AN ASSESSMENT OF THE VALIDITY OF THE MIDTERM AND THE END OF COURSE ASSESSMENT TESTS ADMINISTERED AT NACETTEPE UNIVERSITY, DEPARTMENT OF BASIC ENGLISH

A THESIS PRESENTED BY
HILAL ÖZKAN
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 1999
AN ASSESSMENT OF THE VALIDITY OF THE MIDTERM AND THE
END OF COURSE ASSESSMENT TESTS ADMINISTERED AT
HACETTEPE UNIVERSITY, DEPARTMENT OF BASIC ENGLISH

A THESIS PRESENTED BY
HİLAL ÖŞKEN
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, 1999

Hilal Öşken
ABSTRACT

Title: An Assessment of the Validity of the Midterm and the End of Course Assessment Tests at Hacettepe University, Department of Basic English.

Author: Hilal Ösken

Thesis Chairperson: Dr. William Snyder
Bilkent University, MA TEFL Program

Committee Members: Dr. Necmi Akşit
Michele Rajotte
Bilkent University, MA TEFL Program

This study investigated the validity of tests administered at Hacettepe University, Department of Basic English (DBE). The predictive validity of six midterm tests conducted throughout an academic year was examined in terms of students’ test scores on each midterm test and on the end of course assessment test. The consistency between the end of course assessment test and the course objectives was investigated to determine the content validity of the end of course assessment test.

The differences between midterm tests and the end of course assessment test in terms of their formats and content necessitated a validation study of those tests. The differences between students from different faculties’ test scores also necessitated such a study.

In this study the subjects were 510 C-level (beginner) preparatory students and 34 English instructors at DBE at Hacettepe University.

To gather data for this study, questionnaires were given to 34 English instructors to get their opinions about the course content and about the content of the end of course assessment test. Apart from that, all C-level students’ six midterm and the end of course
assessment test scores (3570 scores) were used to examine the predictive validity of the
tests.

Data from questionnaires were analyzed using frequencies and percentages and
the results were shown in tables. For the comparison of the test scores between
midterm tests and the end of course assessment test Pearson Product Moment
Correlation Coefficient was used. To examine the differences between test scores of
students from different faculties a One-way ANOVA was run. The frequencies of
the course objectives (grammar and reading) were compared with the test items in the
end of course assessment and the results were shown in tables.

The results of the questionnaires indicated that according to the instructors of
DBE, only the grammar and vocabulary parts of the end of course assessment test
represent the course objectives. Regarding listening, speaking and writing skills,
results indicated that instructors wanted these skills to be tested in the end of course
assessment test. The results of the comparison of the test scores indicated that the
degree of overall predictive validity was very high ranging from $r = .90$ to $r = .72$ at
$p < .01$ significance level. However, predictive validity of tests results for students
from different faculties was not always significant. It varied from $r = .90$ to $r = - .35$.
The results of the comparison of the course objectives with the test indicated that the
test items were not constructed according to their frequencies in the course books.

The findings suggest that the end of course assessment test administered at DBE
at Hacettepe University should be revised to include writing, listening and speaking
since these are parts of the course objectives. Another suggestion is that the course
content of DBE should be revised with regard to time spent on language skills and be
reorganized in order to be equally fair to the students from all faculties.
BILKENT UNIVERSITY
INSTITUTE OF ECONOMICS AND SOCIAL SCIENCES
MA THESIS EXAMINATION RESULT FORM
July 31, 1999

The examining committee appointed by the Institute of Economics and Social
Sciences
for the thesis examination of the MA TEFL student
Hilal Ösken
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 validity of the Midterm and the End of Course
Assessment Tests Administered at Hacettepe University, Department
of Basic English.

Thesis Advisor: Dr. Patricia Sullivan
Bilkent University, MA TEFL Program

Committee Members: Dr. William E. Snyder
Bilkent University, MA TEFL Program

Dr. Necmi Akpit
Bilkent University, MA TEFL Program

Michele Rajotte
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. Patricia Sullivan  
(Advisor)

Dr. William E. Snyder  
(Committee Member)

Dr. Necmi Akşit  
(Committee Member)

Michele Rajotte  
(Committee Member)

Approved for

Institute of Economics and Social Sciences

Ali Karaosmanoğlu  
Director
Institute of Economics and Social Sciences
# TABLE OF CONTENTS

LIST OF FIGURES........................................................................................................ viii

LIST OF TABLES........................................................................................................ ix

CHAPTER 1 INTRODUCTION........................................................ .................. 1
  Introduction to the Study....................................................... 1
  Background of the Study ................................................ 2
  Statement of the Problem.............................................. 4
  Purpose of the Study.................................................. 5
  Significance of the Study.................................................. 5
  Research Questions....................................................... 6

CHAPTER 2 LITERATURE REVIEW......................................................... 7
  Definitions and Approaches to Language Testing........ 7
  Types and Purposes of Language Tests........................ 10
    Proficiency Tests................................................... 11
    Placement Tests.................................................... 11
    Achievement Tests................................................. 11
    Diagnostic Tests................................................... 12
  Evaluating Language Tests......................................... 13
    Reliability............................................................... 13
    Validity................................................................. 15
      Content Validity............................................. 16
      Predictive Validity........................................ 18
      Construct Validity............................................ 18
      Concurrent Validity........................................ 20

CHAPTER 3 METHODOLOGY................................................................. 22
  Introduction................................................................... 22
  Subjects......................................................................... 22
    Preparatory Class Students........................................ 22
    English Language Instructors.................................. 23
  Materials....................................................................... 23
    Questionnaires................................................... 23
    Course Books...................................................... 24
    The End of Course Assessment Test......................... 24
    Test Scores...................................................... 24
  Procedures.................................................................... 25
  Data Analysis............................................................... 25

CHAPTER 4 DATA ANALYSIS................................................................. 27
  Overview of the Study.................................................. 27
  Data Analysis Procedure........................................... 28
  Results of the Study.................................................. 29
| Analysis of the Questionnaires          | 29 |
| Analysis of Test Scores                | 33 |
| Analysis of Course Objectives and Test Items | 39 |

**CHAPTER 5**

- **CONCLUSIONS** | 46
- **Summary of the Study** | 46
- **Discussion of Findings and Conclusions** | 47
- **Limitations of the Study** | 51
- **Implications for Further Research** | 51

**REFERENCES** | 53

**APPENDICES** | 55

- **Appendix A:**
  - Test Validity Research Questionnaire | 55
- **Appendix B:**
  - ANOVA and Multiple Comparisons Tables | 58
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 The Correlation Between Midterm Test 1 and 2</td>
<td>34</td>
</tr>
<tr>
<td>2 The Correlation Between Midterm Test 2 and 3</td>
<td>34</td>
</tr>
<tr>
<td>3 The Correlation Between Midterm Test 3 and 4</td>
<td>35</td>
</tr>
<tr>
<td>4 The Correlation Between Midterm Test 4 and 5</td>
<td>35</td>
</tr>
<tr>
<td>5 The Correlation Between Midterm Test 5 and 6</td>
<td>36</td>
</tr>
<tr>
<td>6 The Correlation Between Midterm Test 6 and the End of course Assessment Test</td>
<td>36</td>
</tr>
<tr>
<td>7 Comparison of Course Objectives and Test Items</td>
<td>43</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>DESCRIPTION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DBE English Instructors' Teaching Experience</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>Ranking of Course Components</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Instructors' Perception of the Test Content</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>Instructors' Opinions About Inclusion of Other Language Skills</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>Predictive Validity of Midterm Tests</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>Predictive Validity of the Midterm Tests for Different Faculties</td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>Analysis of Variance for the End of Course Assessment Test</td>
<td>38</td>
</tr>
<tr>
<td>8</td>
<td>Multiple Comparison of Groups for the End of Course Assessment Test</td>
<td>39</td>
</tr>
<tr>
<td>9</td>
<td>Comparison of Course Objectives and Test Items</td>
<td>41</td>
</tr>
</tbody>
</table>
CHAPTER ONE: INTRODUCTION

Introduction to the study

In every educational system, some type of testing is used to evaluate learners' performance or achievement. Tests which are devised to measure learners' performance are defined as instruments of evaluation. These instruments should be a positive part of the teaching-learning process so that future teaching and learning can be more effective (Allan, 1996). Language tests constructed to measure relatively small samples of performance in the case of such a complex thing as language need evaluating since it is difficult to provide a true measure of one particular skill. There are two major concepts that need to be considered in such an evaluation: reliability and validity. Reliability, which is fundamental to the evaluation of any instrument, is the degree to which test scores are free from measurement error and are consistent from one occasion to another (Bachman, 1990; Brown, 1996; Hughes, 1990; Rudner, 1994). Validity is the degree to which a test measures what it is supposed to measure and nothing else (Brown, 1996; Heaton, 1988; Weir, 1990).

The relationship between validity and reliability is complicated. Brown (1996) claims that despite being considered a precondition for validity, the reliability of a test does not necessarily indicate that a test is also valid; a test can give the same result time after time, but not be measuring what it was intended to measure. This leads to the idea that validity is central to test construction. Weir (1990) points out that it is sometimes essential to sacrifice a degree of reliability in order to enhance validity. If validity is lost in an attempt to increase reliability, we end up with a test
which is a reliable measure of something other than what we wish to measure. This is what we want to avoid.

Davies (1990) points out the expansion of the relevance of language testing in two ways: first, there is a growing realization of the need to value validity more than reliability, and second, there is a move to extend the scope of testing to include not only the measurement of learners’ output but also evaluation of courses, materials and projects. Such current views emphasize the importance of determining whether the test is a representative sample of what students have been taught and were expected to learn. In order to improve the validity of language tests, Oller (1983) suggests examining whether the bits and pieces of the language to be tested are really representative of the content of the language taught.

In addition to looking at bits and pieces, it is important to look at test development more broadly. Rudner (1994) suggests that we ask the following questions.

* Did the test development procedure follow a rational approach that ensures appropriate content?

* How similar is this content to the content you are interested in testing?

* What is the overall predictive accuracy of the test? (p.1-3)

Background of the study

In Turkey, at most of the English-medium universities and partly-English-medium universities, there are one-or two-year preparatory programs that teach new entrants English so that they can do their undergraduate studies. Hacettepe University is one of these Turkish universities which has a one-year preparatory program in the Department of Basic English (DBE). The DBE at Hacettepe
University at Beytepe Campus has an academic staff about 100 and an annual intake of about 1000 students. The students of DBE at the Beytepe Campus of Hacettepe University are grouped into three levels: A (intermediate), B (Pre-intermediate), and C (Beginner), and are students from the faculties of Engineering, Education, Economics and Business Administration, Library Science, and Sports Science. Students are placed in A, B, and C level classes according to their grades on the exemption test, which is used as a placement test as well.

In the current situation at DBE, the syllabus is based on the following course books: Front Page 1, 2, 3 (Haines & Carrier) and New First Certificate Masterclass (Hainess & Carrier). The instruction, which focuses on four skills of language (reading, writing, listening and speaking), is based on the stated objectives in these course books.

All students of the DBE are given six midterm tests, an oral assessment test, and the end of course assessment test throughout an academic year. The midterm tests administered every six weeks are constructed according to the objectives of the units covered. These tests focus on measuring students' performance through reading, writing, listening and use of English. Although speaking has a part in the course objectives, it is not measured until the end of the term. The percentages of skills tested and the test formats are as follows:

<table>
<thead>
<tr>
<th>Skill</th>
<th>Percentage</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>30-40%</td>
<td>multiple choice or completion</td>
</tr>
<tr>
<td>Writing</td>
<td>10-15%</td>
<td>composition or letter writing</td>
</tr>
<tr>
<td>Listening</td>
<td>10-15%</td>
<td>multiple choice or completion</td>
</tr>
<tr>
<td>Use of English</td>
<td>30-35%</td>
<td>cloze tests, restatements, completion</td>
</tr>
</tbody>
</table>
The end of course assessment test is administered in June. This test consists of reading, grammar and translation sections. It does not test writing, speaking and listening skills. The end of course assessment test measures students' performance through 100 multiple choice test items. The percentages of the abilities measured are as follows:

- **Reading**: 25-30% multiple choice
- **Grammar**: 55-60% multiple choice
- **Translation**: 10% multiple choice
- **Vocabulary**: 5% multiple choice

The aim of the end of course assessment test is to measure what DBE students accomplish throughout the academic year and that of the midterm tests is to measure what they accomplish every six weeks.

**Statement of the Problem**

Several problems at Hacettepe University, DBE led me to focus on this topic. One is that the effects of excluding writing, listening, and speaking skills from the end of course assessment test are seen in language classrooms at DBE. For example, since these skills are not tested, it may lead some students and teachers to consider them unimportant and therefore, not spend time studying them.

A second concern is the differences between the midterm tests and the end of course assessment test with regard to test format and test content. This led me to investigate the predictive validity of the tests. Another issue is the lack of some skills in the end of course assessment test, which necessitated an investigation of content validity of the end of course assessment test. The fourth concern which had
been observed by me and other teachers as well, is that some faculties are more successful than the others although they have the same instruction.

In addition to these problems, the validity of the tests administered at DBE has never been evaluated, so the current situation at DBE necessitated an investigation of validity of the tests administered.

**Purpose of the study**

The aim of the study is to investigate the degree of validity both in the midterm tests and the end of course assessment test. I will examine the predictive validity of each midterm test on the next one. I will also look at the predictive validity of the last midterm test in order to investigate whether that midterm test is an indicator of the students’ performance in the end of the course assessment test. Since there is a concern that students at some faculties are more successful than others, I will examine the differences among the groups on the end of course assessment test. And finally, I will investigate the content validity of the end of course assessment test to examine whether the test’s content represents the objectives of the course.

**Significance of the Study**

It is hoped that this study will be beneficial for Hacettepe University DBE, and for other universities in Turkey. The language teachers, administrators and students at DBE can benefit from the study in terms of avoiding the negative effects of tests on teaching and learning. It is also hoped that this study will encourage administrators and testers to make some changes on tests, based on the results of the study. Moreover, the administrators and testers may also consider the difference among students from different faculties and may make necessary changes on preparatory course content and on the content of the tests administered at DBE.
Research Questions

This study will address the following research questions:

What is the overall predictive validity of the midterm tests?

To what extent are the scores in the last midterm test related to the scores in the end of course assessment test?

Does the predictive validity of the tests vary across different faculties, and if so how?

In what ways is the end of course assessment test representative of the course content?
CHAPTER TWO: LITERATURE REVIEW

This chapter presents an overview of the literature dealing with second language tests and validity of language tests. First, it provides a basis for the study by discussing definitions and approaches to language testing. Next, it reviews various types of language tests and the purposes of language tests. Thirdly, it takes a close look at the evaluation of language tests. The final section focuses on types of validity.

Definitions and Approaches to Language Testing

Testing is an important part of every teaching and learning experience. Bachman (1991) defines a test as one type of measurement designed to obtain a specific sample of behavior. He distinguishes tests from other types of measurement and claims that language tests can be viewed as the best means of assuring that the sample of language obtained is sufficient for the intended measurement purposes. Genesee and Upshur (1997) define language tests as a description of attributes or qualities of things and individuals by assigning scores to them. 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. Davies (1990) emphasizes the changes in language testing in the 1980s by saying:

Language testing has extended its range of relevance beyond its earlier focus- in two ways: first, by developing measures other than quantitative ones (basically a growing realization of the need to value validity more than reliability) so that qualitative measures of judgement including self-
judgements and control and observation are included in the tester’s reper­
toire; and second, by extending the scope of testing to encompass evalua­
tion, evaluation of courses, materials, projects, using both quantitative and
qualitative measures of plans, processes and input as well as measurement
of learners’ output, the traditional testing approach. (p.74)

Every test has a particular content, usually a reflection of what has been taught or
what is supposed to have been learned. However, evaluation of students’
performance is only one function of a test (Heaton, 1988). Classroom tests also give
teachers a chance to increase learning by making adjustments in their teaching that
enable students to benefit more. In addition, tests enable teachers to evaluate syllabi,
materials, and methods.

In the development of language testing, Heaton (1988) discusses four main
approaches to testing: the essay-translation approach, the structuralist approach, the
integrative approach, and the communicative approach. The first one, the essay-
translation approach which consists of essay writing, translation and grammatical
analysis is also referred to as the “pre-scientific” stage of language testing (Spolsky,
1995, p.5), and as “intuitive and subjective” stage (Madsen, 1983, p.5). Heaton
(1988) claims that no special skill or expertise in testing is required: the subjective
judgement of the teacher is considered to be of paramount importance.

The second stage, the structuralist approach, is also referred to as the “scientific
era” (Madsen, p.6), and the “psychometric-structuralist or modern phase” (Spolsky,
p.4). The structuralist approach emphasizes that language learning is chiefly
concerned with the systematic acquisition of a set of habits. The focus is on the
learner's mastery of the separate elements of the target language. The skills of
listening, speaking, reading and writing are also separated because it is considered essential to test one thing at a time (Heaton 1988, p.17).

The third stage is the integrative approach which involves the testing of language in context and is concerned primarily with meaning and the total communicative effect of discourse. Tests are designed to assess the learner’s ability to use two or more skills simultaneously (Heaton, 1988, p.18).

The fourth stage, the communicative approach, is also referred as the “communicative stage” (Madsen, p.6), and the “psycholinguistic-sociolinguistic or post-modern phase” (Spolsky, p.3). The communicative approach to language testing is also concerned primarily with how language is used, but communicative tests aim to incorporate tasks which approximate as closely as possible those facing the students in real life. Success is judged in terms of the effectiveness of the communication which takes place rather than formal linguistic accuracy (Heaton 1988, p.19). Bachman and Palmer (1996) suggest two principles for communicative testing. First, that there should be a similarity between test performance and language use. And second, that test usefulness, which includes reliability, construct validity, authenticity, interactiveness, impact and practicality, should be considered.

Although the names for each stage or phase are different, a common point shared by Spolsky (1995), Heaton (1988), and Madsen (1983) is that a good test will frequently use features of the communicative approach, the integrative approach and even the structuralist approach, depending on the particular purpose of the test.

In conclusion, the aim of language tests is to enhance teaching and learning. The approaches to language testing which form the framework of language tests have
been expanding with the latest including more of a focus on language use and relationship to "real life."

Types and Purposes of Language Tests

As there are many purposes for which language tests are developed, there are many types of language tests. Davies (1990) claims that "why" is the primary question for language testing studies and he explains the "why" of testing: what we want to do with the information from the tests that is related to the types of tests. Similarly, Brown (1996) states that "the use of testing in language programs is to provide information for making decisions, that is, for evaluation" (p.54).

Bachman (1991) classifies language tests according to five distinguishing features:

1- The purpose
2- The content upon which they are based
3- The frame of reference within which their results are to be interpreted.
4- The way in which they are scored.
5- The specific technique or method they employ (p.70)

He states that this type of classification scheme provides a means for a reasonably complete description of any given language test in order to prevent misunderstanding which could be caused by language tests that refer to only a single feature.

Tests are used to obtain information and the information to be obtained varies from situation to situation, but even so it is possible to categorize tests according to a small number of kinds of information being sought (Hughes, 1990). This categorization will provide useful information both in writing appropriate new tests and deciding whether an existing test is suitable for a particular purpose. Hughes’
categorization includes four types of tests: proficiency, placement, achievement and diagnostic tests. In this section the purpose of each of these tests will be described.

Proficiency Tests

Proficiency tests assess the general knowledge or skills. They are designed to make distinctions between students' performance in order to determine proficiency of students and to place them into the proper level of course. The content of a proficiency test is not related to the content or objectives of a language program (Brown, 1996; Davies, 1996; Henning, 1987; Hughes, 1990).

Placement Tests

Placement tests are used to identify a particular performance level of the student. They aim at grouping students of similar ability levels. Placement tests are specifically related to a program, that is, placement tests must be constructed according to the key features at different levels of teaching in the institution (Brown, 1996; Hughes, 1990).

Achievement Tests

An achievement test measures how much of a language someone has learned in a particular course of study or program of instruction (Brown, 1996; Davies, 1990; Weir, 1993).

There are two types of achievement tests (Hughes, 1990): final achievement tests and progress achievement tests. Final achievement tests are administered at the end of a course of study. The content of these tests must be related to the course objectives with which they are concerned. Progress achievement tests are administered to measure the progress that students are making. These tests must be related to course objectives, as well. Hughes suggests that basing the test content
directly on the objectives of the course has some advantages. First, it obliges course
designers to be explicit about objectives. Second, it makes it possible for
performance on the test to show just how well students have achieved those
objectives. Hughes believes that it is better to base test content on the objectives of
the course rather than on the detailed content of a course since the objectives give
more accurate information about group and individual achievement. According to
him this type of test construction leads to a more beneficial backwash effect on
teaching.

Diagonal Tests

A diagnostic test is designed to provide information about the specific strengths
and weaknesses of students in the specific content domains that are covered in a
language program. The purpose of diagnostic tests is to guide remedial teaching; that
is, diagnostic tests are designed to ascertain what further teaching is necessary

Bachman (1991) claims that any language test has some potential for providing
diagnostic information. He also adds that a placement test can be regarded as a
broad-band diagnostic test in that it distinguishes relatively weak from relatively
strong students. Similarly, Heaton (1988) claims that achievement and proficiency
tests are frequently used for diagnostic purposes. Brown (1995) claims that one well
constructed test designed to reflect the objectives of the course in three equivalent
forms can serve as a diagnostic test at the beginning and middle points in a course
and as an achievement test at the end.
Evaluating Language Tests

Test validity and reliability constitute the two chief criteria for evaluating any test. Bachman and Palmer (1996) state that reliability and validity are critical for tests and are sometimes referred to as essential measurement qualities. This is because these are the qualities that provide the major justification for using test scores as a basis for making inferences or decisions.

Reliability

Reliability refers to the consistency or stability of the test scores obtained (Bachman & Palmer, 1996; Heaton, 1988; Henning, 1987). According to Bachman and Palmer (1996) reliability is an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure.

For Brown (1996), to increase a test's reliability it is better to have “a longer test than a short one, a well designed and carefully written test than a shoddy one, a test with items that have relatively high difference indexes or B-indexes (which indicate the degree to which an item distinguishes between the students who passed the test and those who failed), a test that is clearly related to the objectives of instruction, a test made up of items that assess similar language material than a test that assesses a wide variety of material” (p.222). Bachman (1991) identifies the factors affecting test scores under three headings: test method facets (types of tests and test items), personal attributes (individual characteristics), and random factors (mental alertness, emotional state, unpredictable conditions). Bachman and Palmer suggest that the effects of those potential sources of inconsistency can be minimized through test
design. Hughes (1990) suggests some ways of achieving consistent performances from students. These are as follows:

- Take enough samples of behaviour.
- Do not allow candidates too much freedom.
- Write unambiguous items.
- Provide clear and explicit instructions.
- Ensure that tests are well laid out and perfectly legible.
- Candidates should be familiar with format and testing techniques.
- Provide uniform and non-distracting conditions of administration.
- Use items that permit scoring which is as objective as possible.
- Make comparisons between candidates as direct as possible.
- Provide a detailed scoring key.
- Train scorers.
- Agree acceptable responses and appropriate scores at outset of scoring.
- Identify candidates by number not name.
- Employ multiple, independent scoring. (p.36-42)

In addition to the suggestions above, Carey (1988) adds that one more way to ensure consistency is to determine ways to maintain positive student attitudes toward testing.

Different types of reliability estimates are used to estimate the contributions of different sources of measurement error. The types of reliability as listed by Brown (1996) are: test-retest reliability, equivalent forms reliability and internal consistency reliability. Test-retest reliability is one way of measuring reliability by giving the
students the same test twice to the same group of students. Equivalent forms
reliability measures the reliability of a test by giving parallel tests, that is, two similar
tests with the same type and number of items and the same instructions. Internal
consistency reliability is usually determined through split-half reliability which
randomly assigns test items to two groups and compares the results of the two
groups. Internal consistency is also measured through Kuder-Richardson formulas
which measure reliability in terms of whether all of the items in a test are measuring
the same thing.

Hughes (1990) states that “to be valid a test must provide consistently accurate
measurements. It must therefore be reliable. A reliable test, however may not be
valid at all” (p.42). Brown (1996) points out that test reliability and validity, though
related, are different test characteristics. He also adds that reliability is a
precondition for validity but not sufficient for purposes of judging overall test
quality. Validity must also be carefully examined. According to Hughes (1990)
“There will always be some tension between reliability and validity. The tester has to
balance gains in one against losses in the other”(p.42).

Validity

Validity can be defined as the degree to which a test measures what it is supposed
to measure. According to Alderson, Clapham and Wall (1995) the centrality of the
purpose for which the test is constructed or used cannot be understated. If a test is to
be used for any purpose, the validity of use for that purpose needs to be established
and demonstrated. Carey (1988) suggests that to ensure that tests provide valid
measures of students’ progress, the following questions should be taken into account
during the test design process.
• How well do the behavioral objectives selected for the test represent the
  instructional goal framework?
• How will test results be used?
• Which test item format will best measure achievement of each objective?
• How many test items will be required to measure performance
  adequately?
• When and how will the test be administered? (p.76-77)

Validity can be established in a number of different ways, which leads to
different types of validity and these types are in reality different methods of assessing
validity (Alderson, Clapham & Wall, 1995). Over recent years the increasing interest
in different aspects of validity has led to various names and definitions. Hughes
(1990) states that the aspects of validity are: content validity, criterion-related
validity, construct validity and face validity. Weir (1990) adds one more, washback
validity, and Bachman (1991) discusses the evidential basis for validity including
content relevance and criterion relatedness, which includes concurrent validity and
discusses the aspects of validity under three main titles: content validity, predictive
validity, construct validity and criterion-related validity. Types of validity are
discussed below using Brown’s grouping.

Content Validity. Content validity is the extent to which a sample of skills and
instruction represents a particular domain (Bachman, 1991; Heaton, 1988; Hughes,
1990).
Bachman (1991) 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 (language use tasks) in question.

He puts forward a problem with language tests, which is not having a domain definition that clearly and unambiguously identifies the set of language use tasks from which possible test tasks can be sampled. If this is the case, demonstrating either content relevance or content coverage becomes difficult.

Hughes (1990) points out that the importance of content validity should be taken into consideration for this reason: 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 order to investigate content validity of language tests, the specification of the skills and structures that it is meant to cover should be examined (Heaton, 1988; Hughes, 1990; Weir, 1990). Bachman (1991) claims that the specification of not only the ability domain (reading, writing, vocabulary etc. or micro skills of those domains), but also the test method facets is necessary. Alderson, Clapham and Wall (1995) suggest some ways to investigate content validity:

- Compare test content with specifications/syllabus.
- Questionnaires to, interviews with 'experts' such as teachers, subject specialists, applied linguists.
- Expert judges rate test items and texts according to precise list of criteria. (p.193)

Henning suggests that if two tests are being compared for predictive validity, both tests must be formed for the same purpose. This indicates that predictive validity refers to the extent to which a test can be used to draw inferences regarding achievement.

In order to investigate predictive validity, Alderson, Clapham, and Wall (1995) suggest:

- Correlating students' test scores with their scores on tests taken some time later.
- Correlating students' test scores with success on the final exam.
- Correlating students' test scores with other measures of their ability taken some time later, such as subject teachers' assessments or language teachers' assessments.
- Correlating students' scores with success of later placement (p.194).

Bachman (1991) claims that measures that are valid as predictors of some future performance are not necessarily valid indicators of ability. Heaton (1988) makes a similar claim for the whole concept of empirical validity.

The argument is simply that the established criteria for measuring validity are themselves very suspect: two invalid tests do not make a valid test.

(p.162)

Construct Validity. Heaton (1988) states that if a test is capable of measuring certain specific characteristics in accordance with a theory of language behaviour and
learning, the test has construct validity. The word construct refers to any underlying
ability which is hypothesized in a theory of language ability and theories are put to
test and are confirmed, modified or abandoned (Bachman, 1991; Hughes, 1990).
Henning (1987) claims that the purpose of construct validation is to provide evidence
that underlying theoretical constructs being measured are themselves valid.
Genesee and Upshur (1996) claim that construct validity is probably the most
difficult to understand and the least useful for classroom based evaluation.
Bachman, on the other hand, (1991) claims:

Construct validity is often mistakenly considered to be of importance only
in situations where content relevance cannot be examined because the
domain of abilities is not well specified. With reference to language tests
this misconception has generally been applied to the distinction between
measures of proficiency, which are theory based, and measures of achieve­
ment, which are syllabus based. However, because of the limitations on
content relevance discussed above, even achievement tests must undergo
construct validation if their results to be interpreted as indicators of
ability (p.291).

Construct validity is assessed by:

- Correlating each subtest with other subtests.
- Correlating each subtest with the total test.
- Correlating each subtest with the total test minus self.
- Comparing students’ test scores with students’ biodata and
  psychological characteristics.
-Multitrait-multimethod studies.

- Factor analysis. (Alderson, Clapham & Wall, 1995, p.194)

**Concurrent Validity.** Alderson, Clapham and Wall, (1995) define concurrent validity as follows:

Concurrent validation involves the comparison of the test scores with some other measure for the same candidates taken at roughly the same time as the test. This other measure may be scores from a parallel version of the same test or from some other test; the candidates' self assessments of their language abilities; or ratings of the candidate on relevant dimensions by teachers, subject specialists or other informants (p.177).

Henning (1987) states that concurrent validity is empirical in the sense that data are collected and formulas are applied to generate an actual numerical validity. The coefficient derived represents the strength of relationship with some external criterion measure. The major consideration in collecting evidence of concurrent validity is that of determining the appropriateness of the criterion. It must be ensured that the criterion itself has validity (Bachman, 1991; Henning, 1987).

Although concurrent validity and predictive validity seem similar, predictive validity differs from concurrent validity in that of instead of collecting the external measures at the same time as the administration of the experimental test, the external measures will only be gathered some time after the test has been given.

In this chapter, the literature concerning language testing studies was reviewed. In the first section of this chapter, language testing and approaches to language testing were discussed. In the second section, types and purposes of language tests
were presented. In the third section, evaluation of language tests were examined with regard to reliability and validity. The final section focused on the types of validity.

The next chapter will explain the research design of this study, including information about the subjects involved, materials, and procedures used.
CHAPTER THREE: METHODOLOGY

Introduction

The concern of this study was to examine the degree of validity of tests administered at Hacettepe University, Department of Basic English (DBE).

In this study, the main research questions were: "What is the overall predictive validity of midterm tests? To what extent are the scores in the last mid-term test related to the scores in the end of course assessment test? Does the predictive validity of the midterm tests vary across different faculties? Does the end of course assessment test adequately represent the course content?" To answer the questions, a questionnaire was given to the English Language instructors of DBE at Hacettepe University with respect to content validity of the end of course assessment test. Following that, the objectives of the course with regard to grammar and reading were compared with the test items in the end of course assessment test. Finally, the correlation between test scores was calculated with respect to predictive validity of the mid-term tests. In the following sections, first subjects are introduced, then the materials are explained, followed by procedures and data analysis.

Subjects

Two different groups of subjects were included in this study. These included preparatory class students and English language instructors at DBE at Hacettepe University.

Preparatory Class Students:

For this study C-level preparatory class students were chosen for two reasons: first, they are the largest group in DBE and secondly, the end of course assessment
test is devised according to the C-level students' language level. C-level students enroll in the program as false beginners or zero beginners and at the end of the program they are accepted as upper-intermediate learners.

All C level prep students' midterm tests and the end of course assessment test scores from 1997-1998 were examined. This came to 510 student scores, but due to number of drop outs only 392 scores were used. The scores of all C-level students were involved in this study with the assumption that selected samples might not be the representative of the total population, since in each C-level class there were students from different faculties such as Engineering, Social Sciences and Sports Faculties. In order to answer one of the subquestions of this study all C-level Engineering, Social Sciences and Sports Faculty students’ mid-term tests and the end of course assessment test scores from 1997-1998 were analyzed separately.

English Language Instructors

Twenty-seven C-level English language instructors participated in the study. Their ideas and suggestions were sampled through questionnaires. They had teaching experience of between 5 and 20 years.

Materials

For this research study, questionnaires, the course books, the end of course assessment test, and test scores were used.

Questionnaires

The questionnaire was developed for the English language instructors at DBE at Hacettepe University. The questionnaire was piloted with three English language instructors. The purpose of the questionnaire was to get an idea of the validity of the end of course assessment test. The questionnaire consisted of 11 questions in two
sections. The first section was designed to obtain information on prep school instructors’ teaching experience and prep school course content. In this part the English language instructors were also asked to rank their teaching priorities for language skills on a rank order from 1 to 5. The second part was concerned with the content of the end of course assessment test. In this part the English language instructors were asked to put a cross indicating one of the choices “Yes, No and Somewhat.”

There was also one open-ended item in each section encouraging the subjects to go into more detail or to express different views on the questions asked.

Course Books

The course books, Front Page 1, 2, 3 (Hainess & Carrier) and New First Certificate Masterclass (Hainess & Stewart), were examined to figure out the objectives of the grammar and reading parts of the course. The aim of determining the objectives for these two components of the course was that in the end of course assessment test these two components are the major focus. The objectives of each unit were listed and their frequencies were calculated.

The End of Course Assessment Test

The end of course assessment test’s grammar and reading test items were compared with the objectives of the course with respect to the frequency of the language items studied throughout the academic year.

Test Scores

All C-level students’ six mid-term test scores and the end of course assessment test scores were examined to get an overall idea of the predictive validity of the mid-
term tests. In addition, the predictive validity of the last mid-term test on the end of course assessment was calculated for the students from different faculties.

**Procedures**

After the questionnaires were developed, they were piloted before the actual administration and were assessed for revision of any difficulties in understanding the items of the questionnaire. The necessary changes were made. Before distributing the questionnaires the participants were informed about the purpose of the study. Then the questionnaires were handed out to all C-level English language instructors on January 13 and half of them were collected on the same day. Since some instructors were busy, the remainder was collected in February after the first term break at Hacettepe University. The time duration for filling out each questionnaire was about 15 minutes. There was a response rate of 79%.

After receiving permission from the administrators of DBE, I obtained last year’s test scores and the end of course assessment test from the testing office on January 13. I examined the sheets in order to eliminate students who quit the program or did not take one of the exams. I entered 392 students’ six midterm tests’ scores and the end of course assessment test scores into the computer for a total of 3570 scores.

**Data Analysis**

The questionnaire contained mixed question types. The data obtained from the first section of the questionnaire were the rank order responses which were analyzed in terms of percentage of time spent on skills. Multiple choice questions were analyzed by frequencies and percentages. The answers to the open-ended questions were analyzed by putting them into categories according to recurring themes. Tables and graphs were used to show the results.
The course objectives and the end of course assessment test items were compared in terms of frequencies and percentages of language items studied throughout the academic year. The objectives of the reading skill were identified and counted according to the types of the tasks occurring in course books. The objectives of grammar were identified and counted according to the names of the structures in the table of contents part of the course books.

The students' test scores from the mid-term test and the end of course assessment test were calculated to get their means, standard deviations and correlation coefficient. The results were used to compare the last mid-term test to the end of course assessment test in terms of the predictive validity of the mid-term test on the end of course assessment test. The correlation between tests were computed by means of Pearson Product Moment Correlation Coefficient. The correlation among the six midterm tests were compared as well. The difference among different faculties in the end of course assessment test was computed by means of ANOVA.

The following chapter presents the results of the data analysis and displays all data related to content validity of the end of course assessment test and predictive validity of the mid-term tests administered at Hacettepe University, DBE.
CHAPTER FOUR: DATA ANALYSIS

Overview of the study

This study investigated the predictive validity of the midterm tests and the content validity of the end of course assessment test administered at Hacettepe University, Department of Basic English (DBE).

To collect data, questionnaires (See Appendix A) were given to English instructors at DBE at Hacettepe University. In addition, last year's DBE students’ test scores and the course syllabus were obtained from the test office of DBE. The subjects were C-level (false beginners and zero beginners) DBE students and the English instructors of DBE. Test scores from 510 C-level students were collected but during the data analysis procedure only 392 of the scores were used due to a number of drop outs. There were 34 English instructors of C-level students last year and I asked all of them to participate by filling in a questionnaire. The aim of the questionnaire was to get their ideas about the content of the end of course assessment test.

Students' test scores were used to examine the overall degree of predictive validity of the midterm tests. The degree of predictive validity for students from different faculties was investigated in order to examine the predictive validity of the midterm test for different groups. The difference among different faculties was examined in the end of course assessment test. In addition, the consistency between the course syllabus and the end of course assessment test was examined with regard to content validity of the end of course assessment test.
Data Analysis Procedure

The data were analyzed using the following procedure. First, the results of the questionnaires, which were given to the English instructors were analyzed. The number of the subjects who responded was 27, giving a response rate of 79%. The questionnaire consisted of two sections and 11 items (see Appendix A). The rank order items were analyzed by means of frequencies, and the results were changed into percentages. The mean for each group was calculated first by giving each group a value from 1 to 5 and then the number of respondents in each group was multiplied by the value number. After that, the result was divided by the number of all respondents. Multiple choice items were analyzed by means of frequencies, the results were changed into percentages and finally the results were displayed in tables. The open-ended questionnaire items were analyzed in terms of recurring themes under the headings of reading, writing, listening, speaking. The results were shown in tables with frequencies and percentages.

Next, the predictive validity of the midterm tests was computed by means of Pearson Product Moment Correlation Coefficient to determine the relationship between test scores. Excel was used to compute Pearson Product Moment Correlation Coefficient. The results were shown in scatter-plot graphs. The differences between faculties were computed by means of a One-way ANOVA (See Appendix B) using the SPSS system. In addition, post-hoc comparison was run to determine exactly where the differences lie.

As a last step, the course syllabus and the end of course assessment test were examined in terms of the content validity of the end of course assessment test. First,
the objectives of the grammar and reading parts of the course books were listed and
the frequencies of each language item for the grammar part and subskills of the
reading part were counted. Then, the test items in the end of course assessment test
were examined to determine which structure of grammar and which subskills of
reading they tested. Finally, each test item in the grammar and reading parts was
compared with the objectives of the course with respect to frequencies of the items
studied throughout the academic year.

Results of the Study

Analysis of the questionnaires

The questionnaires were given to 34 C-level English instructors. The number of
the subjects who responded was 27.

The first question gave information about teaching experience of English
Instructors at DBE, which is displayed in Table 1.

Table 1

<table>
<thead>
<tr>
<th>DBE English Instructors’ Teaching Experience (N=27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Teaching Experience</td>
</tr>
<tr>
<td>1-5 6-10 11-15 16-20 Above 20</td>
</tr>
<tr>
<td>Overall Teaching Experience</td>
</tr>
<tr>
<td>Experience</td>
</tr>
<tr>
<td>4 8 9 5 1</td>
</tr>
<tr>
<td>Teaching Experience at DBE at HU</td>
</tr>
<tr>
<td>12 5 4 5 1</td>
</tr>
</tbody>
</table>

Note. HU = Hacettepe University; DBE = Department of Basic English
It is clear from the table that the largest percentage of the teachers in this study have five or fewer years of experience. There is only one person whose teaching experience is over 20 years.

Question 3 asked the English instructors to rank order the components of the course in order of time spent on them. Table 2 presents the DBE C-level English instructors' ranking of the components of the course.

Table 2

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Reading</th>
<th>Speaking</th>
<th>Listening</th>
<th>Writing</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
<td>23 (92%)</td>
</tr>
<tr>
<td>2</td>
<td>17 (68%)</td>
<td>2 (8%)</td>
<td>4 (16%)</td>
<td>-</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>3</td>
<td>6 (24%)</td>
<td>3 (12%)</td>
<td>12 (48%)</td>
<td>4 (16%)</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2 (8%)</td>
<td>10 (40%)</td>
<td>7 (28%)</td>
<td>6 (24%)</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>10 (40%)</td>
<td>1 (4%)</td>
<td>14 (56%)</td>
<td>-</td>
</tr>
<tr>
<td>Mean</td>
<td>2.4</td>
<td>4.12</td>
<td>3.12</td>
<td>4.28</td>
<td>1.08</td>
</tr>
</tbody>
</table>

Note 1. 1= the most time spent, 5= the least time spent

Note 2. See page 28 for explanation of calculating weighted mean.

When the weighted means of the answers are compared, it is observed that the respondents ranked the course components in the following order: grammar (M=1.08), reading (M=2.4), listening (M=3.12), speaking (M=4.12), writing (M=4.28), in decreasing order of time spent. Of the 27 respondents, 2 were not included in the rank order analysis since they gave the same order to more than one item. Of the 25 respondents, 23 (92%) ranked grammar as the component on which the most time spent.
Question 4 was designed to get the ideas of the DBE English instructors about the changes they want in the course syllabus. It was an open-ended question and the responses for the item were analyzed by means of recurring themes, such as writing, speaking, video lessons and skills lessons. Of the 27 respondents, 18 (66.6%) answered the open-ended question. Of the 18 respondents, 9 English instructors (50%) wanted writing and speaking skills to be taught separately, 3 (16.6%) wanted skills-based lessons, 2 (11.11%) wanted video and language laboratory lessons, 2 (11.11%) wanted reading skills to be taught separately and 2 (11.11%) did not want any change in the course syllabus. The results indicate that most English instructors want the language skills to be taught separately, though this response only represents 47.05% of all English instructors.

Questions 5-7 were designed to get teachers’ ideas about whether the content of the end of course assessment test represents the content of the course books. Question 5 was about the grammar part of the test, 6 was about the vocabulary part and 7 was about the reading part of the test. There were three choices to the questionnaire items and the responses were analyzed by means of frequencies and the results were changed into percentages. The results are reported in Table 3.

Table 3

<table>
<thead>
<tr>
<th>Instructors’ Perception of the Test Content (N=27)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>5-Grammar</td>
</tr>
<tr>
<td>6-Vocabulary</td>
</tr>
<tr>
<td>7-Reading</td>
</tr>
</tbody>
</table>
Table 3 indicates that most instructors (over 60%) agree that grammar and vocabulary parts of the course content are well-represented on the end of course assessment test. With regard to reading part of the end of course assessment test, while 11 instructors (40%) agree that it is represented, 8 think that it is not represented and 8 think that it is somewhat represented.

Questions 8-10 were designed to get the English instructors' opinions about the language skills which are not included in the end of course assessment test. The results are displayed in Table 4.

Table 4

<table>
<thead>
<tr>
<th>Skills</th>
<th>Yes</th>
<th>No</th>
<th>Undecided</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-Writing</td>
<td>18 (66.6%)</td>
<td>7 (25.92%)</td>
<td>2 (7.40%)</td>
</tr>
<tr>
<td>9-Listening</td>
<td>21 (77.77%)</td>
<td>4 (14.81%)</td>
<td>2 (7.40%)</td>
</tr>
<tr>
<td>10-Speaking</td>
<td>20 (74.07%)</td>
<td>5 (18.51%)</td>
<td>2 (7.40%)</td>
</tr>
</tbody>
</table>

In questions 8-10, the instructors were asked whether they think there should be writing, listening and speaking sections on the end of course assessment test. Most English instructors agree that writing, listening and speaking skills should be included in the end of course assessment test.

Question 11 asked the instructors whether there are any subskills which should be on the end of course assessment test. Of the 27 respondents only 4 instructors (14.81%) answered this question. Their answers were all different. One instructor (25%) responded no. One instructor (25%) thinks the final test should have different skills
tested in separate hours, e.g. structure, reading, writing and listening. One instructor (25%) thinks that it is unfair to expect more than we teach. One instructor (25%) thinks that this cannot be achieved with the present syllabus. Since 23 instructors did not answer this question, we might assume that most do not think that more subskills should be tested.

Analysis of Test Scores

In order to analyze the relationship between tests, Pearson Product Moment Correlation Coefficient (PPMC) was used. First, the correlation between tests were determined in terms of overall predictive validity of midterm tests. The results are shown in Table 5.

Table 5

Predictive Validity of Midterm Tests (N=392)

<table>
<thead>
<tr>
<th>Midterm Tests</th>
<th>r</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>.91*</td>
<td>390</td>
</tr>
<tr>
<td>2-3</td>
<td>.87*</td>
<td>390</td>
</tr>
<tr>
<td>3-4</td>
<td>.81*</td>
<td>390</td>
</tr>
<tr>
<td>4-5</td>
<td>.80*</td>
<td>390</td>
</tr>
<tr>
<td>5-6</td>
<td>.71*</td>
<td>390</td>
</tr>
<tr>
<td>6-ECAT</td>
<td>.72*</td>
<td>390</td>
</tr>
</tbody>
</table>

* \( p \leq .01 

Note 1. 1-2 indicates the predictive validity of test 1 on test 2.
Note 2. ECAT is End of Course Assessment Test.
The correlation between tests is highly significant although the value of $r$ decreases towards the end. To see the direction of the correlation, each student's mark is placed on the scatterplot diagrams for each midterm test.

Figure 1. The Correlation Between Midterm Test 1 and 2.

Figure 2. The Correlation Between Midterm Test 2 and 3.
Figure 3. The Correlation Between Midterm Test 3 and 4.

Figure 4. The Correlation Between Midterm Test 4 and 5.
The Correlation Between Midterm 5 and 6

Figure 5. The Correlation Between Midterm Test 5 and 6.

The Correlation Between Midterm 6 and The End of Course Assessment Test

Figure 6. The Correlation Between Midterm Test 6 and the End of Course Assessment test.
The scatterplot diagrams indicate that there are several outliers who took zero from the tests. They are possibly Sports Science Faculty students since the predictive validity of tests is very low for this faculty. The reasons for the low correlation for this faculty will be discussed in Chapter 5. However, those outliers do not affect the results of the study.

The results of the scatterplot diagrams indicate that there is a positive correlation between tests. Although the line does not touch all the points it does reflect the general direction of the relationship.

The correlation between tests for different faculties was computed again by means of Pearson Product Moment Correlation Coefficient. The results are displayed in Table 6.

Table 6
Predictive Validity of the Midterm Tests for Different Faculties

<table>
<thead>
<tr>
<th>Faculty</th>
<th>MidT.1</th>
<th>MidT.2</th>
<th>MidT.3</th>
<th>MidT.4</th>
<th>MidT.5</th>
<th>MidT.6</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>.87*</td>
<td>.82*</td>
<td>.78*</td>
<td>.75*</td>
<td>.66*</td>
<td>.62*</td>
<td>153</td>
</tr>
<tr>
<td>Engineering</td>
<td>.90*</td>
<td>.91*</td>
<td>.86*</td>
<td>.87*</td>
<td>.83*</td>
<td>.88*</td>
<td>152</td>
</tr>
<tr>
<td>Library Sc.</td>
<td>.90**</td>
<td>.89**</td>
<td>.68**</td>
<td>.72**</td>
<td>.88**</td>
<td>.89**</td>
<td>34</td>
</tr>
<tr>
<td>Economics and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business Adm.</td>
<td>.94**</td>
<td>.92**</td>
<td>.87**</td>
<td>.66**</td>
<td>.36***</td>
<td>.50****</td>
<td>40</td>
</tr>
<tr>
<td>Sports Sc.</td>
<td>.87**</td>
<td>.23</td>
<td>.19</td>
<td>.19</td>
<td>.81**</td>
<td>-.35</td>
<td>13</td>
</tr>
</tbody>
</table>

*p=<.0005, **p=<.001, ***p=<.05, ****p=<.01.

In this table, the level of significance varies due to the number of test scores in each group. The predictive validity of tests for different faculties is significant except for Sports Science Faculty. The predictive validity of the tests for Sports Science Faculty does not indicate consistency. It is surprising to note that while the fifth midterm test
for Sports Science Faculty yielded a value of .81, the correlation between the sixth
midterm test and the end of course assessment test is -.35. This may be because the fifth
midterm test was appropriate for their level of knowledge and the sixth one was difficult
for them. Another explanation may be that Sports Faculty students join tournaments so
they miss some of the instruction and sometimes they miss the exams.

Table 6 indicated the correlation between midterm 6 and the end of course
assessment test. Since the results were so varied I decided to analyze them in more
details. To examine the differences between groups and within groups on the end of
course assessment test a One-way ANOVA was used. The results of the ANOVA are
displayed in Table 7.

Table 7

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>9130.864</td>
<td>4</td>
<td>2282.716</td>
<td>11.638</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>75910.256</td>
<td>387</td>
<td>196.151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>85041.120</td>
<td>391</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An F ratio greater than 1 is necessary to show any differences at all among groups.
The F ratio in Table 7 indicates that the differences between groups are highly
significant at .000 level. In addition, the mean square between groups (2282.716) is
greater than the mean square within groups (196.151), which indicates that there is some
difference between the five groups. The results indicate that the performance of
different faculties on the end of course assessment test differs across the five groups. To
investigate exactly where the differences lie, a post hoc comparison was computed for
each group. A post hoc comparison is used to compare each group with every other group. The results of the post hoc comparison are displayed in Table 8.

Table 8

**Multiple Comparison of Groups for the End of Course Assessment Test**

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>-2.10</td>
<td>10.50*</td>
<td>-9.19*</td>
<td>10.47</td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>12.60*</td>
<td>-7.10*</td>
<td>12.57*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library Sc.</td>
<td>-19.70*</td>
<td></td>
<td>-19.70*</td>
<td>-3.17</td>
<td></td>
</tr>
<tr>
<td>Economics and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business Adm.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.67*</td>
</tr>
</tbody>
</table>

**Note 1.** Group 5 is Sports Science Faculty.

*p=< .05

Table 8 indicates the mean differences between faculties. The results indicate that the ones which are not significantly different are:

a. Education Faculty and Engineering Faculty,

b. Education Faculty and Sports Science Faculty,

c. Library Science Faculty and Sports Science Faculty.

The highest mean difference is between Economics and Business Administration Faculty and Library Science, which is to be expected because the former has the highest mean and the latter has the lowest.

**Analysis of Course Objectives and Test Items**

In the course books, Front Page 1, 2, 3, and New First Certificate Masterclass, there are 54 reading passages followed by tasks to teach and learn the components of reading
skills. In the end of course assessment test there are two reading passages followed by 17 multiple choice test items. There are also nine items to test reading comprehension. Those 9 items are the ones which are not studied throughout the academic year since those types of reading comprehension items are not in the course books. However, those types of items are practiced towards the end of the second term to make students familiar with the form. Therefore, I did not include those items while analyzing the test items in the reading part of the end of course assessment test since they are not included in the components of the reading in the course books.

The frequencies of objectives of the reading skill taught throughout the year and the number of test items of the reading part of the end of course assessment test are shown in Table 9.
Table 9

Comparison of Course Objectives and Test Items

<table>
<thead>
<tr>
<th>Components of Reading Skills</th>
<th>Frequencies in Course book</th>
<th>Number of Items on Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple choice Comprehension Questions</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>23</td>
<td>-</td>
</tr>
<tr>
<td>Skimming</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Scanning</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td>Inference</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Wh- Comprehension Questions</td>
<td>34</td>
<td>-</td>
</tr>
<tr>
<td>Reference</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Fill in the gaps</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>Finding the Meaning of a Word in Texts</td>
<td>36</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 9 indicates that the most frequent component of reading instruction, which is finding the meaning of a word in texts, is tested with two questions while the second and third most frequent components, Wh- comprehension questions and multiple matching, are not tested at all. The components, reference and inference, are tested nearly the same number of times as their frequencies in the course books although their frequencies are the lowest ones overall.

The grammar components are taken from the syllabus-like book prepared by one of the testers. First, the components were listed and then their frequencies were counted and their percentages were calculated. Second, the items testing grammar in the end of
course assessment test were analyzed to determine what components they are testing. The test items were multiple choice questions and there were 43 items used to test grammar. The first part of grammar test consists of 30 test items. Students are required to fill in the blanks with the correct answer by choosing among four options. Of the 30 test items, two were excluded since they were testing vocabulary. The other four parts were written to test structure and meaning; for example, one part was a dialogue completion, one part was sentence completion, one part was finding the sentence whose meaning is almost the same with the given one.

The frequencies of the grammar components and the number of items testing these components are shown in Figure 7. The components are listed in the form used in the course book.
<table>
<thead>
<tr>
<th>Grammar Components</th>
<th>Frequency in the Course Books</th>
<th>Frequency in the Test</th>
<th>Grammar Components</th>
<th>Frequency in the Course Books</th>
<th>Frequency in the Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Present Tense</td>
<td>15</td>
<td>27.77</td>
<td>Past Participle Clauses</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Present Continuous Tense</td>
<td>9</td>
<td>16.66</td>
<td>Pronouns</td>
<td>6</td>
<td>11.11</td>
</tr>
<tr>
<td>Simple Past Tense</td>
<td>8</td>
<td>14.81</td>
<td>Articles</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Past Continuous Tense</td>
<td>3</td>
<td>5.55</td>
<td>Acverbs</td>
<td>4</td>
<td>7.40</td>
</tr>
<tr>
<td>Present Perfect Tense</td>
<td>3</td>
<td>5.55</td>
<td>W.sh Clauses</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>Past Perfect Tense</td>
<td>2</td>
<td>3.70</td>
<td>Relative Clauses</td>
<td>3</td>
<td>5.55</td>
</tr>
<tr>
<td>Simple Future Tense</td>
<td>3</td>
<td>5.55</td>
<td>Comparative/Superlative Adj.</td>
<td>8</td>
<td>14.81</td>
</tr>
<tr>
<td>Future Perfect Tense</td>
<td>2</td>
<td>3.70</td>
<td>Prepositions</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Future Continuous Tense</td>
<td>2</td>
<td>3.70</td>
<td>Gerunds and Infinitives</td>
<td>8</td>
<td>14.81</td>
</tr>
<tr>
<td>Present Perfect Continuous Tense</td>
<td>2</td>
<td>3.70</td>
<td>Countable/Uncountable Nouns</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>Going to</td>
<td>4</td>
<td>7.40</td>
<td>Sentence Connectors</td>
<td>10</td>
<td>18.51</td>
</tr>
<tr>
<td>Be able to</td>
<td>3</td>
<td>5.55</td>
<td>Question Tags</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Would Like</td>
<td>3</td>
<td>5.55</td>
<td>Reported Speech</td>
<td>6</td>
<td>11.11</td>
</tr>
<tr>
<td>Can’t</td>
<td>6</td>
<td>11.11</td>
<td>Causative</td>
<td>5</td>
<td>9.25</td>
</tr>
<tr>
<td>Can</td>
<td>4</td>
<td>7.40</td>
<td>Quantifiers</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>Have to</td>
<td>5</td>
<td>9.25</td>
<td>Do and Make</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Could</td>
<td>3</td>
<td>5.55</td>
<td>So and Such</td>
<td>3</td>
<td>5.55</td>
</tr>
<tr>
<td>Shall</td>
<td>2</td>
<td>3.70</td>
<td>So/Neither</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Must</td>
<td>6</td>
<td>11.11</td>
<td>Frequency Adverbs</td>
<td>3</td>
<td>5.55</td>
</tr>
<tr>
<td>Mustn’t</td>
<td>3</td>
<td>5.55</td>
<td>Passive</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>Need</td>
<td>3</td>
<td>5.55</td>
<td>If Clauses</td>
<td>5</td>
<td>9.25</td>
</tr>
<tr>
<td>Should</td>
<td>3</td>
<td>5.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ought to</td>
<td>3</td>
<td>5.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had Better</td>
<td>1</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had to</td>
<td>1</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would Rather</td>
<td>1</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>2</td>
<td>3.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Might</td>
<td>2</td>
<td>3.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used to</td>
<td>3</td>
<td>5.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would</td>
<td>1</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Be Used to</td>
<td>1</td>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7. Comparison of Course Objectives and Test Items
We know that some structures are recycled in the course books, that’s why their frequencies are more than the others. However, in each recycling unit something new is added to the known structures, for example, Simple Present Tense is first given with the subject pronouns “I, we, you, and they,” then in the recycling unit “he and she” are added, in the next unit, frequency adverbs are introduced, then Simple Present Tense for future meaning is given. That is, they are not simply recycling units repeating the previous one and in the table of contents part of the course books the name of the structure is Simple Present Tense. Even they are recycling units we do not skip them, we do not teach the same structure again but we spend time to practice the structure recycled. At DBE, priority is given to the most frequent tasks and structures while devising tests (from an informal interview with the testers) so the most frequent items should be equally represented in the test. When the number of test items is compared with the frequencies of the objectives, it is observed that the test items are not prepared according to the frequencies of objectives. For example, a component of grammar, Present Perfect Tense, whose frequency is three is tested by four questions while Simple Present Tense whose frequency is 15 is tested by only one question. Some components’ test item numbers are nearly the same as their frequency numbers, for example, conditionals and relative clauses. Some components are not tested at all although their frequencies are higher than the tested ones, for example, causatives, some tenses and some modals. Of the 54 components, only 23 (42.59%) were tested. This indicates that the content validity of the end of course assessment test is lacking. To remedy this, there should be a better representation of course book items in the test. Instead of
testing one component more than once, more components could have been tested in that
number of test items.

Throughout 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 further research and pedagogical implications.
CHAPTER FIVE: CONCLUSIONS

Summary of the Study

The aim of this research study was to investigate the validity of tests administered at Hacettepe University, Department of Basic English (DBE). The predictive validity of the midterm tests was examined to determine whether each midterm test indicated students' performance on the succeeding midterm and on the end of course assessment test as well. The content validity of the end of course assessment test was examined in terms of consistency between the test content and the course objectives. The end of course assessment test scores were analyzed with respect to differences among different faculties.

In this study, the subjects were 510 C-level (Beginner) students and 34 English instructors who taught C-level students during the year 1997-1998. All C-level students were involved in this study with the assumption that selected samples might not represent the whole population since there were students from five different faculties in each C-level class.

For this study data were collected through questionnaires and students' five midterm tests scores and the end of course assessment test scores. To analyze data, Pearson Product Moment Correlation Coefficient was used to determine the correlation between tests with respect to predictive validity. The results were displayed in tables and in diagrams. To examine the differences between groups a One-way ANOVA was computed. The questionnaires were analyzed by means of frequencies and percentages and the results of the questionnaires were used to
determine the content validity of the end of course assessment test. The results of the questionnaires were displayed in tables.

Discussion of Findings and Conclusions

This section discusses the findings of the study and draws conclusions about the research questions outlined in Chapters 1-4. Each sub-section relates to one of the research questions. Where relevant, references to other reported research in the literature are presented.

The first research question was: What is the overall predictive validity of each midterm test conducted at Hacettepe University, Department of Basic English? The results of the Pearson Product Moment Correlation Coefficient indicated that there was a significant correlation between midterm tests at $p<.01$ significance level. In addition, the relationship between tests was positive. Coefficients either positive or negative up to about +.40 or down to about -.40 indicate fairly weak relationships. Relatively strong correlations are those that range from +.80 to 1.0 or -.80 to -1.0 (Brown 1996). For the predictive validity of the first five midterm tests, the last one was below +.80, the others were above +.80. The reason that predictive validity of midterm 5 on midterm 6 which is lower than .80 may be that midterm 6 was written in a form parallel to the end of course assessment test and was unlike the first five. The components of the fifth midterm test and the sixth midterm test are different. That is, the two tests do not measure the same components. For example, the sixth midterm test does not test writing skill and all test items are in multiple choice format while in the fifth midterm students are required to fill in the blanks or complete sentences with their own words without having options. Possibly due to the differences between the forms and content of the tests, the correlation was only
moderately strong (r = .71). Although the correlation is lower than +.80, it is still statistically significant at p<.01 according to the Pearson Product Moment Correlation Coefficient. Excluding writing skills or using multiple choice test items might have caused the lower correlation that occurred between the fifth and the sixth midterm tests as compared to the high correlation between the first four midterm tests ranging from (+.91 to +.80).

The second research question was: To what extent are the scores in the last midterm test related to the scores in the end of course assessment test? The correlation between the last midterm test and the end of course assessment test was r = .72, which indicates a moderately strong correlation and a statistically significant result. To examine the reasons for this correlation, first, I compared the forms of the tests. The last midterm test and the end of course assessment test were shown to be similar to each other. The format and the types of some test items in midterm 6 are exactly the same in the end of course assessment test (See Appendix C). The same type of test components, such as putting the sentences in order, or finding the irrelevant sentence, are tested in multiple choice items. Another reason that the correlation was expected to be high was the time spent on course components. According to the questionnaire results (See Table 2), teachers spent most classroom time teaching grammar and reading skills. The end of course assessment test is mainly devised on these components; for example, grammar is tested in 43 questions and reading is tested in 26 questions. Since these are the main components of the end of course assessment test; it is not surprising that the relationship is significant.
The third research question was: Does the predictive validity vary among different faculties?

The results of the Pearson Product Moment Correlation Coefficient indicated that the correlation between tests varies among different faculties (see Table 6). It was seen that when the predictive validity of six midterm tests was computed separately for each faculty, the results were striking. There was a strong, positive correlation between tests for Education, Engineering, Library Science and Economics and Business Administration Faculties. The correlation between tests for Sports Science Faculty was low. Moreover, for the Sports Science Faculty, the correlation between the last midterm and the end of course assessment test was negative. Brown (1996) states that differential group studies show that the test differentiates between groups: some groups have acquired the knowledge and some have not. To investigate the differences between groups a One-way ANOVA and post hoc comparison were computed for the end of course assessment test (see Tables 7 and 8). The rationale behind designing a One-way ANOVA is to determine whether the performance on the end of course assessment test differs across the five groups. The results were significant for most groups indicating that they did not perform in the same way on the test. This could be related to the fact that students’ grades for initial entrance to different faculties differ, for example, Engineering and Economics and Business Administration faculties are the ones which require the highest grades of all while Sports Science Faculty accepts students with the lowest grades. The correlation between the tests varies only for Sports Science Faculty. It might be caused by those students’ missing some tests and instruction because of their tournaments. Those students take make-up exams which might include some parts of missing instruction
or might be more difficult than the actual test. It leads to a possible reevaluation of the relationship between the Sports Science Faculty and the rest of the university since the correlation between tests was significant for the others. A further explanation into how or why the groups differ, however, is beyond the scope of this study.

The fourth research question was: In what ways is the end of course assessment test representative of the course content? According to the results of the questionnaires responded to by C-level teachers, the end of course assessment test’s grammar, reading and vocabulary parts represent the course content and most teachers think that writing, listening and speaking skills should be included in the end of course assessment test. The results of the analysis of course objectives and the test items of the end of course assessment test, however, indicate that the test items were not chosen according to the frequencies of course objectives (see Figure 7). The least frequent course objectives were tested more than the most frequent ones. For example, when the components of grammar are compared, the frequency of Present Perfect Tense is three and the number of items testing this component is four. Similarly the frequency of Relative Clause is three and the number of items testing it is three whereas the most frequent components, such as Present Continuous Tense or the modal Must, are not tested at all. The same situation is observed for the reading section as well. For example, the most frequent component of the reading skill which is Wh- Comprehension Question is not tested at all while the least frequent one, Reference, is tested in three questions. The lack of test objectives might have caused such a situation because the number of course objectives is high in terms of grammar and it is impossible to test all the bits and
pieces of grammar. Therefore, the testers might have chosen the main structures to test while ignoring the others. For these reasons, I can say that the end of course assessment test is a limited representative of the course content when the proportions of language items in the course books are compared with the proportions of test items in the end of course assessment test. For example, while the percentage of Tenses in the course books is 83.29%, which is the highest number of all components, their percentage in the test is 19.48%.

Limitations of the Study

The first limitation of the study was the number of faculties involved. The predictive validity of the tests was investigated for the five faculties in Beytepe Campus but there are also other faculties of Hacettepe University in the city centre. So the results of the study cannot be generalized for all C-level faculties. The second limitation was the lack of test objectives and specified types and proportions of test items which were necessary to examine the content validity of the end of course assessment test. The content validity of the test was assessed with only the results from questionnaires and analysis of course objectives and the end of course assessment test’s test items with regard to the frequencies of course objectives. These are not enough to evaluate the content validity of the test as a whole since a detailed item analysis is necessary for such a study. Because of time constraints, this study could not be done.

Implications for Further Research

This study might be helpful for teachers who are interested in language testing since it investigates the validity of tests. Moreover, this study can be a model for other validation studies. The teachers, testers and administrators at DBE at
Hacettepe University should benefit from the analysis of the midterm tests and the end of course assessment test with respect to teaching, learning and testing. It is hoped that through this study, testers and administrators will become more aware of the need for test objectives before devising tests to ensure content validity. It is also hoped that the testers at DBE, Hacettepe University will revise the end of course assessment test to include a better representation of skills studied throughout the academic year.

This research study focused on assessing the validity of tests administered at Hacettepe University, DBE. Further research can be done to investigate the reasons for differences between faculties in connection with tests and the course content as well. Another research study might be done to prepare and conduct a model end of course assessment test including writing, speaking and listening skills.
REFERENCES


APPENDIX A

TEST VALIDITY RESEARCH QUESTIONNAIRE

Dear Colleagues,

I am an MA TEFL graduate student at Bilkent University. I am doing a research on the validity of the midterm tests and the end of course assessment test administrated at Hacettepe University. 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 will be grateful if you would take a few moments to complete the questions.

Thank you,

Hilal Ösken
SECTION 1: GENERAL INFORMATION ON PREPARATORY SCHOOL INSTRUCTORS AND PREPARATORY SCHOOL COURSE CONTENT

Please answer all the questions.

Put a cross (X) in the appropriate box or write in the space provided.

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

2. How long have you been teaching at Hacettepe University?
   ( ) 1-5 years   ( ) 11-15 years
   ( ) 6-10 years  ( ) 16-20 years

3. Please number the following course components from 1-5 according to the amount of time you spend on them in one academic year.

   (1=the most time)

   ..............Reading
   ..............Writing
   ..............Listening
   ..............Speaking
   ..............Grammar

4. Are there any changes you think should be made in Prep school course content?
SECTION 2: QUESTIONNAIRE ON THE END OF COURSE ASSESSMENT
TEST FOR PREPARATORY SCHOOL INSTRUCTORS

5. Does the grammar part of the end of course assessment test represent the grammar parts of the course books you use in your English lessons?

( ) Yes, usually   ( ) No, rarely   ( ) Somewhat

6. Does the vocabulary part of the end of course assessment test represent the vocabulary parts of the course books you use in your English lessons?

( ) Yes, usually   ( ) No, rarely   ( ) Somewhat

7. Does the reading part of the end of course assessment test represent the reading parts of the course books you use in your English lessons?

( ) Yes, usually   ( ) No, rarely   ( ) Somewhat

8. Do you think there should be a writing part of the end of course assessment test?

( ) Yes   ( ) No   ( ) Undecided

9. Do you think there should be a listening part of the end of course assessment test?

( ) Yes   ( ) No   ( ) Undecided

10. Do you think there should be a speaking part of the end of course assessment test?

( ) Yes   ( ) No   ( ) Undecided

11. Are there any subskills you think should be on the end of course assessment test?

If so what?
## APPENDIX B

ANOVA and Multiple Comparisons Tables

### ANOVA

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOTLAR Between Groups</td>
<td>9130,864</td>
<td>4</td>
<td>2282,716</td>
<td>11,638</td>
<td>.000</td>
</tr>
<tr>
<td>NOTLAR Within Groups</td>
<td>75910,256</td>
<td>387</td>
<td>196,151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>85041,120</td>
<td>391</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Multiple Comparisons

Dependent Variable: NOTLAR

Tukey HSD

<table>
<thead>
<tr>
<th>(I) GRUP</th>
<th>(J) GRUP</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-2.10</td>
<td>1.604</td>
<td>.686</td>
<td>-6.47</td>
<td>-2.28</td>
<td>2.28</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10.50*</td>
<td>2.655</td>
<td>.001</td>
<td>3.26</td>
<td>17.75</td>
<td>5.36</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>-9.19*</td>
<td>2.487</td>
<td>.002</td>
<td>-15.98</td>
<td>-2.41</td>
<td>21.11</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>10.47</td>
<td>4.046</td>
<td>.073</td>
<td>-.57</td>
<td>21.51</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2.10</td>
<td>1.604</td>
<td>.686</td>
<td>-2.28</td>
<td>6.47</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>12.60*</td>
<td>2.657</td>
<td>.000</td>
<td>5.35</td>
<td>19.85</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>-7.10*</td>
<td>2.469</td>
<td>.035</td>
<td>-13.89</td>
<td>-31.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>12.57*</td>
<td>4.047</td>
<td>.016</td>
<td>1.53</td>
<td>23.61</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>-10.50*</td>
<td>2.655</td>
<td>.001</td>
<td>-17.75</td>
<td>-3.26</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-12.60*</td>
<td>2.657</td>
<td>.000</td>
<td>-19.85</td>
<td>-5.35</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>-19.70*</td>
<td>3.267</td>
<td>.000</td>
<td>-28.61</td>
<td>-10.79</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>-3.17E-02</td>
<td>4.567</td>
<td>1.000</td>
<td>-12.49</td>
<td>12.43</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>9.19*</td>
<td>2.487</td>
<td>.002</td>
<td>2.41</td>
<td>15.98</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>7.10*</td>
<td>2.489</td>
<td>.035</td>
<td>.31</td>
<td>13.89</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>19.70*</td>
<td>3.267</td>
<td>.000</td>
<td>10.79</td>
<td>28.61</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>19.67*</td>
<td>4.471</td>
<td>.000</td>
<td>7.47</td>
<td>31.86</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>-10.47</td>
<td>4.046</td>
<td>.073</td>
<td>-21.51</td>
<td>-5.7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-12.57*</td>
<td>4.047</td>
<td>.016</td>
<td>-23.61</td>
<td>-1.53</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3.17E-02</td>
<td>4.567</td>
<td>1.000</td>
<td>-12.43</td>
<td>12.49</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>-19.67*</td>
<td>4.471</td>
<td>.000</td>
<td>-31.86</td>
<td>-7.47</td>
<td></td>
</tr>
</tbody>
</table>

* The mean difference is significant at the .05 level.