives with.

Procedures

In the last week of March 2000, after getting permission from the administrators of the FLD, and contacting the core-course, grammar, listening, reading and writing coordinators, their syllabi, course books and classroom handouts (for the grammar), and the supplementary materials (like listening cassettes) were obtained. The first semester midterm tests, which had already been administered, were collected through contacting the Materials Development Office.

The English instructors' opinions and suggestions related to the content validity of achievement tests were sampled through questionnaires. The questions of

the questionnaires were formed by taking the previously obtained course materials as the basis for each specified subskill area.

After the questionnaires were developed, they were piloted at the FLD, a week before the actual administration. The core-course and grammar questionnaires were piloted with two instructors teaching core-course and grammar during the second semester. They were chosen from among the instructors for whom the questionnaires were not actually intended, in order not to skew the questionnaire results. Then, the necessary changes were made, which consisted mostly of the clarification and simplification of the instructions. The actual administration took place in the following week, the second week of April.

On April 15th, the revised final versions of the questionnaires were administered. The participants were informed about the nature of the study by the researcher before distribution of the questionnaires. For this purpose, they were briefed about the aim of the research and the questionnaires individually, since the participants were small in number. Also, each participant was provided with the copies of the two midterm tests so that they could refresh their memory of them, because ten weeks had passed since the second midterm tests were administered.

On April 16th, coordinator-instructors, who are the people directly responsible for separate courses like grammar, were interviewed to determine the objectives for the particular courses.

The contents of the tests were then correlated with the contents of the course materials. To pinpoint the numbers of item frequencies, the items in each unit of the book and on the test were counted. Besides the course books, additional class material like listening cassettes (for the listening course) was searched for item

frequency. The course objectives were evaluated in terms of their consistency with the tests' contents. Following this analysis, an evaluation of the content validity of the tests was made.

Data Analysis

Since there were not many teachers, and also to present the data clearly, the multiple-choice questions of the questionnaires were analyzed by putting each teacher's answer in tables. The answers to the open-ended questions were analyzed by grouping the responses into categories. Tables were used to show the results. To measure content validity, the course content items were specified and counted by analyzing the material of each course. Tests were analyzed to find these items' frequencies on them. The materials were then analyzed to find the exercise types in which the content items were presented, and they were correlated with the task types on the tests. The Pearson Product Moment Correlation analysis was used to correlate the item percentages and to show the significance values of the correlations between the course and test contents.

In the tables, taking their frequency in the material as basis, the items were rank-ordered from the most frequent to the least. The matching between the test and material frequencies of language items studied during the first semester were then evaluated. The relationship between the course objectives and test /material content was evaluated. Questionnaire results were related to the findings of the analyses.

CHAPTER FOUR: DATA ANALYSIS

Data Analysis Procedure

The following procedure was used to analyze the data: First, the responses of the instructors to the questionnaires were analyzed. The number of the subjects who responded was 10, with a response rate of 100%. There were three different questionnaires. A questionnaire was prepared for each course. The core-course questionnaire consisted of three sections and 22 items (see Appendix A). The grammar questionnaire consisted of three sections and 20 items (see Appendix B). The listening questionnaire consisted of three sections and 22 items (see Appendix C). Since the number of the participants was very small, all of the results are grouped and shown in tables, then necessary explanations are made.

Next, the first semester midterm tests' sampling of the courses' content was analyzed through counting the frequencies of the specific content items in the materials and their frequencies on the tests. Then, since different methods of testing the same content may produce very different results and affect the content validity of a test, as suggested by Shohamy (1984) and Brown (1994), the frequencies of the material task types (exercises) versus the test task types were counted and correlated for the core-course and grammar. For the listening course, there was no separate table for the exercise types, because after an initial analysis, the items both in material and on the test were found to be essentially in the same format. The significance of the relationship between the content items and exercise types in the materials and the items on the tests were computed by means of Pearson Product Moment Correlation Coefficient, using Excel. The results are shown in tables.

Finally, the coordinators' perceptions of the course objectives were elicited through interviews. Then, the tape-recordings of these interviews were transcribed and analysed in the light of what has been found to be covered in the materials and on the tests.

Results of the Study

Analysis of the questionnaires

The questionnaires were given to 10 intermediate level English instructors; four for core-course, three for grammar, and three for listening. They all responded. In all of the questionnaires, the first and second questions were about their experience in teaching. Their responses are reported in chapter three.

1st Core-Course Midterm Test

Table 1 below shows the answers given to the questions 3, 4, 5, 10, 11, and 12. Question 3 was: "The following areas are taken from your course material. Which "function" area(s) of core-course content should have been represented more on the first midterm test? Choose from below." The instructors were given the list of 29 function areas which were covered up to the first midterm test to choose from. The next question, question 4, asked the same thing about the form areas, giving a list of 14 forms to choose from. Question 5 was: "Were there any content area(s) on the first core-course midterm test that you did not include on the test although you taught that content in the class?" It required a yes/no answer, and if the answer was yes, the participant was asked to specify those areas.

Question 10 was: "To what extent do you think the content of the core-course questions on the first midterm test matched the core-course content you taught in your class?". Question 11 asked: "To what extent do you think the questions on

the first midterm test matched your course objectives?". Question 12 was: "To what extent do you think the exercise types you used in the classroom matched the question types on the first midterm test?" All these three questions were to be answered by choosing from *very little*, *a little*, *some* and *a lot*. No answers are reported for Question 4 in the table below since no form areas were selected by the teachers as being underrepresented on the test.

Table 1.

Core-course Instructors' Answers to the Questions 3, 5, 10, 11, 12

1st Core-course Midterm	Teacher A	Teacher B	Teacher C	Teacher D
Q.3	--	--	(bb) Talking about imaginary situations	--
Q.5	Functions, listening	Listening	--	--
Q.10	Some	Some	A lot	Some
Q.11	Some	Some	A lot	A lot
Q.12	A little	A little	A lot	Some

All four teachers said that they had not emphasized any of the given areas in their teaching, and also they had not weighted any of the areas on the test more than others (as answers to questions 7, 8, and 9, see Appendix A). Table 1 shows that only teacher C gave an area that should have been represented more: " Talking about imaginary situations." Teachers A and B agreed that "listening" was not included on the test. Teacher A adds that "Functions" were not covered on the test either. Question 6 required their reasons for not including items on the test. As reasons for not including " functions" on the test, teacher A said that there weren't many exercises related to the "functions", and they were not tested in the course book

either. For not including "listening" area on the test, teacher A said that listening was not tested because it was difficult to find suitable material. Teacher B said listening had not been tested because of some technical problems and thinking it was covered in the listening course.

The results shown in Table 1 above indicate that teacher C appears to diverge from all the other teachers, responding *a lot* to all the matching questions, not only for this test, but also for the second midterm test (see Table 2), and was the only teacher to rank the match so highly. While claiming that the second midterm test matched teaching content well, she also provided the longest list of areas that should have been represented on the test (see Table 2). This pattern of answer suggests that teacher C either did not understand the questionnaire or she was not being objective in her evaluation for unknown reasons. Therefore, her views will not be included in further evaluation. Concerning the match of teaching content and course objectives to the test, the remaining teachers generally felt that there was some correspondence. However, they were less positive about the match between instructional exercises and test exercises, with two of the three saying there was only *a little* match. This may be an accurate reflection because the range of exercises that can be done in class will always be greater than what is possible to include on a test, since a test is only meant to sample. Thus, this question is less important to determining validity. Overall, the results suggest the teachers felt that this test was moderately valid.

Core-course 2nd Midterm Test

Table 2 below shows the answers given to the questions 13, 14, 15, 20, 21, and 22. Question 13 was: "The following areas are taken from your course material. Which "function" area(s) of core-course content should have been represented more

on the second midterm test? Choose from below." The instructors were provided with the list of 37 function areas which were covered between the administration of the first midterm test and the second midterm test to choose from. Question 14 asked the same question about the form areas with a list of 18 forms to choose from. Question 15 was: "Were there any content area(s) on the second core-course midterm test that you did not include on the test although you taught that content in the class?" It required a yes/no answer, and if the answer was yes, the participant was asked to specify those areas.

Questions 20, 21 and 22 were asking about the match between the course and test content. Question 20 was: "To what extent do you think the content of the core-course questions on the second midterm test matched the core-course content you taught in your class?". Question 21 inquired: "To what extent do you think the questions on the second midterm test matched your course objectives?" Question 22 was: "To what extent do you think the exercise types you used in the classroom matched the question types on the second midterm test?" All these three questions were to be answered by choosing from four options: *very little*, *a little*, *some* and *a lot*. No results are reported in the table for question 14, as no form areas were selected by the teachers as being underrepresented on the test.

Table 2.

Core-course Instructors' Answers to the Questions 13, 15, 20, 21, 22

2nd Core-course Midterm	Teacher A	Teacher B	Teacher C	Teacher D
Q.13	--	--	(p) saying thank you, (ff) asking for travel information, (gg) correcting a mistaken information, (hh) talking about food, (e) infinitives after direct objects, (p) sequence of tenses	(ff) asking for travel information, (p) sequence of tenses
Q.15	Functions, listening	Listening	--	--
Q.20	Some	Some	A lot	Some
Q.21	Some	Some	A lot	A lot
Q.22	A little	A little	A lot	Some

As is seen in Table 2, two of the instructors agreed that one of the function areas "Asking for travel information", and one form area " Sequence of tenses" should have been represented more on the second midterm test. The instructors gave the same answers as they did for the first midterm to the questions related to the emphasizing of areas in teaching and weighting on the test. They said no area was emphasized or weighted more. Instructors' reasons for not including "functions" and "listening" areas on the second core-course test were the same with the reasons given for the first one.

The answers given to questions 20, 21, and 22 about the second midterm test are exactly the same answers as the ones given for the equivalent questions about the first midterm test. The instructors thought the matching of the course objectives with the items on the test was the highest among the three categories. The matching of the

teaching content with the test content was lower, and the matching of the exercise types with the question types on the test was the lowest, in their opinion. These results suggest that they also felt this test was of moderate validity.

1st Grammar Midterm Test

Table 3 below shows the answers given to questions 3, 4, 6, 8, 9, 10, and 11 on the questionnaire given to grammar teachers. Question 5 was a follow-up for question 4, requiring the reasons for not including the items the instructors had said they had not. Question 7 was a follow-up for the 6th question and was aimed to elicit additional information about 6. The answers to these two questions are not included in the table, but will be presented at relevant places below. Question 3 was: "The following areas are taken from your course material. Which area(s) of grammar course content should have been represented more on the first midterm test? Choose from below." The instructors were given the list of 17 structure areas to choose from which were covered up to the first midterm test. Question 4 was: " Were there any content area(s) on the first grammar midterm test that you did not include on the test although you taught that content in the class?" It required a yes/no answer, and if the answer was *yes*, the participant was asked to specify those areas. Question 6 was: "Before the first midterm test, did you emphasize some area(s) of the grammar course content in teaching over other areas?" If affirmative, the participant was required to supply the areas that she thought she had emphasized. Question 8 was: "Did you think some grammar course areas were more important than the others when you were designing your test, so you weighted them more on the first midterm test?" If yes, she was to specify these areas.

Question 9 was: "To what extent do you think the content of the grammar questions on the first midterm test matched the grammar content you taught in your class?". Question 10 inquired: "To what extent do you think the questions on the first midterm test matched your course objectives?". Question 11 was: "To what extent do you think the exercise types you used in the classroom matched the question types on the first midterm test?" All these three questions were to be answered by choosing from four options: *very little*, *a little*, *some* and *a lot*.

Table 3.

Grammar Instructors' Answers to the Questions 3, 4, 6, 8, 9, 10, 11

1st Grammar Midterm	Teacher A	Teacher B	Teacher C
Q.3	Wh questions	Wh questions	Wh questions
Q.4	Wh questions	Wh questions	Wh questions
Q.6	--	Tenses	--
Q.8	The contrast between tenses	--	--
Q.9	A lot	Some	A lot
Q.10	A lot	A lot	A lot
Q.11	Some	Some	Some

As the answer to the 3rd question on the grammar questionnaire, all three teachers agreed that the "Wh- questions" topic area should have been represented more on the first midterm test. "Wh- questions" was also given as the only area that was not included on the test by all of the teachers, as response to the 4th question. As reasons for not including "Wh- questions" on the first midterm test, teacher A said they had supposed students knew it, for that level it was too simple. Teacher B reasoned that it was given as homework, so was not emphasized in the teaching as important. Teacher C said in a similar vein that the syllabus had not included this

area, so it was just given as homework. Only teacher A said that she had emphasized "The contrast between tenses" in her teaching and, only teacher B said that she had weighted "Tenses" more on the test, as answers to questions 6 and 8 respectively. Answering question 7, teacher A said that she was able to emphasize "the contrast between tenses" on the test too.

Overall, the answers given to these questions show that the teachers agreed about the test and felt that it did adequately represent what they had taught. Those areas they did not cover were justified. The answers to the remaining matching questions in Table 3 confirm this analysis. All the teachers agreed that the test matched course objectives a lot, and two of three felt that it matched teaching content at the same level. They felt less strongly about the match of test tasks to classroom tasks, which may be explained by the issue of sampling discussed earlier. In all, the evidence suggests the teachers felt this test was very valid.

2nd Grammar Midterm Test

Table 4 below shows the answers given to the questions 12, 13, 15, 17, 18, 19, and 20. Question 14 asked the reasons for not including some of the items on the test, and question 16 was a follow-up to 15, inquiring about whether the teachers who said they had emphasized some items in their teaching had also done so on the test. Both of the teachers responded *yes* to question 16. Question 12 was: "The following areas are taken from your course material. Which area(s) of grammar course content should have been represented more on the second midterm test? Choose from below." The instructors were given the list of 15 structure areas to choose from which were covered after the first midterm up to the second midterm test. Question 13 was : "Were there any content area(s) on the second grammar

midterm test that you did not include on the test although you taught that content in the class?" It required a yes/no answer, and if the answer was yes, the participant was asked to specify those areas. Question 15 was : "Before the second midterm test, did you emphasize some area(s) of the grammar course content in teaching over other areas?" If affirmative, the participant was required to specify the areas that she thought she had emphasized. Question 17 asked: "Did you think some grammar course areas were more important than the others when you were designing your test, so you weighted them more on the second midterm test?" If yes, she was to specify these areas.

Questions 18, 19 and 20 were related to the teachers' opinions about the course and test content match. Question 18 was: "To what extent do you think the content of the grammar questions on the second midterm test matched the grammar content you taught in your class?". Question 19 required: "To what extent do you think the questions on the second midterm test matched your course objectives?". Question 20 was: "To what extent do you think the exercise types you used in the classroom matched the question types on the second midterm test?" All these three questions were to be answered by choosing from four options: *very little*, *a little*, *some* and *a lot*.

Table 4.

Grammar Instructors' Answers to the Questions 12, 13, 15, 17, 18, 19, 20

2nd Grammar Midterm	Teacher A	Teacher B	Teacher C
Q.12	--	(a) Adjectives (b) adverbs (m) reported speech	(m) reported speech
Q.13	--	Reported speech	Reported speech
Q.15	Modals, Differences between modals	Perfect modals	--
Q.17	--	Tenses	Active-Passive, Modals
Q.18	A lot	Some	A lot
Q.19	A lot	A little	A lot
Q.20	Some	Some	Some

Once more, the relatively few selections made by teachers for any of the first four questions suggests that in general they felt the test matched their teaching well, but perhaps not as well as the first midterm. Some areas they did commented on are worth further discussion. Teachers B and C both indicated that "reported speech" was not included on the test and should have been. However, teacher A did not mention this at all. Likewise, while teachers A and B reported that aspects of modal verbs were more emphasized in teaching, teacher C said they were given more weight on the test. These results suggest more differences among the teachers than on the first midterm and, specifically, that they differ on how modals should be treated in instruction and testing. In addition, options "o) Compound sentences" and " p) Either-or..." later had to be dismissed from analysis of the content, because the

teachers reported that although these areas were included in the syllabus, they did not have enough time to teach them (see Appendix B).

As reasons for not asking the "Reported speech" on the second midterm test, instructors B and C's answers were similar. While instructor B said it was easy and doing the exercises was enough, teacher C said they had assumed that students already knew about it. (So it was *not* necessary to test it.) From the way the instructors put it, it can be easily inferred that if an item is perceived to be easy by the instructors, it stands much less chance of being asked on the test.

Table 4 above shows that as answers to the questions 18, 19, and 20, while teacher A and C were positive that their teaching and objectives matched the test very much, teacher B was less positive, and for the matching of the objectives with the test, she was negative. But they all agreed that the test representation of the task formats was to a moderate degree. They stated that it was mainly because of their not frequently practicing multiple-choice tasks in their classrooms, but asking such tasks on the test to a great extent (which turned out to be true after the content analysis). Overall, these results support the analysis above, suggesting more differences in the teachers' views of the validity of the test, but a belief in a moderate level of test representativeness of the course content.

1st Listening Midterm Test

According to the answers given to the questions 3, 4, 6 and 8 (see Appendix C), all of the listening teachers were of the same opinion that there were no areas that should have been represented more on the test. They claimed that all of the listening areas specified in the questionnaire were included on the test, and none of them were

weighted more on the test. They also said that they had not emphasized any of the listening areas in their teaching.

The table below presents the answers given to the questions 9, 10, and 11. Question 9 was: "To what extent do you think the content of the listening questions on the first midterm test matched the listening content you taught in your class?". Question 10 required: "To what extent do you think the questions on the first midterm test matched your course objectives?". Question 11 was: "To what extent do you think the exercise types you used in the classroom matched the question types on the first midterm test?" All these three questions were to be answered by choosing from four options: *very little*, *a little*, *some* and *a lot*.

Table 5.

Listening Instructors' Answers to the Questions 9, 10, 11

1st Listening Midterm	Teacher A	Teacher B	Teacher C
Q.9	A lot	A lot	A lot
Q.10	A lot	A lot	A lot
Q.11	A lot	A lot	A lot

As shown in the table above, for the questions 9, 10, and 11, they all agreed that their teaching, objectives and exercises all matched well with the test. This result is in accord with the answers given to the preceding questions and suggest that the teachers all felt the test was very good at representing the course content.

Listening 2nd Midterm Test

Table 6 below shows the answers given to the questions 12, 13, 14,15, 20, 21, and 22. Question 12 was: "The following areas are taken from your course material.

Which listening strategy area(s) of the course content should have been represented more on the second midterm test? Choose from below." The instructors were given the list of 8 structure areas to choose from which were covered after the first midterm up to the second midterm test. Question 13 was: "Which vocabulary area(s) of the listening course content should have been represented more on the second midterm test?," with 11 vocabulary areas to choose from. Question 14 asked: "Which functional area(s) of the listening course content should have been represented more on the second midterm test?," with a list of 10 functional areas. Question 15 was: "Were there any content area(s) on the second listening midterm test that you did not include on the test although you taught that content in the class?" It required a yes/no answer, and if the answer was yes, the participant was asked to specify those areas.

Question 20 was: "To what extent do you think the content of the listening questions on the second midterm test matched the grammar content you taught in your class?". Question 21 was: "To what extent do you think the questions on the second midterm test matched your course objectives?". Question 22 was: "To what extent do you think the exercise types you used in the classroom matched the question types on the second midterm test?" Each of these three questions were to be answered by choosing from four options: *very little*, *a little*, *some* and *a lot*.

Table 6.

Listening Instructors' Answers to the Questions 12, 13, 14, 15, 20, 21, 22

2nd Listening Midterm	Teacher A	Teacher B	Teacher C
Q.12	Listening strategy areas: a, c, e, g, h	Listening strategy areas: a, c, e, g, h	Listening strategy areas: b, c, e, g, h
Q.13	Vocabulary areas: a, c, g, h, j, k	Vocabulary areas: a, b, c, g, h, i, j, k	Vocabulary areas: a, c, g, j, k
Q.14	Functional areas: b, d, e, g, h, i	Functional areas: a, d, f, g, h, i, j	Functional areas: a, d, f, g, h, i
Q.15	Listening strategy areas: a, c, e, g, h Vocabulary areas: a, c, g, h, j, k Functional areas: b, d, e, g, h, i	Listening strategy areas: a, c, e, g, h Vocabulary areas: a, b, c, g, h, i, j, k Functional areas: a, d, f, g, h, i, j	Vocabulary areas: a, c, g, h, j, k
Q.20	A little	Some	Some
Q.21	Some	A lot	A lot
Q.22	A lot	A lot	Some

Note: See Appendix C to see what the letters refer to

The results shown in Table 6 indicate that teachers A and B agree that a large number of listening strategy, vocabulary, and functional areas were not included on the test and should have been represented more. Teacher C generally agreed with them about the need to represent these areas more, but felt that listening strategy and functional areas had all been included on the test. The common listening strategy areas that were chosen by all three teachers which they think were not represented on the test adequately are "c, e, g, h". That is, c) Following a story; predicting and guessing, e) Listening against background noise, g) Matching with your own

interpretation, and h) Following a story; matching with your own expectations. The vocabulary areas common in all of their answers include "a, c, g, j, k". That is, a) Music, c) Visitors; dogs; shops; embarrasment, g) Fears and phobias; crime, j) Buying and selling; household objects, and k) Superstitions; fortune telling. Their common answers for the functional areas include "d, g, h, i". That is, d) Making enquiries; giving directions; small talk, g) Describing a process, h) Describing objects, and i) Describing the past; interpreting a picture. The responses to questions 17 and 19 (see Appendix C) showed that the instructors had not emphasized any of the given listening areas in their teaching and there were no areas that had been weighted more on the test.

The table above demonstrates the instructors were not so positive about the match of the teaching with the test as they were for the first midterm test, but their answers indicated that they thought that the match between the course objectives and the test and also the exercises and the test tasks were good. This result raises the interesting question of why the test matched course objectives well, but not their teaching. While Hughes (1989) says that it is better to match tests to objectives rather than teaching, the ideal is to have all three – objectives, teaching, and the test – coordinated. Still, the overall results suggest that they found the test moderately representative of the course content.

The teachers answers to the question 16, which was inquiring about their reasons for not including some areas on the test, are displayed in Table 7 below.

Table 7.

Listening Instructors' Answers to the Question 16

	Reasons for not including areas on the test		
2nd Listening Midterm	Teacher A	Teacher B	Teacher C
Forming a mental picture	It was very difficult to test	--	--
Following a story; predicting & guessing	Unable to find suitable material for test	--	Not suitable for test
Listening against background noise	Because of technical shortcomings & not suitable for our purposes	Technical facility shortcomings	Even teaching was difficult & it would ensure failure of students
Matching with your own interpretation	Was not one of the primary goals to teach & supplementary	Students had not reached at that level yet	Risky area & difficult to evaluate
Following an explanation & matching with your own beliefs	Students' level was not high enough	It was supplementary, used as pre- & post- listening activity	Risky & difficult to evaluate
Music	--	Because of its incorrect structures and sometimes slang	--

While there is variation among the teachers as to why certain areas were not included on the test, certain themes, such as concern for the students' ability level and lack of technical facilities, do emerge. Perhaps it was because the teachers could provide justifications for the exclusion of some items that they still felt the test was

moderately valid. This may also explain why they felt the match of the test to their teaching was lower than to the course objectives. The excluded areas were ones they could cover in class, but were not able to do on the test.

Analysis of the Course Material Content vs. Test Content

To analyze how the content of a particular course that had been taught up to the midterm tests was represented on the midterm test of that course, item frequencies were counted. For the Listening first midterm test, the occurrence of each phoneme was counted both at sentence and individual word level. That was because these were the two main exercise types in the material in which the phonemes were presented. For the second midterm test, a course book (*Listening 2*) was used, and in it there were three main areas of listening subskills: Listening strategy, vocabulary, and function. Item frequencies for each of these three areas were counted separately. Listening strategy areas were also types of exercises, so the results for the exercise types were explained on the table for listening strategy areas (Table 17a). All results are displayed in tables.

Next, the item frequencies of the grammar course were counted and compared. To do that, first the material which consisted of the handouts that were used in the classroom was analyzed. The material consisted of structural exercises like writing the correct simple past forms of the verbs provided in the parentheses beside given sentences. The items that were required to complete an exercise were counted for each topic. For example, occurrence of the past simple items was determined by counting the blanks that were to be completed only by a past simple item. Usually, the required forms to complete an exercise were given in the

instruction, but when that was not available, the answers (and thus the items) were singled out from the key sheet, and added to the list of items under their relative topic categories. In some of the exercises, introductory examples were given to help the students do the exercise appropriately, but these examples were not counted. Only the items that the students were required to supply or to correct were included. So, the size of the chunks of items that were counted in the material and on the test was the same. For example, if it was a completing-the-sentence type exercise, the size of the item (e.g. a verb) to complete a sentence in an exercise in the course material was the same with the item required on the test. That made the item frequencies in the material and on the test directly comparable. The course book was used only for self-study, and since the given assignments did not cover all the book related to the topics taught in class, and the exact nature of the assignments was unclear, it was not included in the item analysis. I only analyzed the material that was shared by all the instructors and was obligatory for the students to learn. Exercise types were counted again item by item, as single tasks, not as groups of similar exercises. Each counted item that was testing a certain grammar topic was then classified under a certain exercise type.

For the core-course, there were two skill areas specified in the course book: function and form. Subskills under these areas were specifically labeled in the course book (*Spectrum 4*) for each unit. Up to the first midterm test 5 units were covered, and then 8 more units were covered until the administration of the second midterm test. The units were divided into small sections, and thus, for example one section covered only the "complaining" function plus the "so vs. such" form. To find out the frequencies, the number of exercises requiring a certain function use or

form use was counted. Forms and functions were mostly together in the same exercise, so it was possible for an item to be mentioned twice: once under the function category, then the form. Unlike grammar, for the core-course form areas, instead of the smaller chunks like verbs, the number of exercises (as grouped in the course book) was counted. So, the size of the form chunks mentioned for the material frequencies was bigger than the chunks of the test items (unlike the case for the grammar course). But the methods selected for counting the frequencies were consistently used all through the material and the test, and the size of the items in the material and on the test were proportional, so the figures for both were comparable.

Finally, the objectives which were provided by the coordinators in the interviews were presented and discussed in terms of the results of this study.

For all three courses, in counting the frequencies, the recurring items (very few in number) from the first midterm were not included as part of the second midterm content, because for the second midterm tests, the students were only responsible for the items that were covered *after* the first midterm test was administered. The results for each course are presented in tables separately, then the necessary explanations are made.

For the core-course, the material content had been divided into two: function areas and form areas. These areas are presented in separate tables. First, the function and form frequency correlations of the first midterm test, then of the second midterm test will be presented below.

Table 8.(a)

The Correlation Between 1st Midterm Core-course Material Content with Test

Content: Function Areas

1st Core-course Midterm	Frequency in material		Frequency on test	
Asking for and giving an opinion	8	(8%)	0	(0%)
Asking a friend for information	8	(8%)	0	(0%)
Talking about where products are produced	8	(8%)	0	(0%)
Expressing surprise	5	(5%)	0	(0%)
Complaining	5	(5%)	0	(0%)
Making plans	5	(5%)	0	(0%)
Talking about imaginary situations	5	(5%)	0	(0%)
Talking about likes and dislikes	4	(4%)	0	(0%)
Asking about and confirming facts	4	(4%)	0	(0%)
Confirming information	3	(3%)	0	(0%)
Talking about exceptions	3	(3%)	0	(0%)
Talking about habits	3	(3%)	0	(0%)
Expressing anticipation	3	(3%)	0	(0%)
Talking about strenghts and weaknesses	3	(3%)	0	(0%)
Talking about famous landmarks and works	3	(3%)	0	(0%)
Telling about a past experience	3	(3%)	0	(0%)
Expressing duration	3	(3%)	0	(0%)
Talking about your life	3	(3%)	0	(0%)
Expressing a wish	3	(3%)	0	(0%)
Agreeing and disagreeing with an opinion	2	(2%)	0	(0%)
Talking about needs	2	(2%)	0	(0%)
Talking about the content of something	2	(2%)	0	(0%)
Making a decision with someone	2	(2%)	0	(0%)
Reporting a conversation	2	(2%)	0	(0%)
Stating a preference	1	(1%)	0	(0%)
Reacting to news	1	(1%)	0	(0%)
Talking about your class or group	1	(1%)	0	(0%)
Discussing plans	1	(1%)	0	(0%)
Resuming a conversation	1	(1%)	0	(0%)
Reading for overall comprehension	0	(0%)	1	(50%)
Reading for specific information	0	(0%)	1	(50%)
Total	97	(100%)	2	(100%)

(n = 31), r = -0.393, p <.05

It is clear from Table 8 (a) that there is no matching between the function areas covered in the material and the functions on the test at all. The correlation coefficient (r = -0.393) of their frequencies is significantly negative at .05 level. Two functions, "Reading for overall comprehension" and "Reading for specific

information" were tested, and both were missing in the course book. The reason for not asking the functions in the course material previously presented to the students might be that they are much harder to test and score than "reading comprehension."

Table 8.(b)

The Correlation Between 1st Midterm Core-course Material Content with Test Content : Form areas

1st Core-course Midterm	Frequency in material		Frequency on test	
The superlative of adjectives with the present perfect	9	(11%)	0	(0%)
"so" vs. "such"	8	(10%)	0	(0%)
Indefinite compounds with "except" and "but"	8	(10%)	0	(0%)
"while" vs. "during"	7	(9%)	1	(2.5%)
Gerunds	7	(9%)	1	(2.5%)
Tag questions	6	(7%)	0	(0%)
The passive without an agent	6	(7%)	0	(0%)
Relative clauses with "who" and "whose"	6	(7%)	2	(5%)
The passive in the present and past	5	(6%)	6	(15%)
Rejoinders showing interest or surprise	4	(5%)	0	(0%)
The passive in the past tense	4	(5%)	0	(0%)
"the whole", "all", and "every" with time expressions	4	(5%)	0	(0%)
Contrary-to-fact conditionals	4	(5%)	3	(7.5%)
"be used to"	3	(4%)	1	(2.5%)
"would rather"	1	(1%)	1	(2.5%)
Vocabulary	0	(0%)	25	(62.5%)
Total	82	(100%)	40	(100%)

($n = 16$), $r = -0.567$, $p < .05$

A significant negative correlation was found between the form areas in the course book and on the test of the first midterm. The striking point here is that

although vocabulary was not among the form areas in the book, it was the most frequently tested item on the test.

Table 9. The Correlation Between the Material and the Test Task Formats

1st Core-course Midterm		
Task Format (Exercise) Types	**Frequency in material**	**Frequency on test**
Acting out a similar conversation to the one that has just been listened to	26 (16%)	0 (0%)
Listening and practising conversations	25 (16%)	0 (0%)
Listening to and reading a passage and identifying statements as "true", "false" or "it doesn't say"	12 (8%)	0 (0%)
Answering questions and comparing the answers with classmates	10 (6%)	0 (0%)
Discussing questions about books, sports and crafts in groups	10 (6%)	0 (0%)
Listening for specific information	8 (5%)	0 (0%)
Listening to conversations and choosing answers from two alternatives	6 (4%)	0 (0%)
Completing sentences in an appropriate way	5 (3%)	0 (0%)
Matching expressions with the pictures they describe	5 (3%)	0 (0%)
Reading a text and answering the personal questions related to its topic	5 (3%)	0 (0%)
Negotiating answers to the questions related to a text	5 (3%)	0 (0%)
Rephrasing the given expressions	4 (3%)	0 (0%)
Asking/answering questions with a partner by looking at pictures	4 (3%)	0 (0%)
Talking about an experience by using the given expressions	4 (3%)	0 (0%)
Reading two sentences and identifying if they mean the same or different	3 (2%)	0 (0%)
Completing a text by using "so" or "such"	2 (1%)	0 (0%)
Completing sentences by choosing answers from a box	2 (1%)	0 (0%)
Asking questions by doing the necessary changes to given sentences	2 (1%)	0 (0%)
Matching expressions by choosing from two lists	2 (1%)	0 (0%)
Listening and taking notes	1 (0.6%)	0 (0%)
Listening to a conversation and looking at pictures to describe statements	1 (0.6%)	0 (0%)
Completing a text by using the superlatives of the adjectives in parentheses	1 (0.6%)	0 (0%)
Answering questions by using the superlative form of an adjective with the present perfect	1 (0.6%)	0 (0%)
Reacting to the given statements appropriately	1 (0.6%)	0 (0%)

Expressing original statements to get reactions	1 (0.6%)	0 (0%)
Looking at pictures and discussing what you know about them	1 (0.6%)	0 (0%)
Making up sentences about the given places and works by using the passive	1 (0.6%)	0 (0%)
Combining sentences with "who" or "whose" to describe novels	1 (0.6%)	0 (0%)
Determining the positiveness or negativeness of the given adjectives	1 (0.6%)	0 (0%)
Writing a short description of a book	1 (0.6%)	0 (0%)
Rewriting sentences by using "except" or "but"	1 (0.6%)	0 (0%)
Completing a conversation by choosing sentences from a box	1 (0.6%)	0 (0%)
Telling jokes in groups	1 (0.6%)	0 (0%)
Changing the given sentences from active to passive	1 (0.6%)	0 (0%)
Finding out what somebody likes/dislikes most	1 (0.6%)	0 (0%)
Retelling a story in one's own words after hearing it twice	1 (0.6%)	0 (0%)
Completing and making sentences with "while" or "during"	1 (0.6%)	0 (0%)
Matching words and phrases with their definitions	1 (0.6%)	15 (36%)
Completing sentences by using an appropriate word from the given list	0 (0%)	10 (24%)
Completing sentences grammatically and meaningfully by using connectors	0 (0%)	6 (14%)
Completing sentences by using the correct forms of the verbs in parentheses	0 (0%)	6 (14%)
Combining two sentences by using contrary-to-fact conditionals	0 (0%)	3 (8%)
Reading a text and stating its main purpose	0 (0%)	1 (2%)
Completing a chart by using the information from a text	0 (0%)	1 (2%)
Total	159 (100%)	42 (100%)

(n = 44), r = -0.191

Though found to be not significant at .05 level, the correlation coefficient of the first midterm task formats indicates a high negative relationship between the course book and the test. According to the results in Table 9, almost all of the task types on the test were new to the students.

Table 10.(a) The Correlation Between 2nd Midterm Core-course Material Content with Test Content: Function Areas

2nd Core-course Midterm	Frequency in material	Frequency on test
Giving warnings and advice	14 (8%)	0 (0%)
Showing concern	10 (6%)	0 (0%)
Asking a favor of someone	9 (5%)	1 (5.6%)
Giving an excuse	8 (5%)	0 (0%)
Stating a condition	8 (5%)	0 (0%)
Talking about problems	7 (4%)	1 (5.6%)
Stating a prohibition	7 (4%)	0 (0%)
Discussing arrangements and obligations	6 (3%)	0 (0%)
Cancelling plans	6 (3%)	1 (5.6%)
Making a request	6 (3%)	0 (0%)
Offering to do someone a favor	5 (3%)	0 (0%)
Having something done	5 (3%)	0 (0%)
Telling about a past experience	5 (3%)	0 (0%)
Asking for a reason	5 (3%)	0 (0%)
Catching up on what someone's been doing	5 (3%)	0 (0%)
Imagining something	5 (3%)	0 (0%)
Expressing regret	5 (3%)	1 (5.6%)
Making a judgement	5 (3%)	1 (5.6%)
Starting a conversation	4 (2%)	1 (5.6%)
Describing an injury	4 (2%)	0 (0%)
Reporting requests and instructions	4 (2%)	0 (0%)
Talking about moods and feelings	4 (2%)	0 (0%)
Asking for information	4 (2%)	1 (5.6%)
Going shopping	3 (2%)	0 (0%)
Telling a frightening story	3 (2%)	0 (0%)
Agreeing or refusing to do a favor	3 (2%)	0 (0%)
Greeting someone after a long time	3 (2%)	0 (0%)
Asking for travel information	3 (2%)	0 (0%)
Correcting a mistaken impression	3 (2%)	0 (0%)
Apologizing	2 (1%)	0 (0%)
Refreshing your memory	2 (1%)	0 (0%)
Asking permission	2 (1%)	1 (5.6%)
Saying thank you	2 (1%)	0 (0%)
Discussing possible plans	2 (1%)	0 (0%)
Talking about food	2 (1%)	0 (0%)
Making an objection	2 (1%)	0 (0%)
Giving unwanted advice	1 (1%)	0 (0%)
Asking for an opinion	0 (0%)	1 (5.6%)
Reacting to news	0 (0%)	1 (5.6%)
Listening to a conversation to get specific information	0 (0%)	8 (44%)
Total	174 (100%)	18 (100%)

(n =40), r = -0.251

Though not significant at .05 level, the correlation between the function areas in the course book and on the test for the second midterm was found to be highly negative. More than half of the functions required from students to perform on the test were new to the students.

Table 10.(b)

The Correlation Between 2nd Midterm Core-course Material Content with Test Content: Form Areas

2nd Core-course Midterm	**Frequency in material**	**Frequency on test**
Embedded yes-no questions	10 (10%)	0 (0%)
Negative questions	8 (8%)	0 (0%)
The present perfect continuous	8 (8%)	0 (0%)
The present perfect continuous vs. the present perfect	7 (7%)	0 (0%)
Conditional sentences	7 (7%)	0 (0%)
Noun compounds	6 (6%)	0 (0%)
The simple past vs. past continuous with "when" and "while"	6 (6%)	5 (16%)
"be supposed to" vs. "have to" in negative statements	6 (6%)	0 (0%)
"be supposed to"	5 (5%)	0 (0%)
"should have"	5 (5%)	0 (0%)
The causative "have"	4 (4%)	0 (0%)
Infinitives after direct objects	4 (4%)	0 (0%)
"had better" and "had better not"	4 (4%)	0 (0%)
"if" vs. "unless"	4 (4%)	0 (0%)
Reported speech with imperatives	4 (4%)	0 (0%)
Reflexive pronouns	3 (3%)	0 (0%)
Sequence of tenses	3 (3%)	0 (0%)
"too...to" and "not...enough...to"	2 (2%)	0 (0%)
Vocabulary	0 (0%)	27 (84%)
Total	96 (100%)	32 (100%)

($n = 19$) $r = -0.494, p < .05$

Table 10 (b) above shows a significant negative correlation between the form areas of the second midterm material and test at .05 level. The most frequently tested item "Vocabulary", was not among the form areas in the material.

Table 11.

The Correlation Between the Material and the Test Task Formats

2nd Core-course Midterm	Frequency in material	Frequency on test
Task Format (Exercise) types		
Listening and practising conversations	49 (20%)	0 (0%)
Acting out a similar conversation to the one that has just been listened to	37 (15%)	0 (0%)
Listening for specific information to complete a chart	21 (9%)	8 (16%)
Listening to conversations and choosing answers from two alternatives	12 (5%)	0 (0%)
Listening to and reading a passage and identifying statements as "true", "false" or "it doesn't say"	11 (5%)	0 (0%)
Rephrasing the given expressions	11 (5%)	0 (0%)
Playing roles	11 (5%)	0 (0%)
Completing a conversation by using the correct forms of the verbs in parentheses	10 (4%)	0 (0%)
Discussing questions about presented topics in groups	7 (3%)	0 (0%)
Reading a text and answering questions related to its topic	6 (3%)	0 (0%)
Negotiating answers to the questions related to a text	5 (2%)	0 (0%)
Matching pictures with the listened information	4 (2%)	0 (0%)
Asking about and reporting classmates' personal lives	4 (2%)	0 (0%)
Listening to conversations to check implied statements	4 (2%)	0 (0%)
Matching expressions with pictures	3 (1%)	0 (0%)
Telling about a personal experience	3 (1%)	0 (0%)
Completing conversations with "be supposed to" or "have to"	2 (0.8%)	0 (0%)
Completing conversations by choosing sentences from a box	2 (0.8%)	0 (0%)
Completing conversations by using negative questions	2 (0.8%)	0 (0%)
Matching words with their definitions	1 (0.4%)	14 (28%)
Finding words from a text which mean the same as the given words	1 (0.4%)	7 (14%)
Writing what a pronoun in a text refer to	1 (0.4%)	6 (12%)
Looking at given pictures and using the key words to make up a story	1 (0.4%)	5 (10%)
Answering questions by using compound nouns	1 (0.4%)	0 (0%)
Asking/ answering questions by using the given information	1 (0.4%)	0 (0%)
Identifying if the given two words mean the same or not	1 (0.4%)	0 (0%)

Matching expresions by choosing from two lists	1 (0.4%)	0 (0%)
Looking at pictures and discussing what you know about them	1 (0.4%)	0 (0%)
Answering questions by using the information from a table	1 (0.4%)	0 (0%)
Talking about abilities using "if..."	1 (0.4%)	0 (0%)
Writing a letter by using the given information	1 (0.4%)	0 (0%)
Changing sentences to questions using "What if...?"	1 (0.4%)	0 (0%)
Rewriting questions by using the expressions from a box	1 (0.4%)	0 (0%)
Combining pairs of sentences into one sentence	1 (0.4%)	0 (0%)
Guessing the end of a story	1 (0.4%)	0 (0%)
Talking about imaginary situations	1 (0.4%)	0 (0%)
Rewriting sentences beginning with "if..."	1 (0.4%)	0 (0%)
Responding to a friend's problems by using conditional sentences	1 (0.4%)	0 (0%)
Completing a text with reflexive pronouns	1 (0.4%)	0 (0%)
Making a list of things that you are not supposed to and have to do	1 (0.4%)	0 (0%)
Making judgements about situations using "should/should not have"	1 (0.4%)	0 (0%)
Categorizing the listed shopping items properly	1 (0.4%)	0 (0%)
Describing a person to the classmates	1 (0.4%)	0 (0%)
Listening to conversations and putting the events in order	1 (0.4%)	0 (0%)
Completing an open-ended story	1 (0.4%)	0 (0%)
Completing conversations with "if" or "unless"	1 (0.4%)	0 (0%)
Changing expressions into compound nouns	1 (0.4%)	0 (0%)
Changing sentences into passive	1 (0.4%)	0 (0%)
Listening/retelling a story by using the given pictures	1 (0.4%)	0 (0%)
Responding to the given situations appropriately	1 (0.4%)	0 (0%)
Asking permission for given situations by using "Would you mind if...?"	1 (0.4%)	0 (0%)
Asking a classmate for help	1 (0.4%)	0 (0%)
Reading and writing a thank-you note	1 (0.4%)	0 (0%)
Finding sentences that mean the opposite of the given sentences	1 (0.4%)	0 (0%)
Completing a dialogue according to the function given in parentheses	0 (0%)	10 (20%)
Total	239 (100%)	50 (100%)

(n = 55), r = -0.012

The correlation between the task formats in the material and on the test of the second midterm was found to be negative, though not significant at .05 level. Only one of the higher frequency items, "Listening for specific information", was tested, and the testing frequency of this item was almost twice the amount it occurred in the

book. The most frequent format types, namely "Listening and practising conversations", and "Acting out a similar conversation...," were not asked on the test at all.

Table 12.

The Correlation Between 1st Midterm Grammar Material Content with Test Content

1st Grammar Midterm	Frequency in material		Frequency on test	
Topics				
The past simple	272	(19%)	29	(31%)
The past perfect	229	(16%)	6	(6%)
The present simple	181	(13%)	16	(17%)
The present perfect	163	(11%)	11	(12%)
The present continuous	145	(10%)	3	(3%)
The past continuous	122	(9%)	5	(5%)
The future simple ("will" and "going to")	66	(5%)	6	(6%)
"used to"	38	(3%)	0	(0%)
The present perfect continuous	36	(3%)	4	(4%)
The past perfect continuous	33	(2%)	5	(5%)
Wh- questions	28	(2%)	0	(0%)
The future continuous	26	(2%)	3	(3%)
Present continuous for the future	25	(2%)	0	(0%)
Present simple for the future	23	(2%)	0	(0%)
The future perfect	7	(0.4%)	5	(5%)
The future perfect continuous	7	(0.4%)	0	(0%)
"be to + infinitive"	6	(0.3%)	0	(0%)
Total	1407	(100%)	93	(100%)

(n = 17), r = 0.794, p = <.05

The correlation between the frequencies of topics covered in the material up to the first midterm test and their frequencies on the test was found to be significantly positive at .05 level. The most frequent topic both in the material and on the test was

"The past simple." But the second most frequent topic, "The past perfect," and the fifth frequent, "The present continuous" were almost three times less frequent on the test than they were in the material. Other than these two areas, the test was successful at sampling the areas covered in the course material.

Table 13.

The Correlations Between the Material and Test Task Formats

1st Grammar Midterm	**Frequency in material**	**Frequency on test**
Task Format (Exercise) types		
Completing sentences by using the correct forms of the verbs given in parentheses	667 (47%)	30 (32%)
Completing sentences by selecting and using the correct forms of verbs in a box	98 (7%)	0 (0%)
Completing sentences by choosing from two alternatives	96 (7%)	0 (0%)
Fill-in-the-blanks	77 (5%)	0 (0%)
Making sentences by using the given keywords	47 (3%)	0 (0%)
Correcting the underlined verbs if necessary	41 (3%)	0 (0%)
Correcting the tense in a sentence	39 (3%)	0 (0%)
Matching sentences with their interpretations	35 (2%)	0 (0%)
Completing half-sentences appropriately	34 (2%)	0 (0%)
Asking and answering questions according to a given context	26 (2%)	0 (0%)
Answering questions about a text	24 (2%)	0 (0%)
Combining two sentences into one sentence	19 (1%)	0 (0%)
Reading introductory sentences and completing the following ones	19 (1%)	0 (0%)
Filling out the time lines of the events of a text	17 (1%)	0 (0%)
Matching questions with their answers	16 (1%)	0 (0%)
Identifying tense mistakes	16 (1%)	0 (0%)
Writing sentences about the given situations	12 (0.8%)	0 (0%)
Numbering actions according to their occurrrences in time	11 (0.8%)	0 (0%)
Matching the beginnings of sentences with their endings	11 (0.8%)	0 (0%)
Rewriting sentences by using the given connectors in correct tense	10 (0.7%)	0 (0%)
Completing sentences by using the information given in a chart	10 (0.7%)	0 (0%)
Making sentences by using the information given in a chart	9 (0.6%)	0 (0%)

Discussing the meanings of verb forms in sentence pairs	9	(0.6%)	0	(0%)
Finding and correcting the mistakes in sentences	9	(0.6%)	0	(0%)
Writing questions for the answers by using the given keywords	9	(0.6%)	0	(0%)
Reading the answer and writing questions about the underlined words	9	(0.6%)	0	(0%)
Reading sentences and stating which event happened first	6	(0.4%)	0	(0%)
Using "when" or "after" to make sentences about situations	5	(0.3%)	0	(0%)
Using sentences on the left to complete paragraphs on the right	4	(0.3%)	0	(0%)
Fill-in-the-blanks (multiple-choice)	4	(0.3%)	0	(0%)
Looking at drawings and writing appropriate dialogues	4	(0.3%)	0	(0%)
Matching tenses with their correct description	4	(0.3%)	0	(0%)
Looking at diagrams and matching them with the tenses they describe	3	(0.2%)	0	(0%)
Making sentences about future definite plans	3	(0.2%)	0	(0%)
Getting/comparing information in charts by asking questions to a partner	2	(0.1%)	0	(0%)
Completing an open-ended conversation	1	(0.07)	0	(0%)
Writing as many interview questions as possible in 5 minutes	1	(0.07)	0	(0%)
Completing sentences with correct tense choosing from multiple choices	0	(0%)	63	(68%)
Total	1407	(100%)	93	(100%)

($n = 38$), $r = 0.368$, $p < .05$

Although the most frequent task format on the first midterm test was multiple choice, with 68%, it was absent in the material. And as presented in Table 13, from among the many task types frequent in the material, only the most frequent one "Completing sentences by using the correct forms of the verbs given in parentheses" was sampled. Despite these shortcomings in the sampling, the correlation between the frequencies of the exercise types in the material and their frequencies on the test was found to be significantly positive at .05 level. Though it would not represent the diversity of formats in the material again, if all the questions on the test were from the most frequent area, instead of 68% multiple choice, the correlation result would have been much higher.

Table 14.

The Correlation Between 2nd Midterm Grammar Material Content with Test Condent

2nd Grammar Midterm Topics	Frequency in material		Frequency on test	
The Passive	354	(22%)	9	(10%)
Articles	276	(17%)	5	(5%)
Reported speech	178	(11%)	0	(0%)
Present modals	165	(10%)	1	(2%)
Perfect modals	126	(8%)	9	(10%)
Comparative forms	119	(7%)	3	(3%)
Adjectives	75	(5%)	0	(0%)
Adverbs	70	(4%)	0	(0%)
Active vs. Passive	66	(4%)	21	(23%)
Superlative forms	50	(3%)	2	(2%)
Equative forms (as...as)	49	(3%)	1	(2%)
"Some", "Any", "No"	44	(3%)	0	(0%)
Quantifiers	38	(2%)	22	(24%)
"There" and "It"	15	(1%)	0	(0%)
Mixed tense types	0	(0%)	20	(22%)
Total	1625	(100%)	93	(100%)

(n = 15), r = -0.141

The correlation between the frequencies of topics covered in the grammar material up to the second midterm test and their frequencies on the test was found to be negative, though not significant at .05 level. The major difference between this test and the first one, which made the correlation lower, was the inclusion of the "Mixed type tense" questions, which was a topic absent in the material. While less frequent topics like "Quantifiers" and "Passive vs. Active" were heavily tested, a much more frequent topic, "Reported speech" was not tested at all.

Table 15.

The Correlation Between the Material and the Test Task Formats

2nd Grammar Midterm	Frequency in material		Frequency on test	
Task Format (Exercise) types				
Completing sentences by using the correct forms of the words in parentheses	339	(21%)	41	(44%)
Completing sentences by choosing from two alternatives	174	(11%)	0	(0%)
Completing sentences by using the correct forms of the verbs in a list	167	(10%)	0	(0%)
Turning reported sentences into direct, and the direct into the reported speech	124	(8%)	0	(0%)
Completing sentences by putting "a/an", "the" , "some",or "any" where necessary	95	(6%)	0	(0%)
Completing sentences by putting in "the" where necessary	94	(6%)	0	(0%)
Finding and correcting mistakes in a text	69	(4%)	0	(0%)
Changing the active to passive and the passive to active sentences	69	(4%)	0	(0%)
Completing sentences by using the correct forms of the quantifiers in a list	55	(3%)	0	(0%)
Completing sentences by using the information from a table	45	(3%)	0	(0%)
Writing a sentence to report the previous sentence	35	(2%)	0	(0%)
Completing sentences by choosing adjectives from a box	33	(2%)	0	(0%)
Choosing the item that best completes the sentence (multiple-choice)	31	(2%)	52	(56%)
Writing the plural of words	30	(2%)	0	(0%)
Identifying the word which should not be in a sentence	29	(2%)	0	(0%)
Rewriting sentences by putting the adverbs in their correct places	25	(1%)	0	(0%)
Completing sentences according to the given situations	21	(1%)	0	(0%)
Identifying the incorrect form in a sentence	17	(1%)	0	(0%)
Rewriting sentences by putting the adverbs in their correct places	16	(1%)	0	(0%)
Completing sentences by choosing verbs from a box and using information from a chart	15	(0.9%)	0	(0%)
Rewriting sentences by using the given keywords	14	(0.9%)	0	(0%)
Underlining the quantifiers that can be used with the accompanying nouns	10	(0.6%)	0	(0%)
Matching adjectives with the appropriate nouns	9	(0.6%)	0	(0%)
Writing sentences according to a given table	9	(0.6%)	0	(0%)
Writing the opposites of the given adjectives	9	(0.6%)	0	(0%)
Writing passive sentences beginning with "there" using given keywords	9	(0.6%)	0	(0%)
Completing sentences appropriately	9	(0.6%)	0	(0%)
Making compound adjectives to describe sentences	8	(0.5%)	0	(0%)
Making sentences about the given situations	8	(0.5%)	0	(0%)
Turning sentences into indirect questions	8	(0.5%)	0	(0%)
Choosing the correct statement about a sentence from two alternatives	8	(0.5%)	0	(0%)
Describing activities in pictures by using the given keywords	7	(0.4%)	0	(0%)
Describing differences in two pictures by using the given keywords	7	(0.4%)	0	(0%)
Matching requests with the correct responses	7	(0.4%)	0	(0%)
Completing statements about signs	5	(0.3%)	0	(0%)
Writing agreements with the given statements	5	(0.3%)	0	(0%)
Making sentences using the keywords and information from a poster/map	5	(0.3%)	0	(0%)
Changing sentences into two new sentences by using the words in parentheses	4	(0.2%)	0	(0%)
Writing a card for a special day by using as many superlatives as possible	1	(0.1%)	0	(0%)
Total	1625	(100%)	93	(100%)

($n = 39$), $r = 0.451$, $p < .05$

As shown in the table above, the frequencies of the grammar task types in the material which was covered between the first midterm and second midterm test and these task types' frequencies on the 2nd midterm test were found to be positively significant at .05 level. So, the task types for the second midterm can be said to correlate well. But the same problem with the first midterm persisted here, that is, the sampling of "multiple choice" task types on the test, which was one of the less frequent areas in the material.

Table 16.(a)

Correlation of the 1st Midterm Listening Material Content with Test Content

1st Listening Midterm	**frequency in material**		**total frequency**		**frequency on test**		**total frequency**	
	(in words)	(in sentences)			(in words)	(in sentences)		
Consonants								
Distinguishing between /s/ and /z/	12	35	47	(12%)	2	2	4	(13%)
Distinguishing between /p/, /v/ and /f/	23	11	34	(9%)	0	1	1	(3%)
Distinguishing between /tʃ/, /dʒ/ and /ʒ/	8	21	29	(7%)	1	2	3	(10%)
Distinguishing between /h/ and no /h/	24	4	28	(7%)	0	1	1	(3%)
Distinguishing between /b/, /v/ and /w/	16	10	26	(7%)	0	1	1	(3%)
Distinguishing between /s/ and / /	16	6	22	(6%)	0	0	0	(0%)
Distinguishing between /j/ and /dz/	16	4	20	(5%)	0	0	0	(0%)
Distinguishing between /l/,/r/ and /n/	15	5	20	(5%)	0	0	0	(0%)
Distinguishing between /p/ and /b/	10	8	18	(5%)	0	0	0	(0%)
Distinguishing between /t/ and /d/	10	7	17	(4%)	2	3	5	(16%)
Distinguishing between /k/ and /g/	10	7	17	(4%)	2	0	2	(6%)
Distinguishing between /ʃ/ and /tʃ/	12	4	16	(4%)	0	0	0	(0%)
Distinguishing between /l/ and /r/	10	6	16	(4%)	1	0	1	(3%)
Distinguishing between /θ/ and /s/	10	5	15	(4%)	2	1	3	(10%)
Distinguishing between /θ/ and /t/	10	5	15	(4%)	1	2	3	(10%)
Distinguishing between /θ/ and /f/	8	4	12	(3%)	0	0	0	(0%)
Distinguishing between /n/ and /ŋ/	8	4	12	(3%)	1	0	1	(3%)
Saying /m/	12	0	12	(3%)	0	0	0	(0%)
Distinguishing between / ð / and /z/	8	0	8	(2%)	2	0	2	(6%)
Distinguishing between / ð / and /d/	7	0	7	(2%)	1	2	2	(6%)
Distinguishing between /θ/ and /d/	0	0	0	(0%)	0	2	2	(6%)
Total	253	138	391	(100%)	15	16	31	(100%)

($n = 21$) , $r = 0.223$

The correlation between the consonant item frequencies in the material and on the test found to be positive, but not significant at the level of .05. When the table is analyzed, it can be seen that the most frequently taught item, "Distinguishing between /s/ and /z/" made 13% of the total frequencies and was successfully represented on the test by a 13%. But the sampling of other items was not so successful; for example the most frequently tested items /t/ and /d/ (16%) were found in the material at a quarter of that number (4%). But perhaps the most striking difference between the test and the teaching material is that although the distinction between /θ/ and /d/ was absent in the material, it was asked more than five of the other test items, at 6% of the whole test.

Table 16(b).

Correlation of the 1st Midterm Material Listening Content with Test Content

1st Listening Midterm	Frequency in material		Total frequency		Frequency on test		Total frequency	
	in words	in sentences			in words	in sentences		
Vowels								
Distinguishing between /əʊ/ and /o:/	12	28	40	(10%)	0	0	0	(0%)
Distinguishing between /æ/ and /^/	12	24	36	(9%)	0	1	1	(9%)
Distinguishing between /ɪ/ and /e/	10	24	34	(9%)	0	1	1	(9%)
Distinguishing between /ɒ/ and /o:/	8	23	31	(8%)	1	0	1	(9%)
Distinguishing between /ɜ:/ and /a:/	8	23	31	(8%)	1	0	1	(9%)
Saying /aɪ/	9	19	28	(7%)	0	0	0	(0%)
Distinguishing between /i:/and /ɪ/	8	16	24	(6%)	1	1	2	(19%)
Distinguishing between /æ/ and /e/	10	14	24	(6%)	1	0	1	(9%)
Saying /oɪ/	0	22	22	(6%)	0	0	0	(0%)
Distinguishing between /u:/ and /ʊ/	6	13	19	(5%)	0	0	0	(0%)
Distinguishing between /eɪ/ and /e/	12	5	17	(4%)	1	0	1	(9%)
Distinguishing between /eə/ and /ɪə/	12	5	17	(4%)	2	1	3	(27%)
Saying /aʊ/	0	10	10	(3%)	0	0	0	(0%)
Distinguishing between /e/ and /o:/	0	10	10	(3%)	0	0	0	(0%)
Distinguishing between /o/ and /o:/	0	10	10	(3%)	0	0	0	(0%)
Distinguishing between /^/ and /o:/	0	10	10	(3%)	0	0	0	(0%)
Distinguishing between /ɜ:/and /o:/	0	10	10	(3%)	0	0	0	(0%)
Distinguishing between /ɜ:/ and /i:/	0	10	10	(3%)	0	0	0	(0%)
Total	107	276	383	(100%)	7	4	11	(100%)

(n = 18), r = 0.193

Vowels demostrate a lower correlation. The most frequent item in the material (10%) was not tested at all, and the item which was the most frequent on the test, at 27%, had only 4% frequency in the material. All listening teachers had claimed that there was no area in their course content that was not included on the test. However, I found four vowel areas in the course content which were not included on the test, another factor causing a poorer test representation of the vowels than the consonants.

Table 17.(a)

The Correlation Between 2nd Midterm Listening Material Content with Test Content

Listening Strategy Areas	**Frequency in material**		**Frequency on test**	
Listening for specific information	24	(47%)	25	(100%)
Matching with your own interpretation	7	(13%)	0	(0%)
Following an explanation;matching with your own beliefs	5	(10%)	0	(0%)
Following a story; predicting and guessing	5	(10%)	0	(0%)
Forming a mental picture	4	(8%)	0	(0%)
Preparing an appropriate response	3	(6%)	0	(0%)
Following a story; following an explanation	3	(6%)	0	(0%)
Listening against background noise	1	(2%)	0	(0%)
Total	52	(100%)	25	(100%)

($n = 8$), $r = 0.973$, $p < .05$

Table 17 (a) shows that there is a significant positive correlation between the listening strategy item frequencies in the material and their frequencies on the 2nd midterm test. Although only one of the areas, that is, "Listening for specific information" was sampled on the test, it was the most frequent item in the course book, and this led this part of the listening test to have the highest positive correlation.

Table 17. (b)

The Correlation Between 2nd Midterm Listening Material Content with Test Content

Vocabulary areas	**Frequency in material**		**Frequency on test**	
The past;homes;art	8	(14%)	0	(0%)
Superstitions; fortunetelling	8	(14%)	0	(0%)
Music	6	(11%)	0	(0%)
Childhood; accidents; games	6	(11%)	0	(0%)
Buying and selling; food and drink	5	(9%)	10	(40%)
Visitors; dogs; shops; embarrasment	5	(9%)	0	(0%)
Travel; public places; parties	5	(9%)	0	(0%)
Buying and selling; household objects	5	(9%)	0	(0%)
Cities; landscapes	4	(7%)	0	(0%)
Fears and phobias; crime	3	(5%)	0	(0%)
Answerphones	1	(2%)	15	(60%)
Total	56	(100%)	25	(100%)

(n =11), r = -0.581

The correlation between the vocabulary item frequencies in the course book and on the test is negative but not significant at .05 level. That means the test shows a very low validity in terms of sampling the vocabulary areas. The least frequent vocabulary item in the material, "Answerphones," was tested as the most frequently.

Table 17. (c)

The Correlation Between 2nd Midterm Listening Material Content and Test Content

Functional areas	**Frequency in material**		**Frequency on test**	
Interpreting a picture;matching information with pictures	19	(36%)	0	(0%)
Imagining & remembering scenes	11	(21%)	0	(0%)
Giving explanations;making deductions	6	(10%)	0	(0%)
Describing scenes	4	(7%)	0	(0%)
Describing a process	3	(6%)	0	(0%)
Describing objects	3	(6%)	0	(0%)
Narration	3	(6%)	10	(40%)
Making arrangements	2	(4%)	15	(60%)
Requests;agreeing and refusing;asking personal information	1	(2%)	0	(0%)
Making enquiries;giving directions;small talk	1	(2%)	0	(0%)
Total	53	(100%)	25	(100%)

(n = 10), r = -0.251

The correlation between the material frequencies of the functional areas with their test frequencies was found to be negative, though not significantly at .05 level. Similar to the vocabulary areas, one of the least frequent functional area in the material, "Making arrangements" was tested the most. The other of the two areas tested was "Narration," which was again not so frequent in the material. The most frequent functional areas, "Interpreting a picture; matching information with pictures," and "Imagining & remembering scenes" were not tested at all.

Analysis of the Course Objectives

Objectives for the Core - course Course

The coordinator of the first semester intermediate level core-course provided the objectives for the course as:

1. Success in the exchange of information
2. Creating real-life situations with the help of the course book
3. Proceeding from comprehension through listening or reading to production through speaking or writing

These objectives seem in line with the activities presented in the book, but the testing of these "information exchanging" activities was ignored. The tests were dense with grammar items and vocabulary, and few tasks requiring listening or reading. Also, the second objective, "creating real-life situations" is obscure and needs clarification. Though it was stated as an objective, use of the production skills, i.e. speaking and writing, was not tested at all. In fact, these objectives are not specific enough to be useful for testing.

Objectives for the Grammar Course

The coordinator of the first semester intermediate level grammar course supplied the objectives for the course as:

1. Having students give grammatically, situationally and semantically correct responses
2. To help students to understand and use English
3. Having students repeat structures that were taught in class
4. Having students use the grammatical structures in everyday English

5. Improving students' proficiency level so that they will be able to follow the courses in their majors
6. Training students for the TOEFL

The objectives given above for the grammar were too general to have been tested with the grammar tests administered during the first semester. Also, these objectives are not in agreement with what was found to be the grammar material content. The material totally consists of structural exercises, and has nothing to do with the "communicative use" of English. The exercises do not prepare students for the TOEFL, or any TOEFL-like proficiency test as the one the students have to pass et the end of the academic year, because a very small amount of the exercises was in multiple-choice format, as shown in Table 13 and 15.

Objectives for the Listening Course

The coordinator of the first semester intermediate level listening course gave the objectives for the course as:

1. Having students discriminate phonemes
2. Having students learn the functions stated in the book

The results of the analysis of the first midterm material and test show that the first objective of the listening course mentioned above was in line with the content the students were presented in the first midterm material and its test. For the second objective, out of ten functions, two of the less frequent function areas were found to be tested. The most emphasized functions in the book were not tested at all. So, the students ability to use *all* of the functions stated in the book was not measured, and the amount covered on the test is disproportional to the content of the book.

In this chapter, the quantitative and qualitative analysis of the data collected were presented. The next chapter discusses and explains the results, describes limitations of the study and suggests some implications for curriculum development and points for further study.

CHAPTER FIVE : CONCLUSIONS

Overview of the Study

The aim of this study was to investigate the content validity of the midterm tests administered in the first semester at Anadolu University Foreign Languages Department (AUFLD). The content validity of the midterm tests of the core-course, grammar and listening courses was examined by collecting and analyzing data from multiple sources of information. The course instructors' views about the midterm tests' representativeness of the courses' content were elicited through questionnaires. The course materials' content and the tests' content were analyzed in terms of consistency between them. Also, the objectives for each of the three courses were analyzed for their consistency with the content areas the tests were used to measure.

This study aimed to find answers to the questions below:

a) What are the instructors' opinions about the midterm tests' representativeness of the courses' content?

b) To what extent do the sampled areas on the midterm tests and the course content areas correlate?

c) To what degree do the midterm tests' content and the course objectives correlate?

Results

The first research question was: What are the instructors' opinions about the midterm tests' representativeness of the courses' content?

The instructors in general thought the midterm tests' representativeness of the courses' content was moderate to high.

The questionnaire results showed that all of the four core-course instructors thought in general they had represented what they had taught to a high degree on

both of the midterm tests. The only difference between the two tests was that two of the instructors counted more function areas that should have been represented on the second midterm test. In their teaching, none of them had emphasized any of the areas specified in the course book over any other area. They all agreed that their tests had not included "function" and "listening" areas, although these areas were part of the material they were teaching during the first semester. But they had different answers about the matching of the exercise types to the task formats on the test. Two of the instructors said the matching was between "some" to "a lot", and the other two said there was little matching in terms of task types. The instructors who said "a little" added that this was because their test had focused on vocabulary and structural knowledge, while the book emphasized the language functions. Overall, their view of the matching between the course material content and the tests' content was positive.

For the grammar course, all of the three instructors agreed that the "Wh questions" topic was not included on the first midterm test and should have been represented more. For the second midterm test, two of the teachers agreed that "reported speech" was not included on the test and should have been. The other teacher said everything was included and fairly represented on the test. Although grammar instructors were not in 100% agreement over the representation of the given areas, they were in total agreement that the test content matched the teaching content and their course objectives well. But they also agreed that the exercises in their classrooms matched with the tests only to "some" degree. They all agreed that it was because of not using "multiple-choice" task types in their classes, though they heavily used that type on the midterm tests. This might be because the final

proficiency test in the FLD is a TOEFL-like, multiple-choice test, and the grammar course-coordinator stated one of their objectives was to prepare the students for TOEFL. So, they wanted to prepare the students for such multiple-choice test tasks. One other reason for their emphasizing the multiple-choice task format might be that besides its ease in scoring, this was the test material most readily available to them to use on the tests. Overall, their answers showed that they thought the match between their teaching and testing was positive.

The listening instructors were in total agreement on every question related to the first midterm test. They claimed that all of the listening areas were equally weighted and included on the test, and the match between the test and their teaching content and their objectives was very high. In the first listening midterm, there were two types of exercises: distinguishing phonemes at sentence level and individual word level. They also agreed that the test matched the exercises done in the classroom very well, which was found to be true, because the instructors had followed the same pattern of task types both in the material they used and the test they administered. But they had some reservations about the validity of the second midterm test. They pointed out many areas from the three sections (strategy, vocabulary, function) of the teaching content as missing on the test. Their answers for the "matching of teaching, objectives, and exercises" showed that they thought the second midterm test's validity was lower than the first one, but still fairly high. The lowest matching they reported was between the teaching content and the second midterm test content. Overall, the listening instructors thought that their midterm tests reflected the course content well enough.

The second research question was: To what extent do the sampled areas on the midterm tests and the course content areas correlate?

The analysis of the course content and the midterm tests' content showed that, other than the first grammar midterm test and the listening strategy areas of the second listening midterm test, the tests' sampling of the courses' content was disproportional and inadequate.

The analysis of the first core-course midterm test and material frequencies showed that for the first midterm, the function areas in the course book and the tested areas did not match at all, as reported by the instructors. What they tested as functions, that is, "Reading for overall comprehension" and "Reading for specific information" were not mentioned in the book. As for the other half of the course content, that is, "Forms," the most frequently tested area was "Vocabulary knowledge," which had zero frequency in the material. As a result, the correlation between the first midterm course material content and the test content was found to be significantly low. Also, the relationship between the formats students were required to be engaged in by the course book and the formats on the test were found to be unrelated. Almost none of the task formats on the test had been introduced to the students previously.

Though at first glance it looks positive, and more valid than the first core-course midterm test, the second one was found to be have no better validity than the first one. There were some of the functional areas sampled, but again the most frequent function that was tested was not from the couse book. The results of the second midterm "Form areas" showed that this area was even less representative of the material, because again the most frequently tested area was vocabulary, which

has zero frequency in the material. An analysis of the task format types revealed that the two most frequent task types in the material were not tested at all, and out of 55 different task types previously presented to students, only five were tested, and three of these were not highly frequent in the material. The sixth task type represented on the test, "Completing a dialogue according to the function given in parentheses," was not among the exercises in the material, and was newly introduced on the test. But, though the instructors were more hesitant about the exercises' matching of the test task types in their answers to the questionnaire, this area turned out to correlate the most positively. The most negatively significant correlations came from the "forms" area, which was peculiar, because it is usually easier to gear the distribution of form areas to the test than it is for the function areas.

The first grammar midterm test and material structure areas yielded a high positive correlation. This was mostly due to the testing of just one item: the past simple. This item was the most frequent both in the material and on the test. An analysis of the first grammar midterm test tasks showed a significant positive correlation with the exercises in the material, despite the fact that the most frequent task type on the test was absent from the material. High positive correlation of the tasks was a result of only one item (there were just two total on the test) that is, "Completing sentences by using the correct forms of the verbs given in parenthesis." The remaining 37 task types found in the material were not used on the test.

The second grammar midterm test and material structure areas' correlation was negative though not significant. That was because the less frequent areas in the material were asked much more frequently, and also one frequently tested area, "Mixed tenses," was not in the materials. For example, the material frequency of

"Quantifiers" is 2%, but it was 12 times more frequently tested on the test, with 24%. The task format types' correlation for the second midterm grammar test revealed a high positive correlation. But although positive, most of the task types in the materials were absent on the test. There were only two task types on the test, and only one of them matches well with the material. Overall, the second midterm test was found to have a lower representation of the course content than the first one.

The first listening midterm test was analyzed on two levels: consonants and vowels. Though not significant, a positive correlation was found between the consonant frequencies on the test and in the material. Vowels correlated slightly less positively than the consonants did. Here, though not stated in the syllabus, I found five extra vowel discrimination items in the materials, all of which are Turkish phonemes. They are missing on the test. The testing of most of the items was found to be not proportional to their frequencies in the material. The most striking material-test disproportions are: "Distinguishing between /t/ and /d/", with 4% frequency in the material and 16% frequency on the test, and "Distinguishing between /eə/ and /ıə/" with a frequency of 4% in the material and 27% on the test.

The second listening midterm test was analyzed in three parts: Listening strategy areas, vocabulary areas and functional areas. The first of these, the "listening strategy areas" correlation gave the highest positive results in the study. It does not mean that strategy items in the material were proportionally represented on the test. The main reson for the high positive match seem to be that "Listening for specific information" area was found to match both in the material and on the test as the most frequent item. As for the vocabulary items, the test represented the less frequent items in the material more. For example, "Answerphones" was the least

frequent item in the material, but the most frequent on the test. So, the vocabulary items correlated highly negatively. As for the functional areas, their representation on the test was very similar to that of the vocabulary items. Less frequent items were asked more on the test. For example, while "Making arrangements" had a frequency percentage of 4% in the materials, it made up 60% on the test. So, based on this analysis, the overall validity of the second listening midterm test can be said to be low, in spite of the fact that the listening strategy areas correlated well.

The third research question was: To what degree do the midterm tests' content and the course objectives correlate?

The course objectives stated by the course coordinators matched the tests' content to a very limited extent. Only one of the objectives for the first listening midterm test was specific enough and matched the test well. In fact, a comparison of the objectives to the tests' content was very difficult, because the given objectives were too general and obscure.

The core-course class was stated to have: success in the exchange of information, creating real-life situations with the help of the course book, and proceeding from comprehension through listening or reading to production through speaking or writing, as the objectives. When the content of the course book is analyzed, the activities in the book seem to be in line with the statements above, but the test content represents the same to a very limited extent. The tests from the core-course are mostly reading-, vocabulary-, and partly listening-centered. But, in the course book, listening and speaking skills were much more frequent than the others. Thus, the tests can be said to not adequately reflect the course objectives stated above. The tests do not test any ability of the students to cope with real-life

situations or successfully exchanging information. They can only be said to have tested the "the receptive skills of the students," and thus stayed at the stage of "comprehension," without "proceeding to production".

Objectives for the first semester grammar course were provided by the coordinator as: to have students give grammatically, situationally and semantically correct responses, to help students to understand and use English, to have students repeat structures that were taught in class, to have students use the grammatical structures in everyday English, to improve students' proficiency level so that they will be able to follow the courses in their majors, and to train students for the TOEFL. The course material analyzed does not reflect any of the objectives stated above, especially the "use" of the structures, because students were only required to provide the correct structures; they did not have to be meaningful, or use everyday English. Except for the multiple-choice question format on the tests, the material for the grammar course does not present any specific of preparation for a standard test.

Listening objectives were presented by the coordinator as: having students discriminate phonemes, and having students learn the functions in the book. The first objective can be said to have been well-represented by the first midterm material and test content, but the second objective, related to the functions in the book, seems to have not been followed much in the testing. Only two of the less frequent function areas, "making arrangements" and "narration" were tested on the second listening midterm test, and no reason was provided in the questionnaire for these items having been selected to be asked. In particular, the "narration" function may not be so important for the students preparing for academic studies in their respective fields.

Discussion

This section discusses the findings of the study and draws conclusions about the research questions outlined in Chapters 1- 4. References to other research findings in the literature are made at relevant points.

Some responses to the questionnaires seemed to conflict with what was found in the materials, and sometimes even among themselves. The core-course instructors were most positive about the match between their objectives and the test, but this is surprising because they do not seem to have set, agreed upon objectives. For example, their reasons for not including an area on the test rather than another varied to a great extent. One of the grammar instructors said she had emphasized "the contrast between tenses" in teaching and weighted it more on the test, but analysis of the tests shows that it was not tested until the second midterm test. One grammar instructor identified "tenses" as the area to have been weighted more on the test, but, what another instructor had responded, "active vs. passive," was weighted more. The reasons listening instructors gave for not including the areas they mentioned were also not consistent. Instructors either did not give a reason, or they gave different reasons for the exclusion of the same item. Some instructors gave the "difficulty in evaluation" as one of the reasons for not having asked a particular item. Putting in the test what is the easiest to assess, or simply by choosing the items for the test which are most readily available is not a good way of testing, and the validation for this type of tests is usually not possible, because they become arbitrary (Bachman & Palmer, 1996; Hughes, 1989).

Another interesting finding that came up during the analysis of the questionnaires was about the difficulty of the test items. The instructors reported not

asking the items they defined as easy, and tried to include the items that they expected the students to have more difficulty in dealing with, even if they had not emphasized these difficult areas in their teaching. For example, the grammar instructors had not tested their students' knowledge of the "Wh questions" and "Reported speech," because they said they thought they were too simple for the intermediate level. This problem of tending not to test the learning areas where students should be possible to do well can not be justified. In fact, inclusion of these supposedly simple items can improve the overall validity of an achievement test and might give teachers a better ground for making judgements about their students' learning (Brown, 1996; Hughes, 1989).

On both the grammar and listening tests, there were areas emphasized that were not especially problematic for learners. For example, on the first grammar midterm test, the emphasis on "the past simple" is hardly justifiable. The research on both Turkish (Aycan, 1990; Mergen, 1999; Şahin, 1993) and on other (Hinkle, 1992; Richards, 1979) L1 background students show that "the past simple tense" is not a great problem for the learners of English as a foreign language, but "the present perfect" is much bigger problem. While there is no tense in Turkish that corresponds to the English "present perfect," there is not much difference between Turkish simple past tense and English simple past (Aycan, 1990; Şahin, 1993). On the other hand, though so problematic for the Turkish learners of English, "the present perfect" was tested only one-third of what "the simple past" was. On the second grammar midterm test, as did the core-course teachers about theirs, grammar teachers had checked the match between task types on the test and in the materials as a bit lower

than the objectives-test match and material-test match, but it turns out that there is not much problem with the task types of either midterm test.

For the first listening midterm consonants area, justifying the more frequent test items could be difficult, because they all exist in Turkish. For example, for the consonants, the most frequently tested item was "the distinguishing between /s/ and /z/," but this area was not among the problematic phonemic areas for Turkish EFL learners, as reported by research findings (Bada, 1997; Kaçmaz, 1993). It is not the inclusion of these easy areas on the tests that is a problem, but their overemphasis.

Also, the correlation results for both of the listening tests do not affirm the listening instructors that the correlation between the course content and test content would match well. Even if they represented some of the areas on the test, as with the other courses, this representation seems to be disproportional. Listening strategy areas yielded the most positive results, and that was not because the listening instructors had included everything they covered in the classroom on their second listening midterm test. But there was a better proportion of the items than the other areas. Perhaps this might be a clue for the teachers who prepare their own tests to decide which points to highlight both in their teaching and testing, since it may not be practical and necessary to include every item covered in class on the test. But first, the most important items need to be decided on in the light of the course objectives, then covered on the test accordingly. Without having any clear objectives, the instructors participating in this study were also found (through the responses to the questionnaires) to be too dependent on their course books in their teaching and testing. For example, though they did not see the point in teaching it,

the listening instructors said they had taught "Listening against background noise" because it was in the book.

Even the course coordinators consulted the course books to vocalize their course objectives during the interviews. The apparent gap between the stated objectives and the tests' and materials' content might be accounted for by their being retrospectively produced and too general to match with the specific course contents.

Overall, the results collected from the three sources; questionnaires, material and tests, and interviews were conflicting. Only the results from the questionnaires were positive about the tests' validity. The other two sources indicated a much lower validity of the midterm tests. Lack of clearly defined testing criteria and course objectives seems to be the main factor causing such a conflict among the results

Implications

The results suggest that even teachers at the same level, teaching the same course usually had different justifications for sampling one item over another. Making the tests together did not solve the problem. But perhaps if the teachers had negotiated and decided about the items to be included on their tests earlier by looking at their course materials, this problem could be overcome. Determining the test items would inevitably require them to be clear about what they want from their students to perform as a result of their teaching. That means, to have tests which have higher validity, testers need to define their objectives first. Then the items on the test should be directly related to these objectives (Hughes, 1989). Actually, teachers seemed to have some objectives in mind during the test design, but they had not recorded them in any way. But as (Brown, 1995, 1996) suggests, it is very

important that the objectives be written, so that they can be referred to and revised continuously.

The results of this study also showed that there is a big gap between the weighting of the items in the material and on the test. Well-defined objectives would act like a bridge, and the gap between the teaching and testing would be much smaller. Having clear, well-defined objectives helps teachers to teach and test their students better (Brown,1995, 1996; Hughes, 1989). Actually it is a cycling process; testing helps teachers to refine their objectives through enabling them to see which points their students have difficulty with and need to be emphasized more in students' learning, and having clearer, more focused objectives helps testers to design better tests, for example, by weighting the items in parallel with the primary objectives of the program (Brown, 1995; Graves, 2000). Clear objectives provide the criteria for teachers in deciding which language content areas to weight on the test over the others.

More valid testing also requires having test specifications agreed upon by all the teachers involved in the testing of a particular course (Alderson, Clapham, & Wall, 1995). By having test specifications, teachers are able to know and be confident about what they are testing and why they are testing it. Having this starting point, then they can constantly revise these specifications by taking the needs of the students into account. A test specifications document, giving the details of how and on what to test learners, is the primary source of data for any content validation. Eventually, what a validator needs is written data. Even if a test's validity is high, testers still may want to better it; then, they need to have specifications for

their test in a written form. So, training the instructors about defining their course objectives and designing their own test specifications seems necessary.

One further implication might be that since the instructors in the FLD do not seem to have enough time to spend on testing, and they devote most of their time to teaching, a testing office focusing just on test preparation could work towards more validity by bringing a conformity among the tests at different proficiency levels, and help produce a better representation of the teaching content on the tests in the FLD. Since testing requires a certain degree of specialization, the testing office workers could develop that kind of specialization by specifying the content domain and improving tests over and over by taking the particular testing needs in their particular teaching context into account (which, in this case, is the FLD context). Although some might feel that a testing office will be too distant from the classroom, with good coordination among teachers and testers this problem can be overcome (Brown, 1995).

In short, in order to have more valid achievement tests, objectives, teaching, and testing should all be closely related (Bachman,1991; Brown, 1995; Davies,1990; Hughes,1989). It follows that valid language testing starts at the program level and requires the improvement of curriculum design (Brown,1995).

Limitations of the Study

The biggest difficulty I had in doing this study was the unavailability of any test specifications for the courses. Especially for the reading and writing courses, there were no written data of the course content available other than the tests, so they had to be discarded. Since the available objectives were too convoluted and not refined enough, they were not suitable for a content validation either. That

left me with only the option of getting my hands on any kind of data about the content of the courses available otherwise. In addition, I was only able to cover the first semester of the courses in my department, and perhaps a study covering a longer period would have produced different results.

Further Research

Further research can be done on an in-depth analysis of the achievement test validity of a particular course. Testers may be given forms containing the course content areas to put them into rank order, then those rankings can be compared with the rank-order of the frequencies of the same items on the test. Content validity of speaking and writing achievement tests can be investigated, which was not able to be done in this study. The relationship between the patterns in teachers' test preparation and validity can be investigated. Another correlational study may investigate the relationship between students' academic English needs and weighting of the language test items in terms of those needs. Finally, students' perceptions of certain achievement tests' face validity can be triangulated with the testers' perceptions of the same tests' validity and the statistical analysis results of their content validity.

REFERENCES

Alderson, J. C. (1988). Testing English for specific purposes: How scientific can we get? In A. Hughes (Ed.), *Testing English for University Study* (pp.16 – 28). ELT Documents: 127. Modern English Publications.

Alderson, J. C., & Clapham, C. (1995). Assessing Student Performance in the ESL Classroom. *TESOL Quarterly, 29,* 184-186.

Alderson, J. C., & Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Anastasi, A. (1988). *Psychological Testing*. London: Macmillan.

Aycan, N. (1990). *Errors in Tense in the Written English of Turkish Students with Special Reference to the Semantic Bases*. Unpublished Master's Thesis. Bilkent University, Ankara.

Bachman, L. F. (1989). The development and langauge 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. (1991a). *Fundamental Considerations in Language Testing*. Hong Kong: Oxford University Press.

Bachman, L. F. (1991b). What Does Language Testing Have to Offer?. *TESOL Quarterly, 25,* 671 – 700.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice*. UK: Oxford University Press.

Bada, E. (1997). Teacher and learner awareness of the voice factor in English sounds. In *Teacher Factor in Foreign Language Classes*. Turkey: TÖMER Gaziantep Branch, 66 – 74.

Basanta, C. P. (1995). Coming to Grips with Progress Testing: Some Guidelines for Its Design. *Forum, 33,* 55 – 64.

Bejar, I. (1983). *Achievement Testing: Recent Advances*. CA: Sage Publications.

Bond, L. (1996). Norm-and Criterion-Referenced Testing. Available: Online *ERIC Document Reproduction Service, ERIC Digest* No: ED 410316.

Brown, H. D. (1994). *Principles of Language Learning and Teaching*. New Jersey: Prentice Hall Regents.

Brown, J. D. (1989). Improving ESL Placement Tests: Using Two Perspectives. *TESOL Quarterly, 23,* 65 - 83.

Brown, J. D. (1995). *The Elements of Language Curriculum*. Boston: Heinle & Heinle Publishers.

Brown, J. D. (1996). *Testing in Language Programs*. NJ: Prentice Hall Regents.

Carey, L. M. (1988). *Measuring and Evaluating School Learning*. Newton, Mass: Allyn and Bacon, Inc.

Childs, R. A. (1989). Constructing Classroom Achievement Tests. *ERIC Digest*. Available: http://www.ed.gov/databases/ERIC_Digests/ed 315426.html

Davies, A. (1988). Procedures in language test validation. In A. Hughes (Ed.), *Testing English for University Study* (pp. 29 – 35). ELT Documents:127. Modern English Publications.

Davies, A. (1990). *Principles of Language Testing*. Padstow, Cornwall: Basil Blackwell Ltd.

Doff, A., & Becket, C. (1991) *Listening 2*. Cambridge: Cambridge University Press.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Graves, K. (2000). *Designing Language Courses*. Boston, MA: Heinle & Heinle Publishers.

Hamp-Lyons, L. (1996). Second language writing: Assessment Issues. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp.69 – 87). Cambridge: Cambridge University Press.

Heaton, J.B. (1988). *Writing English Language Tests*. Hong Kong: Longman Group UK Ltd.

Heaton, J.B. (1990). *Classroom Testing*. N.Y: Longman Group UK Ltd.

Henning, G. (1987). *A Guide to Language Testing*. Cambridge, Mass: Newbury House Publishers.

Hills, J. R. (1976). *Measurement and Evaluation in the Classroom*. Columbus, OH: Charles E. Merrill Publishing Company.

Hinkle, E. (1992). L2 Tense and Time Reference. *TESOL Quarterly, 26,* 557- 572.

Hudson, T., & Lynch, B. (1984). A Criterion-referenced Measurement Approach to ESL Achievement Testing. *Language Testing, 1,* 171 – 201.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Ingram, D. E. (1985). Assessing proficiency: An overview on some aspects of testing. In K. Hyltenstam, & M. Pienemann (Eds.), *Modelling and Assessing Second Language Acquisition* (pp.215 – 276). UK: Multilingual Matters Ltd.

Johnson, K., & Johnson, H. (Eds.). (1998). *Encylopedic Dictionary of Applied Linguistics*. Oxford: Blackwell Publishers Ltd.

Jordan, R. R. (1997). *English for Academic Purposes*. Cambridge: Cambridge University Press.

Kaçmaz, T. (1993). *An Analysis of the Pronunciation Problems of Turkish Learners of English*. Unpublished Master's Thesis. Bilkent University, Ankara.

Landy, F. J. (1987). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41,* 1183 – 1192.

Lynch, B. K. & Davidson, F. (1994). Criterion-Referenced Language Test Development: Linking Curricula, Teachers, and Tests. *TESOL Quarterly, 28,* 727 – 743.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

Mergen, G. (1999). *An Analysis of Turkish EFL Students' Errors in Present Perfect Tenses.* Unpublished Master's Thesis. Bilkent University, Ankara.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Waimer & H. Brown (Eds.), *Test Validity*. Hillsdale, NJ: Erlbaum.

Murphy, K. R. & Davidshofer, C.O. (1991). *Psychological Testing: Principles & Applications*. Englewood Cliffs, N.J: Prentice-Hall International, Inc.

Picket, D. (1984). *Authenticity in Foreign Language Testing*. S. Anivan (Ed.) Singapore: SEAMO Anthology Series.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Richards, J.C. (1979). Introducing the Perfect: An Exercise in Pedagogic Grammar. *TESOL Quarterly, 13,* 495 – 501.

Rudner, L. M. (1994). Questions to Ask When Evaluating Tests. Available: (Online) *ERIC Document Reproduction Service*. Doc. No: ED 385607.

Shohamy, E. (1984). Does the Testing Method Make a Difference? The Case of Reading Comprehension. *Language Testing, 1,* 147 – 170.

Spolsky, B. (1968). Language Testing: The Problem of Validation. *TESOL Quarterly, 2,* 88 – 94.

Spolsky, B. (1978). Linguists and language testers. In B. Spolsky (Ed.), *Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics.

Spolsky, B. (1985). The Limits of Authenticity in Language Testing. *Language Testing, 2,* 31 – 41.

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

Stevenson, D. K. (1981). Beyond faith and face validity: The multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests. In A.S Palmer, & P.J.M. Groot, & G.A. Trosper (Eds.), *The Construct Validation of Tests of Communicative Competence* (pp.37 – 61). Georgetown University, Washington, D.C: TESOL.

Şahin, M. K. (1993). *Error Analysis of Tense and Aspect in the Written English of Turkish Students*. Unpublished Master's Thesis. Bilkent University, Ankara.

Teasdale, A. (1996). Content validity in tests for well-defined LSP domains: An approach to defining what is to be tested. In M. Milanovic, & N. Saville (Eds.), *Studies in Language Testing 3* (pp.211 – 230). Cambridge: Cambridge University Press.

Underhill, N. (1982). The great reliability validity trade-off: Problems in assessing the productive skills. In B. Heaton (Ed.), *Language Testing* (pp.17 – 23). UK: Modern English Publications Ltd.

Warshawsky, D., Byrd, D.R.H. (1994). *Spectrum 4*. Upper Saddle River, NJ: Prentice Hall Regents.

Weir, C. J. (1990). *Communicative Language Testing*. Trowbridge, UK: Prentice Hall International (UK) Ltd.

Weir, C. J. (1995). *Understanding & Developing Language Tests*. UK: Phoenix ELT

APPENDIX A

TEST CONTENT VALIDITY RESEARCH QUESTIONNAIRE

Dear Colleagues,

I am an MA TEFL graduate student at Bilkent University. I am doing a research on the content validity of the first semester midterm tests administered at Anadolu University Foreign Languages Department. I am interested in your opinions about the tests. Your responses will help me a great deal with my research. Your responses will be kept confidential. You do not have to give your name and no one will know your specific answers to these questions. I would be grateful if you would take a few moments to complete the questions.

Thank you,

Harun Serpil

QUESTIONNAIRE FOR THE CORE-COURSE TEACHERS

1. How long have you been teaching?
   a) 1-5 years
   b) 6-10 years
   c) 11-15 years
   d) 16-20 years

2. How long have you been teaching at Anadolu University?
   a) 1-5 years
   b) 6-10 years
   c) 11-15 years
   d) 15-20 years

Questions about the first midterm:

3. The following areas are taken from your course material. Which "function" area(s) of core-course content should have been represented more on the first midterm test? Choose from below.

   a) Asking for and giving an opinion
   b) Expressing surprise
   c) Agreeing and disagreeing with an opinion
   d) Confirming information
   e) Talking about likes and dislikes
   f) Stating a preference
   g) Talking about habits
   h) Complaining
   i) Expressing anticipation
   j) Talking about strengths and weaknesses
   k) Asking about and confirming facts
   l) Talking about famous landmarks and famous works
   m) Reacting to news
   n) Talking about needs
   o) Asking a friend for information
   p) Talking about the content of something
   q) Talking about your class or group
   r) Making a decision with someone
   s) Talking about exceptions
   t) Reporting a conversation
   u) Making plans
   v) Discussing plans
   w) Telling about a past experience
   x) Expressing duration
   y) Resuming a conversation
   z) Talking about your life
   aa) Expressing a wish
   bb) Talking about imaginary situations
   cc) Talking about where products are made or produced

4. Which "form" area(s) of core-course content should have been represented more on the first midterm test? Choose from below.

   a) The superlative of adjectives with the present perfect
   b) Rejoinders showing interest or surprise
   c) Tag questions
   d) Be used to
   e) *So* vs. *such*
   f) Gerunds
   g) The passive in the past tense
   h) The passive without an agent
   i) Relative clauses with *who* and *whose*
   j) Indefinite compounds with *except* and *but*
   k) *The whole*, *all*, and *every* with time expressions
   l) *While* vs. *during*
   m) Contrary-to-fact conditionals
   n) The passive in the present and past

5. Were there any content area(s) on the first core-course midterm test that you did not include on the test although you taught that content in the class?
   a) If no, skip to question 7.
   b) If yes, please state those area(s) below.

6. What were your reason(s) for not including that core-course content on the first midterm test? (e.g. you may have asked it on the quiz)

7. Before the first midterm test, did you emphasize some area(s) of the core-course content in teaching over other areas?
   a) If no, skip to question 9.
   b) If yes, what were those area(s)?

8. Were you able to emphasize those area(s) on the first midterm test too?

   a) Yes b) No

9. Did you think some core-course areas were more important than the others when you were designing your test, so you weighted them more on the first midterm test?
   a) If no, skip to question 10.
   b) If yes, please state those areas.

10. To what extent do you think the content of the core-course questions on the first midterm test matched the core-course content you taught in your class?

    a) very little b) a little c) some d) a lot

11. To what extent do you think the questions on the first midterm test matched your course objectives?

    a) very little b) a little c) some d) a lot

12. To what extent do you think the exercise types you used in the classroom matched the question types on the first midterm test?

    a) very little b) a little c) some d) a lot

Questions about the second midterm:

13. The following areas are taken from your course material. Which "function" area(s) of core-course content should have been represented more on the second midterm test? Choose from below.
    a) Offering to do someone a favor
    b) Asking a favor of someone
    c) Having something done
    d) Making a request
    e) Going shopping
    f) Telling about a past experience
    g) Telling a frightening story
    h) Apologizing
    i) Starting a conversation
    j) Asking for a reason
    k) Refreshing your memory
    l) Cancelling plans
    m) Giving an excuse
    n) Asking permission
    o) Agreeing or refusing to do a favor
    p) Saying thank you
    q) Discussing arrangements and obligations
    r) Stating a condition
    s) Stating a prohibition

t) Describing an injury
u) Reporting requests and instructions
v) Talking about problems
w) Showing concern
x) Talking about moods and feelings
y) Giving warnings and advice
z) Greeting someone after a long time
aa) Catching up on what someone's been doing
bb) Discussing possible plans
cc) Imagining something
dd) Giving unwanted advice
ee) Asking for information
ff) Asking for travel information
gg) Correcting a mistaken impression
hh) Talking about food
ii) Making an objection
jj) Expressing regret
kk) Making a judgement

14. Which "form" area(s) of core-course content should have been represented more on the second midterm test? Choose from below.

a) The causative "have"
b) Noun compounds
c) The simple past vs. past continuous with *when* and *while*
d) Negative questions
e) Infinitives after direct objects
f) Had better and had better not
g) Be supposed to
h) If vs. unless
i) Be supposed to vs. have to in negative statements
j) Reported speech with imperatives
k) The present perfect continuous
l) Reflexive pronouns
m) The present perfect continuous vs. the present perfect
n) Conditional sentences
o) Embedded yes-no questions
p) Sequence of tenses
q) Too...to and not...enough...to
r) Should have

15. Were there any content area(s) on the second core-course midterm test that you did not include on the test although you taught that content in the class?
a) If no, skip to question 17.
b) If yes, please state those area(s) below.

16. What were your reason(s) for not including that core-course content on the second midterm test? (e.g. you may have asked it on the quiz)

17. Before the second midterm test, did you emphasize some area(s) of the core-course content in teaching over other areas?
    a) If no, skip to question 19.
    b) If yes, what were those area(s)?

18. Were you able to emphasize those area(s) on the second midterm test too?

    a) Yes b) No

19. Did you think some core-course areas were more important than the others when you were designing your test, so you weighted them more on the second midterm test?
    a) If no, skip to question 20.
    b) If yes, please state those areas.

20. To what extent do you think the content of the core-course questions on the second midterm test matched the core-course content you taught in your class?

    a) very little b) a little c) some d) a lot

21. To what extent do you think the questions on the second midterm test matched your course objectives?

    a) very little b) a little c) some d) a lot

22. To what extent do you think the exercise types you used in the classroom matched the question types on the second midterm test?

    a) very little b) a little c) some d) a lot

APPENDIX B

QUESTIONNAIRE FOR THE GRAMMAR TEACHERS

1. How long have you been teaching?
   c) 1-5 years c) 11-15 years
   d) 6-10 years d) 16-20 years

2. How long have you been teaching at Anadolu University?
   c) 1-5 years c) 11-15 years
   d) 6-10 years d) 15-20 years

Questions about the first midterm:

3. The following areas are taken from your course material.Which area(s) of grammar course content should have been represented more on the first midterm test? Choose from below.

   a) Simple present tense
   b) Present continuous tense
   c) Simple Present tense vs. Present continuous tense
   d) Simple past tense
   e) Used to
   f) Past continuous tense
   g) Wh- questions
   h) Present continuous tense and Simple present tense with future meaning
   i) Present perfect tense
   j) For, since, yet, already
   k) Present perfect tense vs. Simple past tense
   l) Present perfect continuous tense
   m) Present perfect continuous tense vs. Present perfect tense
   n) Future perfect tense
   o) Future continuous tense
   p) Past perfect tense
   q) Past perfect continuous tense

4. Were there any content area(s) on the first grammar midterm test that you did not include on the test although you taught that content in the class?
   a) If no, skip to question 6.
   b) If yes, please state those area(s) below.

5. What were your reason(s) for not including that grammar course content on the first midterm test? (e.g. you may have asked it on the quiz)

6. Before the first midterm test, did you emphasize some area(s) of the grammar course content in teaching over other areas?
   a) If no, skip to question 8.
   b) If yes, what were those area(s)?

7. Were you able to emphasize those area(s) on the first midterm test too?

   a) Yes b) No

8. Did you think some grammar course areas were more important than the others when you were designing your test, so you weighted them more on the first midterm test?
   c) If no, skip to question 9.
   d) If yes, please state those areas.

9. To what extent do you think the content of the grammar questions on the first midterm test matched the grammar content you taught in your class?

   a) very little b) a little c) some d) a lot

10. To what extent do you think the questions on the first midterm test matched your course objectives?

   a) very little b) a little c) some d) a lot

11. To what extent do you think the exercise types you used in the classroom matched the question types on the first midterm test?

a) very little b) a little c) some d) a lot

Questions about the second midterm:

12. The following areas are taken from your course material. Which area(s) of grammar course content should have been represented more on the second midterm test? Choose from below.

a) Adjectives
b) Adverbs
c) Equatives (as...as)
d) Comparatives
e) Superlatives
f) Some-Any-No
g) Quantifiers
h) Articles/ Definite-Indefinite
i) Modals (I) (Present Modals)
j) Modals (II) (Perfect Modals)
k) The Passive
l) It is said that.../He is said to.../...is supposed to...
m) Reported speech
n) Compound sentences
o) Either-or/ Neither-nor/Both-and/Not only-but also

13. Were there any content area(s) on the second grammar midterm test that you did not include on the test although you taught that content in the class?
a) If no, skip to question 15.
b) If yes, please state those area(s) below.

14. What were your reason(s) for not including that grammar course content on the second midterm test? (e.g. you may have asked it on the quiz)

15. Before the second midterm test, did you emphasize some area(s) of the grammar course content in teaching over other areas?
    a) If no, skip to question 17.
    b) If yes, what were those area(s)?

16. Were you able to emphasize those area(s) on the second midterm test too?

    a) Yes b) No

17. Did you think some grammar areas were more important than the others when you were designing your test, so you weighted them more on the second midterm test?
    a) If no, skip to question 18.
    b) If yes, please state those areas.

18. To what extent do you think the content of the grammar questions on the second midterm test matched the grammar content you taught in your class?

    a) very little b) a little c) some d) a lot

19. To what extent do you think the questions on the second midterm test matched your course objectives?

    a) very little b) a little c) some d) a lot

20. To what extent do you think the exercise types you used in the classroom matched the question types on the second midterm test?

    a) very little b) a little c) some d) a lot

APPENDIX C

QUESTIONNAIRE FOR THE LISTENING TEACHERS

1. How long have you been teaching?
   e) 1-5 years c) 11-15 years
   f) 6-10 years d) 16-20 years

2. How long have you been teaching at Anadolu University?
   r) 1-5 years c) 11-15 years
   s) 6-10 years d) 15-20 years

Questions about the first midterm:

3. The following areas are taken from your course material. Which area(s) of listening course content should have been represented more on the first midterm test? Choose from below.
   a) Identifying the correct words in a sentence on the basis of phoneme differentiation
   b) Identifying individual consonants in words:
   b.1- Distinguishing between / s / and / ʃ /
   b.2- Distinguishing between / ʃ / and / tʃ /
   b.3- Distinguishing between / tʃ / , / dʒ / and / ʒ /
   b.4- Distinguishing between / j / and / dʒ /
   b.5- Distinguishing between / s / and / z /
   b.6- Distinguishing between / h / and no / h /
   b.7- Distinguishing between / p / and / b /
   b.8- Distinguishing between / t / and / d /
   b.9- Distinguishing between / k / and / g /
   b.10-Distinguishing between / l / and / r /
   b.11-Distinguishing between / l / and / r / and / n /
   b.12-Distinguishing between / p / and / v / and / f /
   b.13-Distinguishing between / b / and / v / and / w /
   b.14-Distinguishing between / θ / and / f /
   b.15-Distinguishing between / θ / , / ð / , / s / and / z /
   b.16-Distinguishing between / θ / , / ð / , / t / and / d /
   b.17-Distinguishing between / n / and / /
   b.18-Saying / m /
   c) *Identifying individual vowels in words*:
   c.1- Distinguishing between / i: / and / ɪ /
   c.2- Distinguishing between / ɪ / and / e /
   c.3- Distinguishing between / æ / and / e /
   c.4- Distinguishing between / æ / and / ^ /
   c.5- Distinguishing between / ɜ: / and / a: /
   c.6- Distinguishing between / əʊ / and / o: /
   c.7- Distinguishing between / u: / and / ʊ /
   c.8- Distinguishing between / ɒ / and / o: /
   c.9- Distinguishing between / eɪ / and / e /

c.10- Distinguishing between / eə/ and /ɪə/
c.11- Saying / aɪ /
c.12- Saying / oɪ /
c.13- Saying / aʊ /

4. Were there any content area(s) on the first listening midterm test that you did not include on the test although you taught that content in the class?
a) If no, skip to question 6.
b) If yes, please state those area(s) below.

5. What were your reason(s) for not including that listening course content on the first midterm test? (e.g. you may have asked it on the quiz)

6. Before the first midterm test, did you emphasize some area(s) of the listening course content in teaching over other areas?
a) If no, skip to question 8.
b) If yes, what were those area(s)?

7. Were you able to emphasize those area(s) on the first midterm test too?

a) Yes b) No

8. Did you think some listening course areas were more important than the others when you were designing your test, so you weighted them more on the first midterm test?
e) If no, skip to question 9.
f) If yes, please state those areas.

9. To what extent do you think the content of the listening questions on the first midterm test matched the listening content you taught in your class?

a) very little b) a little c) some d) a lot

10. To what extent do you think the questions on the first midterm test matched your course objectives?

a) very little b) a little c) some d) a lot

11. To what extent do you think the exercise types you used in the classroom matched the question types on the first midterm test?

a) very little b) a little c) some d) a lot

Questions about the second midterm:

12. The following areas are taken from your course material. Which listening strategy area(s) of the listening course content should have been represented more on the second midterm test? Choose from below.
    a) Forming a mental picture
    b) Preparing an appropriate response
    c) Following a story; predicting and guessing
    d) Following a story; following an explanation
    e) Listening against background noise
    f) Listening for specific information
    g) Matching with your own interpretation
    h) Following an explanation; matching with your own beliefs

13. Which vocabulary area(s) of the listening course content should have been represented more on the second midterm test? Choose from below.
    a) Music
    b) Buying and selling; food and drink
    c) Visitors; dogs; shops; embarrasment
    d) Travel; public places; parties
    e) Cities; landscapes
    f) Answerphones
    g) Fears and phobias; crime
    h) Childhood; accidents; games
    i) Buying and selling; household objects
    j) The past; homes; art
    k) Superstitions; fortune telling

14. Which functional area(s) of the listening course content should have been represented more on the second midterm test? Choose from below.
    a) Imagining scenes; remembering scenes
    b) Requests; agreeing and refusing; asking personal information
    c) Narration
    d) Making enquiries; giving directions; 'small talk'
    e) Describing scenes
    f) Making arrangements

g) Describing a process
h) Describing objects
i) Interpreting a picture; matching information with pictures
j) Giving explanations; making deductions

15. Were there any content area(s) on the second listening midterm test that you did not include on the test although you taught that content in the class?
    a) If no, skip to question 17.
    b) If yes, please state those area(s) below.

16. What were your reason(s) for not including that listening course content on the second midterm test? (e.g. you may have asked it on the quiz)

17. Before the second midterm test, did you emphasize some area(s) of the listening course content in teaching over other areas?
    a) If no, skip to question 19.
    b) If yes, what were those area(s)?

18. Were you able to emphasize those area(s) on the second midterm test too?

    a) Yes b) No

19. Did you think some listening areas were more important than the others when you were designing your test, so you weighted them more on the second midterm test?

    a) If no, skip to question 20.
    b) If yes, please state those areas.

20. To what extent do you think the content of the listening questions on the second midterm test matched the listening content you taught in your class?

    a) very little b) a little c) some d) a lot

21. To what extent do you think the questions on the second midterm test matched your course objectives?

a) very little b) a little c) some d) a lot

22. To what extent do you think the exercise types you used in the classroom matched the question types on the second midterm test?

a) very little b) a little c) some d) a lot

APPENDIX D

CORE-COURSE OBJECTIVES TRANSCRIPT

- What were your objectives for the first semester's "core-course" at the intermediate level?
- *Core–course Coordinator*:

In the first term, Spectrum 4 was used for intermediate level students. And our course book provided practice in all poor communication skills. And it had special focus on listening and speaking aspects. The success in the exchange of information was one of our objectives. Another objective was trying to create a real-life situation with the help of our course book. One of our most important objective was... proceeding from comprehending language through listening or reading to producing it through speaking or writing. But some of the functions, and structures in the conversations...in the units, were RECEPTIVE and they didn't become productive until later units or levels. Related to this fact, our quizzes and midterms were mostly based on testing of the functions taught during the classes. Some technical problems also prevented us from using listening PARTS in our exams, although our book is mostly based on listening activities. Thanks.

APPENDIX E

GRAMMAR COURSE OBJECTIVES TRANSCRIPT

- What were your objectives for the first semester's grammar course at the intermediate level?
- *Grammar Course Coordinator*:

  OK. Students are generally expected to give grammatically correct responses as well as semantically and situationally correct ones. So, in our... in the first semester, the aim of our intermediate grammar course is...uhm to help students to understand and use English... through supplementary materials which were prepared accordingly, and their grammar book, this is Focus on Grammar , intermediate one,...umm that contains speaking, reading and writing activities. Some of them were given as homework in order to make the students REPEAT those structures that were taught in class. Uhh... we had both controlled and communicative exercises, because we wanted the students to use the grammatical structures in everyday English, in usage.Uhm... furthermore, we wanted to improve their proficiency level to follow the English instructions and materials in their majors next year in their faculties. We also aimed to train them for the proficiency exam, such as TOEFL, uhm which will be given at the end of the Intensive English Program. That's it, I think.
- Thanks.
- Not at all.

APPENDIX F

LISTENING COURSE OBJECTIVES TRANSCRIPT

- What were your objectives for the first semester's listening course at the intermediate level, please?
- *Listening Course Coordinator*:
  Our objective was to teach sound discrimination. We taught consonants and vowels by describing the sounds themselves first. Then the students heard words having similar sounds in a sentence. They tried to distinguish the correct sound, to decide which word they have heard. They also heard short dialogues and passages having words with similar sounds, and they were to distinguish the correct sound. In the first midterm, we prepared the questions the same way we we practiced in class, for example, they heard... a sentence saying, uh..."There are dirty chairs in the room." And they had two alternatives: a) dirty, b) thirty. So, they were asked to distinguish between /∂/ and /θ/ sounds. After finishing teaching and testing sound discrimination, the syllabus included the units, from books selected for every level, we covered the functions suggested by the book. And in the second midterm, for...levels beginner, elementary, lower-intermediate, intermediate, and upper-intermediate, there were three different exams. Uhh... one for beginner and elementaries, the other for lower-intermediate levels and the third, intermediates and upper-intermediates. While preparing the EXAMS... for the three different levels, we took the FUNCTION suggested in the books into consideration. But of course we CHOSE some functions as we asked just two parts of questions in the midterm. And the format of the questions, was similar to the ones we DID in class, I mean the ones in the book. Charts, comprehension questions, true-false statements, etc.
- Thanks.