To the memory of my beloved, self- sacrificing father,

Erol Orhan Küçük

THE RELATIONSHIP AMONG FACE VALIDITY, RELIABILITY AND PREDICTIVE VALIDITY OF UNIVERSITY EFL PREPARATORY SCHOOL ACHIEVEMENT TESTS

The Graduate School of Education
of
Bilkent University

by

FUNDA KÜÇÜK

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

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

June 2007

BİLKENT UNIVERSITY

GRADUATE SCHOOL OF EDUCATION

MA THESIS EXAMINATION RESULT FORM

June 15, 2007

The examining committee appointed by the Graduate School of Education

for the thesis examination of the MA TEFL student

Funda Küçük

has read the thesis of the student.

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

Thesis title	: The Relationship among Face Validity, Reliability and Predictive Validity of University EFL Preparatory School Achievement Tests
Thesis Advisor	: Visiting Asst. Prof. Dr. JoDee Walters Bilkent University, MA TEFL Program
Committee Members	: Asst. Prof. Dr. Julie Mathews-Aydınlı Bilkent University, MA TEFL Program
	Asst. Prof. Dr. Arda Arıkan Hacettepe University, Faculty of Education Department of Foreign Languages Teaching; Division of English Language Teaching

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

_____

(Visiting Asst. Prof. Dr. JoDee Walters)
Supervisor

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

_____

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

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

_____

(Asst. Prof. Dr. Arda Arıkan)
Examining Committee Member

Approval of the Graduate School of Education

_____

(Visiting Prof. Dr. Margaret Sands)
Director

ABSTRACT

THE RELATIONSHIP AMONG FACE VALIDITY, RELIABILITY AND PREDICTIVE VALIDITY OF UNIVERSITY EFL PREPARATORY SCHOOL ACHIEVEMENT TESTS

Funda Küçük

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

Supervisor: Asst. Prof. Dr. JoDee Walters

Co-Supervisor: Asst. Prof. Dr. Julie Mathews-Aydınlı

June 2007

This study examined the relationship between face validity and relatively more objective measures of tests, such as reliability and predictive validity. The study also examined the face validity, reliability and predictive validity of the achievement tests administered at Zonguldak Karaelmas University Preparatory School.

The instruments employed in this study were two questionnaires and C-(beginner) level students' test scores. First, instructors and students were given questionnaires to define the degree of face validity and reliability of the achievement tests. Second, the correlations between students' first term averages, second term averages, cumulative averages and the end-of-course assessment scores were examined to find the degree of predictive validity.

Analysis of data revealed that face validity does not contradict with more objective measures of tests, such as reliability and predictive validity. However, face validity and reliability analyses revealed some important weaknesses in the local testing system. These weaknesses would not have been revealed if the researcher had looked at only face validity, or only reliability, or only predictive validity of the tests. Therefore, it is very important to look at tests from multiple perspectives, and get information from a variety of sources. Additionally, it has been found that the face validity, reliability and predictive validity of the achievement tests administered at Zonguldak Karaelmas University Preparatory School are high in spite of the weaknesses that were revealed in the analysis.

Key Words: Achievement Tests, Face Validity, Reliability, Predictive Validity

ÖZET

ÜNİVERSİTE İNGİLİZCE YABANCI DİL HAZIRLIK OKULU BAŞARI SINAVLARININ GÖRÜNÜŞ GEÇERLİĞİ,GÜVENİRLİĞİ VE YORDAMA GEÇERLİĞİ ARASINDAKİ İLİŞKİ

Funda Küçük

Yüksek Lisans, Yabancı Dil Olarak İngilizce Öğretimi Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. JoDee Walters

Ortak Tez Yöneticisi: Yrd. Doç. Dr. Julie Mathews-Aydınlı

Haziran 2007

Bu çalışma görünüş geçerliği ile güvenirlik ve yordama geçerliği gibi nispeten daha nesnel sınav ölçütleri arasındaki ilişkiyi incelemiştir. Çalışma Zonguldak Karaelmas Üniversitesi Hazırlık Okulu'nda yapılan başarı sınavlarının görünüş geçerliği, güvenirliği ve yordama geçerliğini de incelemiştir.

Bu çalışmada kullanılan veri toplama araçları iki anket ve C- (başlangıç) seviyesindeki öğrencilerin sınav notlarıdır. İlk olarak, sınavların görünüş geçerliği ve güvenirlik derecesini belirlemek üzere okutmanlara ve öğrencilere anketler verilmiştir.Sonra, yordama geçerliği derecesini bulmak amacıyla öğrencilerin birinci dönem ortalamaları, ikinci dönem ortalamaları, genel ortalamaları ve final notları arasındaki korelasyonlara bakılmıştır.

Veri analizi görünüş geçerliğinin, güvenirlik ve yordama geçerliği gibi nesnel sınav ölçütleriyle çelişmediğini ortaya koymuştur. Fakat, görünüş geçerliği ve güvenirlik analizleri yerel sınav sistemindeki bazı önemli kusurları ortaya çıkarmıştır. Eğer araştırmacı sınavların yalnız görünüş geçerliği, ya da yalnız güvenirliği, ya da yalnız yordama geçerliğini inceleseydi bu kusurlar ortaya çıkarılamazdı. Bu nedenle, sınavları çok yönlü incelemek ve çeşitli kaynaklardan bilgi edinmek çok önemlidir. Ayrıca, analiz sonucunda ortaya çıkarılan kusurlara rağmen, Zonguldak Karaelmas Üniversitesi Hazırlık Okulu'nda uygulanan başarı sınavlarının görünüş geçerliği, güvenirliği ve yordama geçerliğinin yüksek olduğu saptanmıştır.

Anahtar Kelimeler: Başarı Sınavları, Görünüş Geçerliği, Güvenirlik, Yordama Geçerliği

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor, Assist. Prof. Dr. JoDee Walters, for her contributions, invaluable feedback and patience throughout the completion of this thesis. I also wish to thank to Assist. Prof. Dr. Julie Mathews-Aydınlı, the director of the MA TEFL program, for her continual support in my studies and for having such a big heart full of love for us.

I am grateful to the former director of Zonguldak Karaelmas University Prep School, Assist. Prof. Dr. Nilgün Yorgancı Furness, for her encouragement and support. I am grateful to the Rector, Prof. Dr. Bektaş Açıkgöz and the former Vice Rector, Prof. Dr. Yadigar Müftüoğlu, who gave me permission to attend this program as well. Furthermore, I would like to express my sincere gratitude to the current coordinator of the Prep School, Okşan Dağlı, and the assistant coordinators, Eda Baki and Mustafa İnan, who provided me with the necessary help to conduct my study.

I am thankful to Selin Marangoz Çıplak, Ayşe Kart, İnan Tekin, İsmail Aydoğmuş and Yalçın Dayı, who felt equal responsibility with me in the distribution and the collection of the questionnaires. I also want to thank my colleagues at Prep School and the former prep class students who agreed to take part in the study, for their willingness. Additionally, I owe thanks to Nuray Okumuş and Evren Köse, who helped me in editing the questionnaire.

I am sincerely grateful to my dearest friend, Özlem Karakaş and her family members, Ali Muhlis Karakaş, İlknur Karakaş and Özge Karakaş, for their unconditional love and support, which motivated me during this challenging process.

My special thanks go to my dearest brother, Muzaffer Küçük, for his deep love, moral support and never-ending trust in me, and my genuine thanks go to my cutest sister, Tuğba Küçük, for her affection, encouragement, understanding and guidance during this busy year. Finally, I wish to thank my devoted mother, Türkan Küçük, for her everlasting love, caring and patience. Without my family, it would have been impossible for me to survive this year.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER I: INTRODUCTION

Introduction

Achievement tests are tests which gather information during, or at the end of, a course of study in order to examine if and where progress has been made in terms of the objectives of teaching (McNamara, 2000, p. 6). In large educational institutions, testing offices, rather than individual teachers, design achievement tests in order to ensure standardization. Unfortunately, a large number of instructors do not trust these tests and the testers (Bachman & Palmer, 1996; Cohen, 1994; Hughes, 2003). Not only instructors, but also test takers and other stakeholders may distrust the tests and the testers.

This situation necessitates validating the tests. The most complicated criterion of an efficient test is validity. It refers to the degree to which inferences drawn from test scores are proper, meaningful and useful in terms of the goals of the test (Gronlund, 1998, cited in Brown, 2004, p. 22). There are both subjective and objective measures of validity. To begin with, 'face validity', which is a subjective measure, entails learning the personal opinions, intuitions and unscientific remarks of instructors, test takers and other stakeholders about the tests. According to Weir (1990),

> if a test does not have face validity, it may not be acceptable to the students taking it, or the teachers and receiving institutions who may make use of it. Furthermore, if the students do not accept the test as valid, their adverse reaction to it may be that they do not perform in a way which truly reflects their ability. (p. 26)

However, some testers consider face validity as irrelevant. Furthermore, these testers dismiss face validity, since it is not based on facts (Alderson, Clapham, & Wall, 1995). On the other hand, predictive validity, which is an objective measure, refers to the degree to which a test can predict the possible future success or failure of the test takers (Bachman, 1990; Hughes, 2003). If a test predicts future success well, it is believed that the inferences drawn from this test are trustworthy. Thus, such a test is labeled as valid. The second objective measure is reliability, which is defined as the degree of consistency between the scores of one test with itself or with another test (Brown et al., 1999, p. 168).

The aim of this study is to find out how well face validity reflects relatively more objective measures of tests: reliability and predictive validity. In order to do so, the perceptions of students and instructors of the face validity of the achievement tests conducted in Zonguldak Karaelmas University Preparatory School were investigated, and the correlation between the two groups' perceptions was explored. Furthermore, students' achievement test scores were compared with one another at various times throughout the academic year to determine the degree of predictive validity. The test construction and testing conditions were also assessed to define the degree of reliability in terms of the performance of the students. Additionally, the study examined whether the current testing system permits scorer reliability or not. Lastly, the correlations between face validity and predictive validity and face validity and reliability were inspected. Thirty English instructors and fifty two students participated in this survey study. Data were collected by distributing one questionnaire to the instructors and

another questionnaire to the students. Achievement test scores obtained from the Preparatory School also served as data.

Background of the Study

Language tests are tests constructed to measure test takers' knowledge of and skills in a foreign language in educational programs. According to Bachman (1990), the fundamental purpose of language tests is to collect information for taking decisions about people, such as students and instructors, and decisions about the program.

Although all language tests collect information for taking decisions, there are differences amongst them. For instance, they differ in terms of their purpose, frame of reference, design, scoring procedure and method (Bachman, 1990, p. 70). In short, there are various test types. Among these test types, achievement tests are employed most frequently in educational institutions. Brown (1996) defines achievement tests as tests which are administered to learn how well students have achieved the instructional goals of a course (p. 14).

However, sometimes achievement test results may not accurately reflect the students' language knowledge and skills. Therefore, constant assessment of achievement tests is needed. One way to do this is to examine the good qualities they possess. Bachman and Palmer (1996) define these qualities as reliability, validity, authenticity, interactiveness, washback impact, and practicality (p. 38).

As Bachman (1990) states, among the good qualities, the fundamental quality to consider while constructing, administering and interpreting language tests is validity (p. 289). In general, validating a test refers to gathering scientific data and logical arguments to show that the test is proper in terms of the goals of the assessment. There

are several validity types, and each validity type entails collecting data in different ways. Predictive validity and face validity are two of these validity types.

Predictive validity indicates that the test predicts the possible future success or failure of the test takers (Bachman, 1990; Hughes, 2003). In other words, it is believed that the inferences made from a test are reliable if the test accurately predicts the success of those who take it. To investigate predictive validity, students' test scores can be correlated with their scores on tests taken some time later.

Face validity is the second type which can be employed to discuss validity evidence. It refers to the degree to which the test seems valid in terms of testing what it has to test (Alderson et al., 1995; Cohen, 1994; Hughes, 2003). Investigation of face validity requires learning the subjective judgments and perceptions of the stakeholders of the tests.

While validity is a fundamental quality of tests, reliability is a precondition for validity because test scores that are not reliable cannot provide suitable grounds for valid interpretation and use (Bachman, 1990). According to Hughes (2003), there are two essential concepts involved in reliability: 'scorers' reliability' and 'reliability in terms of the test takers' performance'. Scorers' reliability refers to the degree to which test scores are free from measurement error (Rudner, 1994). Sources of measurement error for the scorers are time pressure, inefficient rating scales and so on (Alderson et al., 1995, p. 128). The second concept, reliability in terms of the test takers' performance, refers to the extent to which test scores of a group of test takers are consistent over repeated test applications (Berkowitz, Wolkowitz, Fitch & Kopriva, 2000, cited in Rudner & Schafer, 2001). In other words, if the same person took the same test more than once, and if there

is an inconsistency between his or her scores, it can be said that the test has a low reliability level (Hughes, 2003). Some reasons for the inconsistency between the scores of the test takers are unclear instructions, ambiguous questions and so on (Hughes, 2003, p. 44).

Several researchers have conducted studies in an attempt to assess the validity and reliability of various tests. Among these researchers, some have looked at the reliability of tests, some have explored the predictive validity of tests, and some have investigated the face validity and content validity of tests.

First of all, five researchers, namely Brown (2003), Cardoso (1998), Manola and Wolfe (2000) and Nakamura (2006), have looked at the reliability of tests. Brown (2003) explored the scorers' reliability of a speaking test. Cardoso (1998) explored the reliability of the reading section of English language tests administered in the State University of Campinas in Brazil as part of the university entrance examination. Additionally, Manola and Wolfe (2000) explored the reliability of the essay writing section of the TOEFL, investigating whether the essay medium could affect the reliability of the scores and the accuracy of the inferences drawn from these scores. Finally, Nakamura (2006) investigated the reliability of the pilot English placement test developed for Keio University Faculty of Letters in Japan in order to determine what changes were needed in order to arrive at the final version.

Furthermore, some researchers have explored the predictive validity of tests. For instance, Yeğin (2003) examined the predictive validity of the Başkent University English Proficiency Exam (BUEPE) by using Item response theory (IRT) -based ability estimates. Dooey (1999) also explored the predictive validity of tests, investigating the

predictive validity of the IELTS (International English Language Testing System) test as an indicator of future academic success. Next, Ösken (1999) looked at the predictive validity of midterm achievement tests administered at Hacettepe University, Department of Basic English (DBE).

Lastly, some researchers have investigated the face validity and content validity of tests. For example, Ösken (1999), in her previously mentioned study, examined the face validity and content validity of the end-of-course assessment administered at Hacettepe University, Department of Basic English (DBE). The next researcher who investigated both face validity and content validity of tests is Serpil. Serpil (2000) looked at the face validity and content validity of midterm achievement tests, administered at Anadolu University, School of Foreign Languages. Another researcher who looked at the face validity and content validity of tests is Nakamura (2006), who examined the face validity and content validity of a pilot English placement test in his previously mentioned study. However, none of the above mentioned studies have explicitly compared face validity with relatively more objective measures of tests such as reliability and predictive validity.

Statement of the Problem

The involvement of instructors and students in the assessment process has been studied (Bachman & Palmer, 1996), and the validity and reliability of tests administered to measure English knowledge and skills as a second language have also received attention in the literature (Brown, 1996; Davies, 1990; Hughes, 2003; Kunnan, 2000; McNamara, 2000). However, the field still lacks research studies which focus on how

well face validity reflects relatively more objective measures: reliability and predictive validity.

At the testing office of Zonguldak Karaelmas University Preparatory School, achievement tests are prepared by the instructors who work in this office in rotation, in addition to their teaching assignments. I personally worked as a member of the office for three consecutive years, and also served as the assistant director of the testing office during the last two years. My experience suggests that all the testing office members did their best to construct well-designed tests. Nevertheless, I have observed a possible problem with face validity. In other words, the achievement tests were not representing the course content in the eyes of both the students and the instructors. This was because much of the curriculum was not reflected in the exams, which might have led the students to distrust the assessment system. Furthermore, the testing system has never been assessed for validity and reliability. Consequently, some language instructors can be suspicious about the tests and the testers. In fact, their suspicion may be well founded in some respects. However, there is some doubt whether it can be assumed that if a test is not appropriate in the eyes of the stakeholders, it is not valid and reliable. Therefore, I would like to learn whether the opinions of the instructors and students about the tests conducted in this institution correlate with the results of relatively more objective measures such as predictive validity and reliability.

Research Questions

This study addresses the following questions:

1. To what extent do the achievement tests possess face validity?
    - To what extent do the achievement tests represent the course content in the eyes of the instructors?
    - To what extent do the achievement tests represent the course content in the eyes of the students?
    - Is there a difference between the two groups' perceptions of the achievement tests' representativeness of the course content?
2. To what extent do the achievement tests possess reliability?[1]
    - To what extent does the current testing system permit scorer reliability?
    - To what extent do the structure of the tests and the testing conditions permit students to accurately demonstrate their language knowledge and skills?
3. To what extent do the achievement tests possess predictive validity?
    - How well do the achievement tests conducted in the first term predict success in the second term?
    - How well do the achievement tests conducted throughout the year predict success in the end-of-course assessment?
4. How closely does the face validity of the achievements tests reflect the reliability and predictive validity of these tests?

[1] In this thesis, scorers' reliability was determined by asking specific questions to the scorers about scoring practices, and reliability in terms of the test takers' performance was determined by asking specific questions to the students about the structure of the tests and the testing conditions.

Significance of the Study

Practitioners make judgments about language tests by assessing their appeal, or "face validity", due to lack of time, resources or competence. However, no research studies have been conducted on how reliable face validity is. In other words, there is a lack of research in the field of foreign language teaching that focuses on how closely a subjective measure, face validity, reflects relatively more objective measures of a test, such as reliability and predictive validity. Therefore, this study may contribute to the literature. In addition, if it is observed that face validity reflects reliability and predictive validity well at the end of this study, administrators and testers may place more stock in the opinions of the stakeholders. On the other hand, if the opposite is observed, relatively more objective measures such as reliability and predictive validity may be employed to assess the tests rather than solely relying on face validity.

At the local level, this study aims to learn the attitudes of the instructors and students towards the current assessment system in Zonguldak Karaelmas University and evaluate the achievement tests conducted in the same institution. The institution will benefit from the study since the strengths and weaknesses of the existing testing system will be defined in the process. The observed weaknesses may lead to changes in the system, and the strengths may serve as an example to other institutions. This study may also lead to further studies on validity and reliability of tests administered to measure English knowledge and skills as a second language.

Key Terminology

Stakeholders: People who are interested in the administration or impacts of a particular test, such as the test takers, their instructors and parents/ families, the test designers and their customers, the receiving institutions (e.g., Ministries of Education and of Immigration) in the case of a selection test (Brown et al., 1999, p. 184).

Conclusion

In this chapter it was aimed to introduce the study through a statement of the problem, research questions, the significance of the study, and the key terms. Moreover, the general frame of the literature review was drawn.

The second chapter of the study will be a review of the literature which includes the definition, uses, types and good qualities of language tests, and previous research studies conducted on the validity and reliability of language tests. In the third chapter, setting, participants, instruments, data collection procedures and data analysis will be presented. In the fourth chapter, the data analysis procedures and the findings will be reported. Lastly, the fifth chapter will display the overview of the study, discussion of findings, pedagogical implications, limitations of the study, and implications for further research.

CHAPTER II: LITERATURE REVIEW

Introduction

This study attempts to investigate how well face validity reflects relatively more objective measures: reliability and predictive validity. The study also aims to examine the predictive validity, face validity and reliability of tests administered at Zonguldak Karaelmas University Preparatory School. This chapter of the thesis reviews the literature on the uses, types and good qualities of language tests, and previous research studies conducted on reliability and validity of language tests.

Uses of Language Tests

Language tests are tests used to measure language skills or competence, and a defining characteristic of language tests is that they include specified tasks through which language skills are elicited (Bachman, 1990; Brown et al., 1999). Language tests generally offer information for taking decisions about individuals and programs (Bachman, 1990; McNamara, 2000). The decisions about individuals include decisions about students and teachers.

To begin with, tests are used to admit and place students into appropriate courses (Bachman, 1990; Hughes, 2003; McNamara, 2000; Norris, 2000). Bachman (1990) states that tests are also used to assess teachers' performance by administrators. Since some teachers are not native speakers of the language, administrators wish to obtain information about these teachers' language proficiency before employing them (p. 61).

Furthermore, language tests are used to make decisions about the programs. These tests define the degree to which course objectives are being accomplished, and demonstrate the effectiveness of syllabus design and pedagogy (Bachman, 1990; Hughes, 2003; McNamara, 2000; Norris, 2000). In other words, the performance of students on tests provides evidence of the extent to which the desired goals of the program are being achieved (Bachman, 1990, p. 62). If it is observed that the course objectives are not being achieved, decisions can be taken to change the existing program.

Language tests are also used to gather data in research studies which are related to the nature of language proficiency, language processing, language acquisition, and language teaching (Bachman, 1990, p. 67). In such research studies, language tests are used to provide information for comparing the performances of individuals with different characteristics or under different conditions of language acquisition or language teaching, and for testing hypotheses about the nature of language proficiency (Bachman, 1990; McNamara, 2000).

Apart from these uses Cohen (1994, cited in Norris, 2000) indicates that,

> language tests are used to diagnose areas of learner need or sources of learning difficulties, reflect on the effectiveness of materials and activities, encourage student involvement in the learning process, track learner development in the L2, and provide students with feedback about their language learning progress for further classroom-based applications of language tests. (p. 3)

Kinds of Language Tests

Achievement tests are the focus of this study. However, other kinds of language tests will also be discussed since such a classification may help the stakeholders to assess the appropriateness of the tests they administer, construct or take, and to gain insights about testing.

Proficiency tests

Proficiency tests are tests which evaluate the general knowledge or abilities compulsory or necessary to enter or to be exempt from a group of similar institutions (Brown, 1996, p. 10). They are not based on a specific syllabus of study followed by test takers in the past, but rather try to measure test takers' general level of language mastery (Brown et al., 1999; Hughes, 2003; Kuroki, 1996).

Proficiency tests differ in nature, since the term 'proficient' has two different meanings. In some proficiency tests, 'proficient' means being adept at the language for a particular purpose (Heaton, 1990; Hughes, 2003). In other words, such proficiency tests look forward to the actual ways in which the candidates will use English in the future time (Heaton, 1990; Hughes, 2003). Thus it is possible to say that these tests measure the candidates' proficiency in various specific disciplines such as life sciences, medicine, social studies, physical sciences and technology (Heaton, 1990, p. 17). The Interuniversity Foreign Language Examination (ÜDS) administered in Turkey which has two forms (medicine and social sciences) can be given as an example for this category.

In other proficiency tests no discipline or program is borne in mind while constructing the test. In these proficiency tests, the concept of 'proficiency' is more general and covers all disciplines (Brown, 1996; Hughes, 2003). British examples of

these tests are the Cambridge First Certificate in English examination (FCE) and the Cambridge Certificate of Proficiency in English examination (CPE) (Hughes, 2003, p. 12).

Proficiency tests can affect students' lives greatly especially when entrance issues are concerned. Therefore, proficiency decisions should never be considered as something unimportant. Additionally, taking quick and careless decisions about these tests is highly unprofessional (Brown, 1996, p. 11).

Placement tests

Placement tests, as their name suggests, are employed to provide data that will help to place students at the level of the teaching program which is most suitable to their abilities (Bailey, 1998; Hughes, 2003). This allows the students to be grouped according to their language ability at the beginning of a course (Brown, 1998; Heaton, 1990). Teachers benefit from this grouping practice because it enables their classes to consist of students with rather similar levels. As a result, teachers can give their full attention to the problems and learning points appropriate for that level of students (Brown, 1996, p. 11).

To be most efficient, placement tests should include the characteristics of the teaching context (e.g., the proficiency level of the classes, the methodology and the syllabus type (Bailey, 1998; Brown, 1996; Brown et al., 1999; Heaton, 1990; Hughes, 2003). This means that a grammar placement test, for example, may not be most suitable, if the syllabus is task-oriented (Brown et al., 1999, p. 145). Brown (1996) states that if there is an inconsistency between the placement test and the syllabus, the danger is that the groupings of similar ability levels will simply not occur (p. 13). Therefore, such placement tests may not serve their purposes.

Finally, testers should design the placement tests in a way which will enable the teachers to sort students into groups easily by just looking at the scores (Heaton, 1990, p. 15). With this purpose in mind, testers can prepare scales showing which student should go to which level. On the other hand, if placement tests are not designed well, they can be a burden for the teachers rather than facilitating their business.

Diagnostic tests

Diagnostic tests are designed to determine whether instructional objectives of courses have been achieved, like achievement tests. However, they differ from achievement tests since they are administered at the beginning or middle of a course, not at the end (Brown, 1996, p. 15). Furthermore, it should be kept in mind that diagnostic tests are mostly constructed in parallel to the syllabuses of specific classes, like achievement tests (Bailey, 1998, p. 39).

Diagnostic tests are often used to figure out the strengths and weaknesses of students, and this is done for a number of purposes (Bailey, 1998; Brown, 1996; Brown et al., 1999; Heaton, 1990; Hughes, 2003). First of all, figuring out the strengths and weaknesses of students helps teachers recognize the areas where remedial instruction is essential (Brown et al., 1999; Heaton, 1990). Next, they are invaluable for self-instruction. This is because diagnostic tests indicate the gaps in the students' language domain. Thus students are directed to the sources of information, exemplification and practice relating to their gaps before it is too late (Brown, 1996; Hughes, 2003).

Brown (1996) claims that the most efficient diagnostic tests are those which report the students' performance of each objective by percentages (p. 15). However, very few tests are constructed especially for diagnostic purposes. In practice

achievement or proficiency tests are widely used for diagnostic purposes, because it is difficult and time consuming to design a detailed diagnostic test (Brown et al., 1999; Heaton, 1990). Unfortunately, teachers may not obtain reliable information from achievement or proficiency tests. Hughes (2003) explains the reason for this as follows:

> It is not so easy to obtain a detailed analysis of a student's command of grammatical structures- for example, whether she or he had mastered the present perfect/past tense distinction in English. In order to be sure of this, we need a number of examples of the choice the student made between the two structures in every different context that we thought was significantly different and important enough to warrant obtaining information on. A single example of each is not enough, since a student might give the correct response by chance. (p. 15)

Finally, Heaton (1990) indicates that it is only necessary to give remedial teaching for the whole class, if at least a quarter of the class has difficulty with a specific aspect of the language (p. 13). In other words, if fewer than 25% of students have problems concerning the language, teachers can treat their weaknesses in groups or privately.

Achievement tests

Achievement can be simply defined as "the mastery of what has been learnt, what has been taught or what is in the syllabus, textbook, materials etc." (Brown et al., 1999, p. 2). However, it is not that easy to define achievement tests since there are two approaches for constructing achievement tests: the alternative approach and the syllabus-content approach.

According to the alternative approach, achievement tests are tests conducted to show how well students have accomplished the instructional objectives (Brown, 1996, p. 14). In other words, achievement tests should be based on the objectives of a course

(Bailey, 1998; Hughes, 2003). Hughes (2003) explains the advantages of this approach as follows:

> First it compels course designers to be explicit about objectives. Secondly, it makes it possible for performance on the test to show just how far students have achieved these objectives. This in turn puts pressure on those responsible for the syllabus and for the selection of books and materials to ensure that these are consistent with the course objectives. Finally, tests based on objectives work against the perpetuation of poor teaching practice, something which course-content- based tests fail to do. (p. 13)

On the other hand, according to the syllabus-content approach, the content of a final achievement test should match the course syllabus or the books and other course materials (Brown et al., 1999; Heaton, 1990; Hughes, 2003). Heaton (1990) claims that,

> if teachers set an achievement test for several classes as well as their own class, they should take care to avoid measuring what they themselves have taught – otherwise they will favor their own classes. By basing their test on the syllabus or the course book rather than their teaching, their test will be fair to students in all the classes being tested. (p. 14)

Additionally, Brown et al. (1999) argue that the opinion that an achievement test should measure success on course objectives rather than on course content is not popular, since such an approach spoils the achievement-proficiency distinction (p. 2). In other words, the most apparent difference between achievement and proficiency tests is that the former measures the success of students with reference to a specific course syllabus. However, while constructing the latter no particular syllabus is taken into consideration. Consequently, if this distinction disappears, it will be hard to differentiate between achievement and proficiency tests.

As can be understood from the above mentioned statements, both approaches to achievement tests have some shortcomings. However, since information on student

achievement is crucial in teaching, teachers should decide between these two approaches by defining the needs of their students and the instructional environment.

Regardless of which approach they are based on, achievement tests are used for teaching and learning purposes. First, they are used to make changes in the curriculum and to assess those changes (Brown, 1996; Weir, 1995). In addition to this, they are used to define how successful students have been in mastering the objectives (Hughes, 2003; Brown, 1996; Weir, 1995). Weir (1995) states that this knowledge helps teachers decide whether to move on to the next unit. For example, if teachers see that students have learned a unit completely, they feel free to proceed on to the next one (p. 167). Achievement tests are also conducted at the end of a learning session, a school year or a whole school or college career, and the results obtained are often used for decision taking purposes, especially selection (Brown et al., 1999, p. 2).

Good Qualities of Tests

Bachman and Palmer (1996) indicate that 'usefulness' is the most significant feature of tests, and six test qualities contribute to test usefulness: reliability, validity, authenticity, interactiveness, washback impact and practicality (p. 38). They further add that since all these qualities contribute to test usefulness they cannot be evaluated separately. Consequently, mentioning good qualities of language tests at this phase may not only inform the stakeholders but also enable them to see how good qualities of tests interact with each other. Although all six qualities will be discussed here, special attention will be paid to validity and reliability of tests. The reason for this is that achievement tests administered at Zonguldak Karaelmas University will be examined in

terms of their validity and reliability in this study, and consequently it will be clear in the end whether these tests serve their intended purposes or not.

Validity

Validity in general refers to the properness of a test or any of its constituent parts as a measure of what it is supposed to measure (Alderson et al., 1995; Brown, 1996). In other words, if a test measures what it should measure, it can be considered as valid (Alderson et al., 1995; Bachman, 1990; Brown, 1996; Brown et al., 1999; Hughes, 2003; Kuroki, 1996). Kuroki (1996) supports this point of view by saying,

> if a test designed to measure students' listening ability requires candidates to write complete sentences in response to a question, the validity may be in question because such a tests in fact measures not only candidates' listening ability but also their grammatical knowledge. (p. 7)

Tests can be validated in various ways or by using various methods. For that reason, different validity types have been established to describe these different ways (Alderson et al., 1995, p. 171). Validity types can be listed as follows: face validity, construct validity, content validity, criterion related validity, predictive validity, and concurrent validity. It is generally wished to validate tests by using as many of these types as possible. This is because the trust put in a test is directly proportional to the amount of evidence obtained to validate it (Alderson et al., 1995; Brown et al., 1999).

On the other hand, the modern approach contradicts the above mentioned view. The modern approach considers validity as a unitary concept, and avoids categorizing it according to the methods it employs (Bachman, 1990; Brualdi, 1999). While it is common to talk about content, criterion and construct validities as distinct types of validity, they should be considered as complementary types of evidence according to

this approach (Bachman, 1990, p. 243). Furthermore, Messick (1989, cited in Kunnan, 1998) , in order to emphasize the unitary nature of validity, has described it "as an integrated evaluative judgment of the degree to which empirical evidence and the theoretical rationale support the adequacy and the appropriateness of inferences and actions based on test scores or other modes of assessments" (p. 19).

Another difference between the traditional and modern validity approaches is that the modern view has diverted the focus of validity from the test itself to the use of test scores (Kenyon and Van Duzer, 2003). However, the traditional conception of validity has always been criticized for not paying attention to the meaning of the scores as a basis for action and the social implications of score use (Messick, 1996b, cited in Brualdi, 1999).

Finally, Bachman (1990) concludes that,

> it is still necessary to gather information about content relevance, predictive utility, and concurrent criterion relatedness, in the process of developing a test. However, it is important to recognize that none of these by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores. And while the relative emphasis of the different kinds of evidence may vary from one test use to another, it is only through the collection and interpretation of all relevant types of information that validity can be demonstrated. (p. 237)

Face validity

Face validity is "the surface credibility or public acceptability of a test" (Ingram, 1977, cited in Bachman, 1990, p. 287). It refers to the degree to which the test is valid in the eye of the examinees who take it, the administrative staff that make judgments on its use, and other technically untrained observers (Alderson et al., 1995; Anastasi, 1982, cited in Weir, 1990; Brown et al., 1999; Davies, 1990). In other words, face validity

relies on the subjective perceptions of untrained stakeholders (Alderson et al., 1995; Davies, 1990).

In order to possess face validity a test should meet some criteria. First, it should have content validity in the eyes of the stakeholders. Brown (2004) notes that if a test measures knowledge and skills which are directly related to the course content, chances of achieving face validity improve (p. 27). Next, the test should measure what it has to measure (Alderson et al., 1995; Brown et al., 1999; Cohen, 1994; Hughes, 2003). For instance, a test which claims to measure pronunciation ability but does not urge the test takers to speak must be considered as inappropriate in terms of face validity (Hughes, 2003, p. 33).

There are two opposing views about face validity. Some assessment experts strongly believe that face validity is important in testing. For example, Davies (1990) states that tests possess a public essence. Therefore, they should have face validity. According to him, if there is an inconsistency between face validity and other types of validity, face validity should be the first to be looked for (p. 23). Alderson et al. (1995) add that face validity is crucial since if a test does not seem valid, the stakeholders may not pay much attention to it and its implications. Moreover, the examinees may not demonstrate their full potential under such a condition (p. 173). Anastasi (1982, cited in Weir, 1990, p. 26) also took a similar line by saying "especially in adult testing it is not sufficient for a test to be objectively valid. The test also needs face validity to function effectively in practical situations."

On the other hand, some assessment experts disregard face validity. For instance, several researchers (Bachman & Palmer (1981a), E. Ingram (1977) and Lado (1961), all

cited in Weir, 1990) have described face validity as useless (p. 26). Cronbach (1984, cited in Bachman, 1990), who agrees with the above mentioned experts, has the following to say about face validity:

> Adopting a test just because it appears reasonable is bad practice; many a good looking test has had poor validity. Such evidence warns against adopting a test solely because it is plausible. Validity of interpretations should not be compromised for the sake of face validity. (p. 286)

Alderson et al. (1995) indicate that tests can be examined for their face validity in several ways, for instance, by conducting questionnaires or interviewing the stakeholders about their perceptions of the tests (p. 173). As a result, it can be concluded that the properness of test components can be defined by analyzing both quantitative (data gathered from the questionnaires) and qualitative (data gathered from the interviews) data.

Construct validity

Construct validity is a type of validity which examines how closely a test measures a theoretical construct or attributes (Alderson et al., 1995; Anastasi, 1982, cited in Weir, 1990; Brown et al., 1999; Hughes, 2003). Brown (2004) exemplifies the case as follows:

> Let's suppose you have created a simple written vocabulary quiz which covers the content of a recent unit and asks students to correctly define a set of words. Your chosen items may be a perfectly adequate sample of what was covered in the unit, but if the lexical objective of the unit was the communicative use of vocabulary, then the writing of definitions certainly fails to match the construct of communicative language use. (p. 25)

Alderson et al. (1995, p. 195) define the procedures to evaluate the construct validity as follows:

- Correlate each subtest with other subtests.
- Correlate each subtest with total test.
- Correlate each subtest with total minus self.

Construct validity is seen as an umbrella term which comprises criterion and content validity (Bachman, 1990). For that reason, it seems appropriate to explain both criterion and content validity hereby to provide a better understanding of construct validity.

Content validity

Content validity is the extent to which the test content matches the target domain to be measured. This type of validity requires the researchers to analyze the test content systematically (Alderson et al., 1995; Brown, 1996; Brown et al., 1999; Hughes, 2003; Anastasi, 1982, cited in Weir, 1990). Thus, a common way to assess the content validity of a test is to compare its content with a teaching syllabus or a domain specification (Alderson et al., 1995; Brown et al., 1999; Davies, 1990, Hughes, 2003). Brown (2004) indicates that,

> the most feasible rule of thumb for achieving content validity is to test performance directly, not indirectly. For example, if the test is intended to test learners' oral production of syllable stress and the given task is to have learners mark stressed syllables in a list of written words, this is considered as indirect testing of oral proficiency. A direct test of syllable production requires that students actually produce target words orally. (p. 24)

Both face validation and content validation processes require evaluating the content of the tests. However, while face validity relies on the subjective perceptions of untrained stakeholders, content validation necessitates learning the objective judgments

of testing experts (Alderson et al., 1995; Davies, 1990). This is the basic difference between face validity and content validity.

Criterion related validity

Brualdi (1999) states that in terms of an achievement test, criterion-related validity refers to the degree to which a test can be employed to make inferences concerning achievement (p. 1). In order to assess criterion-related validity, the degree to which the test scores agree with one or more outcome criteria is defined. (Bachman, 1990; Brown, 2004; Brualdi, 1999; Hughes, 2003).

There are two types of criterion-related validity: predictive validity and concurrent validity (Brown, 2004; Brown et al., 1999; Hughes, 2003; Weir, 1990). In the case of predictive validity the criterion is some future performance which teachers want to forecast. However, in the case of concurrent validity the criterion manner and the administration of the test occur at the same time (Bachman, 1990; Brown, 1996; Weir, 1990).

Predictive validity

Predictive validity is the extent to which a test can forecast test takers' future success (Hughes, 2003; Rudner, 2004). Additionally, in terms of an achievement test, predictive validity refers to the degree to which a test can be employed to make inferences concerning achievement (Rudner, 2004, p. 3).

Placement tests, admission tests and language aptitude tests can be explored for their predictive validity (Brown, 2004, p. 25). Furthermore, such validation practices can also be employed to assess proficiency tests (Hughes, 2003; Davies, 1990). Thus, examining the extent to which a proficiency test foretells a student's ability to get

through a graduate course at a British university can be considered as an example of predictive validation practice (Hughes, 2003, p. 29).

Alderson et al. (1995, p. 194) suggest the following to measure predictive validity:

- Correlate students' tests scores with their scores on tests taken some time later.
- Correlate students' test scores with success in final exams.
- Correlate students test scores with other measures of their ability taken some time later, such as language teachers' assessments.
- Correlate students' scores with success of later placement.

Unfortunately, one might encounter some problems while examining the predictive validity. The first problem is that proficiency levels of the students may increase between the first and second tests. Next, there may not be another English test in the study setting with which to correlate the results of the test in question (Alderson et al., 1995, p. 181). Hughes, (2003) has also mentioned another problematic aspect of predictive validity and has defined the validity coefficient as follows:

> How helpful is it to use final outcome as the criterion measure when so many factors other than ability in English (such as subject knowledge, intelligence, motivation, health and happiness) will have contributed to every outcome? For this reason, where outcome is used as the criterion measure, a validity coefficient of around 0,4 (only 20 per cent agreement) is as high as one can expect. (p. 29-30)

Concurrent validity

Concurrent validation necessitates comparing the test scores with another measure for the same tests takers taken nearly simultaneously. This other measure can be scores obtained from a similar copy of the same test or from another test (Alderson et al., 1995, p. 177). Brown et al. (1999) state that it is only suitable to employ an existing

test as the criterion, if it is a simplified version of the original one (e.g., a tape-based test as a substitute for 'live' oral proficiency interview) (p. 30).

Alderson et al. (1995, p. 193) indicate the ways to evaluate the concurrent validity as follows:

- Correlate the students' test scores with their scores on other tests.
- Correlate the students' test scores with teachers' rankings.
- Correlate the students' test scores with other measures of ability such as students' or teachers' ratings.

Reliability

Reliability is the degree to which the test scores of the test takers are consistent (Bachman & Palmer, 1996; Cohen, 1994; Kenyon and Van Duzer, 2003; Rudner, 1994; Shohamy, 1997). Rudner (1994, p. 3) emphasizes the significance of reliability by saying "fundamental to the evaluation of any instrument is the degree to which test scores are consistent from one occasion to another and are free from measurement error." It can be understood from this statement that there are two constituents of test reliability: 'the performance of the test takers from one occasion to the other' and 'scorers' reliability'. Kenyon and Van Duzer (2003, p. 5) explain the former as follows: if a test taker takes a test and takes the same test after a certain amount of time, the test taker should get about the same score on both occasions. Additionally, according to Brown (2004) 'scorers' reliability' refers to the extent to which different scorers yield consistent scores from the same exam paper.

To begin with, the following have been suggested to ensure 'reliability in terms of the test takers' performance'.

1. Prepare items which are independent. If items in a test are dependent on each other, most probably a test taker who cannot answer a specific question will not be able to answer another question as well (Hughes, 2003, p. 44). Hughes further adds that as this is the case, the performance of the test taker will be reduced.

   Moreover, if dependent items are employed, similar knowledge will be tested. By doing so, extra information about the test taker's performance will not be obtained which will make the test less reliable.

2. Do not prepare tests that are too long. Brown (2004) states that if the tests are too long, the test takers may become tired towards the end of the test. Thus, they may answer some of the questions incorrectly (p. 22). On the other hand, Hughes (2003) argues that if the test will be employed to take important decisions, it should be longer. He believes that accurate information can only be obtained through longer tests (p. 45).

   When these two opposing views are concerned it can be concluded that tests should not be too short or too long. Furthermore, the length of a test can be determined in terms of the importance of the decisions which will be taken after the administration of the test.

3. Provide explicit and unambiguous instructions (Axman, 1989; Genesee & Upshur, 1996; Hughes, 2003). Genesee and Upshur (1996) state that the instructions should be as simple and clear as possible, otherwise they will not serve their real purpose and will be another testing device from the point of the test takers (p. 201).They also indicate that time given to the test takers by the

instructors to complete the test should be stated in tests. In this way, the test takers can use their time more efficiently.

4. Leave out items which do not differentiate between weaker and stronger test takers. Otherwise, the reliability of the test will be spoiled (Hughes, 2003, p. 45). On the other hand, Hughes claims that some easy items should also be included at the beginning of the tests so as not to frighten the test takers in advance.
5. Do not write ambiguous items (Axman, 1989; Brown, 2004; Hughes, 2003). Axman (1989) states that if the language in a test is ambiguous, the test takers may misinterpret the questions. Thus, they may not be able to demonstrate their language knowledge fully (p. 2).
6. Make sure that tests are well laid out (Genesee & Upshur, 1996; Hughes, 2003, Weir, 1995). Genesee and Upshur (1996) suggest the following to ensure that tests are well laid out:
    a) Ensure that the test is legible.
    b) Make sure that pictures and other graphic designs are clear and easy to interpret (p. 203, 243-244).

Apart from the above mentioned issues, suitable testing conditions should be provided in order to ensure reliability (Brown, 1996; Brown, 2004). Brown (1996, p. 189) has defined these testing conditions as "location, ventilation, noise, lighting and temperature." The location of test takers may affect their scores. For instance, test takers sitting far from the cassette player may not hear well in a listening exam. For this reason, they might show a bad performance. Similarly, in testing environments with little air,

test takers might be less successful. Distracting sounds may also lower the test takers' performance. The reason for this is that the test takers will most probably find it hard to concentrate on the task they are doing in noisy places. Additionally, the amount of light can change in different parts of a room (Brown, 2004, p. 21), and darkness may reduce the performance of the test takers. Lastly, if the testing environment is too cold or too hot, test takers may not be able to accurately demonstrate their language knowledge and skills.

Other researchers have also made some suggestions as to how suitable testing conditions can be provided for the test takers. These suggestions are as follows:

1. Ensure that test takers are accustomed to the format and the test techniques. If test takers are experienced about such issues, they will be able to perform better and show their real capacity (Hughes, 2003, p. 47).
2. Make sure that the timing defined for a specific test is appropriate. If it is too short, the test takers' best performance may not be elicited (Brown, 2004; Weir, 1995). Conversely, if the time allocated for the test is too long, tests takers may attempt to cheat, and for that reason the obtained scores may not reflect their real performance.
3. Provide standard timing conditions for all classes which take the same test (Hughes, 2003, p. 48).
4. Make sure that information about how much the given test will affect the students' final grade (weighting of the test) is always announced (Genesee & Upshur, 1996, p. 201).

Hughes (2003) has put forward the following to ensure the 'scorers' reliability' which is the second constituent of reliability:

1. Write items which promote objective scoring: One way to avoid subjectivity is by structuring the tests takers' answers by presenting a part of it. In other words, the test takers may be asked to write a single word as the correct response rather than writing a full sentence. For instance, instead of asking 'What is closely related to success?' the question can be as follows:

   ……..is closely related to success. (Answer: motivation)

2. Prepare a detailed answer key: There can be more than one answer for some questions. Then, the testers should transcribe all the possible answers into the key. Furthermore, there are some questions which might cause disputes among the scorers.

   For example: He is not a reporter. He cannot interview the singer.

   If he is a reporter, he could interview the singer.

   As you see, the first part of the response is incorrect; however the second part is correct. The scorers may not be able to decide what to do in this case. Therefore, testers should identify such questions at the outset and indicate how to mark them.

3. Make sure that the scorers are trained: This is important especially where subjective scoring is concerned.

4. Identify the tests takers by numbers instead of their names: Some teachers may be inclined to favor some students or they may be prejudiced against some students. Therefore, more objective scoring can be ensured if teachers do not know whose paper they are scoring.

5. Appoint multiple scorers: If the exam is a high stakes exam (the scores obtained from high stakes exam are generally used to take very important decisions), it is ideal to employ more than one scorer. It is also advisable to appoint multiple scorers when scoring is subjective (Hughes, 2003, p. 45-46, 48-50).

In addition to the items mentioned above, Weir (1995) recommends having standardization meetings before scoring the papers (p. 27). The reason for this is that the testers may not have anticipated and transcribed all the possible answers into the key, and thus the scorers might experience difficulties. These standardization meetings can help overcome the difficulties which might be encountered during scoring.

It is also highly advisable to assign invigilators or scorers to the classes other than their own teachers. The reason for this is that when their invigilators are their own teachers, students feel more secure and become more inclined to cheat. The second concern is while marking, teachers sometimes find it hard to score their own students' papers objectively.

The last significant issue which should not be overlooked is the time allocated for scoring. In my opinion, the time given for marking should not be too short. If the allocated time is not enough, then the scorers may be pressurized and inclined to mark the papers improperly, and such a manner might spoil the scorers' reliability.

Practicality

Practicality is the degree to which the existing resources meet the resources that are needed in the design, construction and administration of a test. This relation can be shown as follows:

Practicality= $\frac{\text{Existing resources}}{\text{Needed resources}}$ (Bachman & Palmer, 1996, p. 36).

Bachman and Palmer further add that if practicality ≥ 1, the test can be considered as practical. Conversely, if practicality < 1, it can be deduced that the test is not practical.

According to Kuroki (1996) if a test is easy and cheap to develop, conduct and score, it is practical. For instance, a one hour interview which tests speaking skills in crowded classes cannot be labeled as practical (p. 8). Brown et al., (1999) talk about the concept of 'practicality' in a similar manner as follows:

> The term practicality covers a range of issues, such as the cost of development and maintenance, test length, ease of marking, time required to administer the test (individual or group administration), ease of administration (including availability of suitable interviewers and raters, availability of appropriate room or rooms) and equipment required (computers, language laboratory, etc.). (p. 148)

Practicality is a significant consideration which testers should not overlook. The reason for this is that, no matter how valid and reliable tests may be, if they are not practical, it is more appropriate not to conduct them (Brown et al., 1999, p. 148).

On the other hand, Genesee and Upshur (1996) believe that although practicality is important, tests should not be chosen only because they are practical. Technically, reliability and especially validity are more valuable than practicality, and without validity tests are useless (p. 57).

Washback

The term ‘washback’ refers to the effects of testing on teaching and learning (Hughes, 2003, p. 1). It is also known as backwash among language testing specialists.

Washback is generally considered as being either positive (beneficial) or negative (harmful) (Taylor, 2005). If a test promotes learning and teaching, it will give rise to positive washback. For instance, Brown et al. (1999) indicate that adding an oral interview section to an examination may promote conversational language use in the classroom (p. 225). On the other hand, if the test hinders learning and teaching, it will lead to negative washback. Brown and Hudson (1998, p. 667) state that “if assessment procedures in a curriculum do not respond to the curriculum’s goals and objectives, the tests are likely to create a negative ‘washback’ effect on those objectives and on the curriculum as a whole.” In other words, if there is an inconsistency between the tests and the content and the objectives of the courses, negative washback might occur.

Saif (2006) indicates that positive washback can be improved by conducting washback investigations which examine the test development process. Saif further adds that these investigations should be conducted in a way which enables the researcher to define the learners’ needs and the stakeholders’ goals and strategies from the very beginning (p. 3).

Authenticity

Authenticity is the extent to which test tasks pertain to real life language use (Brown & Hudson, 2002; Halleck & Moder, 1995; Hoekje & Linnel, 1994; Lewkowicz, 2000). Bachman and Palmer (1996) indicate the significance of authenticity by saying,

> we consider authenticity to be an important test quality because it relates the test task to the domain of generalization to which we want our score interpretations to generalize. Authenticity thus provides a means for investigating the extent to which score interpretations generalize beyond performance on the test to language use in the TLU [Target language use] domain. (p. 23-24)

Brown (2004, p. 28) indicates that in order to determine the degree of the authenticity of the tests the following features can be assessed:

- The language in the tests should be as natural as possible.
- Items should be contextualized rather than being isolated.
- Topics should be meaningful and interesting for the learners.
- Tasks should represent or nearly represent real-life tasks.
- Some thematic organization should be provided for the items, e.g. a story.

However, Bachman (1990), who recognized the complexities of authenticity, argues that authenticity is not absolute. Alderson et al. (1996, cited in Cumming & Maxwell, 1999) also agree with this idea, and they suggest that tasks are not necessarily either authentic or inauthentic but lie on a continuum which is determined by the extent to which the assessment task related to the context in which it would be normally performed in real-life. Finally, Spolsky (1985, cited in Bachman, 1990) criticizes authenticity by saying,

> however hard the tester might try to disguise her/his purpose in a speaking test, it is not to engage in genuine conversation with the candidate, ….but rather to find out something about the candidate in order to classify, reward or punish her/him. (p. 320-321)

Interactiveness

Interactiveness is defined as the amount and type of involvement of the test taker's distinctive characteristics while accomplishing a test task (Bachman & Palmer, 1996). Spence-Brown (2006, p. 3) states that "these characteristics consist of language skills, thematic knowledge and efficient schemata."

Since it is a significant test quality, some testers try to ensure interactiveness while constructing their tests. Purpura (1995) indicates that these testers should design tasks which require the test takers to exercise their thematic, language and strategic (metacognitive, cognitive, social and affective strategies) knowledge to provide a genuine interaction (p. 5).

Unfortunately, although some tests may seem to be interactive, they may turn out not to possess such a quality. Spence-Brown explains the issue as follows:

> In some cases a study of the discourse and the results of de-briefing interviews reveal almost no interaction of the student with the interviewee. Some of the students do not react to or even comprehend the interviewees' responses, although, interestingly, in some cases they are able to fake the appearance of interaction using pre-prepared responses. On the surface, the language they produce appears fluent and often appropriate. However, the lack of authenticity in the interactions (compared with similar interviews in real life) threatens the validity of the test as evidence of oral interaction skills. (p. 9)

Research Studies on Validity and Reliability

Several research studies have already been conducted in an attempt to assess the validity and reliability of tests (Brown, 2003; Cardoso, 1998; Dooey, 1999; Manola & Wolfe, 2000; Nakamura, 2006; Ösken, 1999; Serpil, 2000; Yeğin, 2003), and examining how other researchers have assessed various kinds of tests can shed light on the process which was followed in this research study.

First of all, some researchers have looked at the reliability of tests. To begin with, Nakamura (2006) investigated the reliability of the pilot English placement test developed for Keio University Faculty of Letters in Japan in order to determine what changes were needed in order to arrive at the final version. The test was taken by 809 freshman university students who were enrolled in the Faculty of Letters. The reliability was verified by the results of the internal consistency coefficient (Classical Test Theory: 0.8) and also by the information pertaining to the very few misfitting items of the test (Item Response Theory).[2] Nakamura concluded that the reliability of the pilot proficiency test was partially supported. In other words, the reliability of the test was explained relatively convincingly.

Cardoso (1998) also explored the reliability of tests, examining the reliability of the reading section of English language tests administered in the State University of Campinas in Brazil as part of the university entrance examination. Two reading tests - one from Unicamp's (The State University of Campinas' Entrance Exam) 1994 English exam and the other from Unicamp's 1995 English exam (in 1995 the exam was shortened)- which explicitly favored authenticity and proficiency were employed in the study. In this study the internal consistency of test scores obtained from Unicamp's 1994 and 1995 English exams was statistically analyzed. The data gathered for the study consisted of individual item scores and total test scores of all students taking the English tests, which amounted to 16,813 students in 1994 and 11,378 in 1995. It was observed

[2] It might be useful to explain classical test theory and item response theory a bit at this point. According to the classical test theory an observed score (on a test) consists of a true score and an error score, and the aim of a test is to ensure scores which are as close as possible to true scores. For that reason, a lot of effort is put into test construction to promote test reliability. On the other hand, IRT is a model which has gone beyond classical test theory, and it allows expression of the relationship of item difficulty and individual ability within a single framework (Brown et al., 1999, p. 22).

that both tests were reliable (the reliability of the test administered in 1994: 0.912, the reliability of the test administered in 1995:0.83). It was also aimed to find out the effects of the modification of test length on the reliability coefficient in this study. The researcher stated that the smaller reliability coefficient found for the 1995 exam did not seem to imply that the shorter test was substantially less reliable, and the 1995 reliability being smaller than the 1994 coefficient only substantiated that the longer tests should have a higher reliability. In order to confirm this, the standard error of measurement difference between the two tests (8.16 points) was used to perform a type of significance test, and it was found that the difference between the mean scores of the exams was not statistically significant. Thus, Cardoso concluded that both tests were reliable.

Brown (2003) investigated a different aspect of test reliability, that of scorer reliability. The aim of this study was to find out how different strategies used by different interviewers resulted in qualitatively different performances in (and hence ratings for) the two interviews. The subjects were two interviewers who differed significantly in terms of their difficulty, and a single candidate. In order to find two interviewers who employ different strategies while interviewing the students, the researcher analyzed the IELTS Speaking Module interviews which formed the basis of a previous study (Brown & Hill, 1998). After the analysis Brown chose Pam (the easiest interviewer) and Ian (the most difficult interviewer). The two interviews were conducted on the same day and involved the same candidate, 'Esther'. Pam used 'topic priming' in an attempt to make the upcoming interview question understandable. She also used prompts which consisted of either open-ended questions or requests for the candidate to produce an extended piece of talk. On the other hand, although Ian was obviously

attempting to elicit extended responses, he was less explicit in his questioning than Pam. Thus, Esther often misinterpreted the pragmatic force of Ian's prompts, and her typically brief responses were often followed by a long pause while Ian waited for a response or formulated his next question. These pauses gave the discourse a sense of disfluency. Consequently, in general, Esther appeared to be more proficient when interviewed by Pam than when she was interviewed by Ian. This study revealed that interviewer training has been overlooked. Additionally, Brown stated that the test administrators should ensure that interviewers' styles are not so diverse. Finally, Brown concluded that differences in interviewer behavior are related to the construct. Therefore, in order to increase the reliability of test scores and the validity of test use, clear and unambiguous theoretical definitions of the abilities should be provided, and the conditions or operations which will be followed in eliciting and observing performance should be specified carefully.

Other researchers who examined the reliability of tests from the perspective of scorer reliability are Manola and Wolfe (2000). They explored the reliability of the essay writing section of the TOEFL, investigating whether the essay medium could affect the reliability of the scores and the accuracy of the inferences drawn from these scores. The aim of this study was to find out to what degree raters' judgments were affected by computer based and hand writing essay mediums for the Test of English as a Foreign Language (TOEFL). The participants involved in this study were 152,951 TOEFL examinees: 51.5% who chose to write their TOEFL essays on a word processor and 48.5% who chose handwriting as their essay medium. The papers of these examinees were scored by two independent groups of trained judges. A number of analyses were

made to define how much the ratings were affected by construct-irrelevant variance. Analyses revealed that the raters found it slightly easier to agree on word processed essays than the handwritten ones, and the scores of the hand written essays were less reliable than the scores of the word processed ones. Thus, Manola and Wolfe suggested that the inferences drawn from hand written essays had low validity. Therefore, they concluded that it was not fair to make decisions about the examinees by using the scores obtained from them.

In both Nakamura's (2006) and Cardoso's (1998) studies reliability was established by examining the internal consistency of test scores. However, in the present study it is aimed to establish reliability by learning the students' perceptions of reliability in terms of their performance in the exams and instructors' perceptions of the scorers' reliability. In this respect, Manola and Wolfe's (2000) and Brown's (2003) studies are similar to the study described in this thesis, since they also investigated scorers' reliability.

Additionally, some researchers have explored the predictive validity of tests. For instance, Yeğin (2003) examined the predictive validity of the Başkent University English Proficiency Exam (BUEPE) by using Item response theory (IRT) -based ability estimates. The study made use of the BUEPE September 2000 data which included the responses of 699 students. Predictive validity was established by using the DEC (Departmental English courses) passing grades of a total number of 371 students. It was found that the correlations between BUEPE total scores and DEC first semester and second semester passing grades were moderately high, with a slightly lower correlation for DEC second semester passing grades. This is consistent with Pack's (1972, cited in

Marvin & Simner, 1999) finding. According to Pack, TOEFL scores are related to the grade obtained in the first English course taken but not related to grades obtained in subsequent English courses. Finally, it is indicated in Yeğin's study that in general the Proficiency Exam in question had a moderately high predictive validity.

Dooey (1999) also explored the predictive validity of tests. She investigated the predictive validity of the IELTS (International English Language Testing System) test as an indicator of future academic success. The subjects were 65 foreign students and 23 native students. Analyses were made by using IELTS scores and semester weighted averages (SWAs). For example, scatter plots were generated to present the possible relationships between the year averages and IELTS overall scores, year averages and IELTS subtest scores, SWAs for Semester 1 and SWAs for Semester 2 and SWAs and IELTS subtest scores. Lastly, correlations between the SWAs and IELTS scores (by subtest and overall) were computed. The analyses revealed that overseas students who did not fully meet the admission criteria in terms of their English level were still successful academically, and 15 out of 23 native English speakers who did not experience any difficulty with English became unsuccessful academically. Thus, Dooey argued that high IELTS scores did not ensure future academic success, and factors other than linguistic ability must have affected the native students' performance. Dooey's study is similar to the study described in this thesis, in that two SWAs (SWAs for Semester 1 and SWAs for Semester 2) were correlated with each other. However, unlike Dooey's study, the correlations between year averages and final examination scores were also computed.

Ösken (1999) also looked at the predictive validity of tests. She examined the predictive validity of midterm achievement tests administered at Hacettepe University, Department of Basic English (DBE). The achievement test scores (obtained from six midterms and one end-of-course assessment) of the students who enrolled in the course in the 1997-1998 academic year were compared with one another to define the predictive validity. The study indicated that the mid-term achievement tests had only a moderate amount of predictive validity, and the author speculated that this was because of the differences between their forms and contents.

The previous two studies described illustrate different ways of looking at predictive validity. Yeğin's study differs from both these two studies and the present study since it employed Item response theory (IRT) -based ability estimates to examine predictive validity. Furthermore, both Yeğin and Dooey investigated the predictive validity of proficiency tests as an indicator of future academic success (after the program). However, Ösken's study and the present study examined the predictive validity of midterms as an indicator of success in other midterms and the end-of-course assessment (within the program).

Lastly, some researchers have investigated the face validity and content validity of tests. For example, Ösken (1999), in her previously mentioned study, examined the face validity and content validity of the end-of-course assessment administered at Hacettepe University, Department of Basic English (DBE) in the 1997-1998 academic year. Questionnaires were distributed to the instructors to investigate their perceptions of whether the end-of-course assessment represented the contents of the course books. Additionally, the number of test items was compared with the frequencies of course

objectives to define the content validity.[3] The data gathered from the questionnaires suggest that the end-of-course assessment represented the course contents in the eye of the instructors. However, the results of the analysis of course objectives and test items of the end-of-course assessment indicate that the test items were not chosen according to the frequencies of course objectives. Therefore, Ösken indicates that the end-of-course assessment was a limited representative of the course contents when the proportions of language items in the course books were compared with the test items in the end-of-course assessment. Ösken states that the mismatch between face validity and content validity may have been due to the lack of test objectives. Furthermore, according to Ösken, the number of course objectives was high in terms of grammar, and it was impossible to test all bits and pieces of grammar. Therefore, the testers might have chosen the main structures to test while ignoring the others.

The next researcher who investigated both face validity and content validity of tests is Serpil. Serpil (2000) looked at the face validity and content validity of midterm achievement tests administered at Anadolu University School of Foreign Languages. Instructors were asked to fill out questionnaires to discover their perceptions of the tests' representativeness of the intermediate course material content and teaching objectives. In addition to this, the instructors were interviewed to find out the teaching objectives. Then, first a comparison between the content of the tests and the course materials was made. Next, a comparison between the content of the tests and teaching objectives was made. The findings indicate that the instructors in general thought that the midterm tests'

[3] The closer examination of test items and course contents, in other words, the higher specificity of the examination is what makes this stage of the study about content validity rather than face validity.

representativeness of the courses' content was moderate to high. However, it was found that the degree of the tests' representativeness of the course material was low, especially when the exercise types were considered. Furthermore, a low correlation between the content of the tests and the teaching objectives was observed. In other words, in this study, face validity did not appear to predict an objective measure: content validity. Serpil speculated that the lack of clearly defined testing criteria and course objectives was the main factor causing such a conflict among the results.

Another researcher who looked at the face validity and content validity of tests is Nakamura. Nakamura (2006) examined the face validity and content validity of a pilot English placement test in his previously mentioned study. In this study, face validity was examined through an informal questionnaire and discussions with 809 freshman university students. Most of the students agreed that the test in question had face validity. Content validity was established through a discussion about the test items. The instructors discussed how well the test items reflect the content of the text book they were using and the content of their teaching. All the English instructors involved in the test construction process agreed that the pilot placement test possessed content validity. Nakamura concluded that both of the presuppositions were partially supported. In other words, face validity and content validity were explained relatively convincingly.

In Nakamura's (2006) study content validity was established through a discussion among the instructors about how well the test items reflect the content of the text book they were using and the content of their teaching. However, in Serpil's (2000) and Ösken's (1999) studies a more systematic way was followed in that comparisons between the contents of the tests and the course materials and comparisons between the

contents of the tests and teaching objectives were made by calculating the percentages of both the taught and tested items. Content validity is not aimed to be established in the study presented in this thesis. Where face validity is concerned, Nakamura (2006) investigated only students' perceptions of the content of the tests, and Serpil (2000) and Ösken (1999) investigated only instructors' perceptions of the content of the tests to establish face validity in their studies. However, in the study described in this thesis both students' and instructors' perceptions of the content of the tests are aimed to be examined and compared with each other to establish the degree of face validity. What is more, in Nakamura's (2006) study face validity has been confirmed by content validity. On the other hand, in both Ösken's (1999) and Serpil's (2000) study there is a conflict between face validity and content validity. However, none of these researchers have made comments about how well face validity reflects other relatively more objective measures such as reliability.

The researcher has reviewed issues of test reliability and validity in the literature, as well as the literature about the ways that assessments are evaluated for reliability and validity in this chapter. It has been seen that no research studies have been conducted before on the topic of how well face validity reflects a selection of relatively more objective measures such as reliability and predictive validity. The study described in the next chapter aims to conduct such a comparison.

Conclusion

This chapter was a review of the literature covering the uses, kinds and good qualities of language tests, and previous research studies conducted on reliability and validity of language tests. The next chapter will describe the methodology of the study in terms of its setting, participants, instruments and data collection procedures.

CHAPTER III: METHODOLOGY

Introduction

This study is an exploratory study which focuses on the relationship between face validity and relatively more objective measures of tests, such as reliability and predictive validity. The study also attempts to investigate the instructors' and students' perceptions of the validity and reliability of tests administered at Zonguldak Karaelmas University Preparatory School.

This study addresses the following research questions:

1. To what extent do the achievement tests possess face validity?
    - To what extent do the achievement tests represent the course content in the eyes of the instructors?
    - To what extent do the achievement tests represent the course content in the eyes of the students?
    - Is there a difference between the two groups' perceptions of the achievement tests' representativeness of the course content?
2. To what extent do the achievement tests possess reliability?
    - To what extent does the current testing system permit scorer reliability?
    - To what extent do the structure of the tests and the testing conditions permit students to accurately demonstrate their language knowledge and skills?

3. To what extent do the achievement tests possess predictive validity?

- How well do the achievement tests conducted in the first term predict success in the second term?
- How well do the achievement tests conducted throughout the year predict success in the end-of-course assessment?

4. How closely does the face validity of the achievements tests reflect the reliability and predictive validity of these tests?

This chapter of the thesis will cover the setting, participants, instruments, data collection procedures and data analysis.

Setting

The study was conducted at Zonguldak Karaelmas University Foreign Languages Compulsory Preparatory School. The aim of this school is to help undergraduate students to acquire English language knowledge and skills, since the students will have to follow some of the courses in their own departments in English. Students with different educational backgrounds attend this school. In other words, both students who have already attended preparatory schools in other institutions and students who have not received any English language education attend this school. For that reason, students first take a proficiency exam at the beginning of each academic year. If they pass this exam, they are exempted from attending the preparatory class. On the other hand, if they are unsuccessful, they must take the placement test which is administered to place the students at the appropriate levels of the program. There are three levels at the

Preparatory School: B level (Lower intermediate), C+ (Beginner) and C- (True Beginner).

The program adopted by this institution is a skill-based one. English instructors at the Preparatory School teach the Quartet set (Q Group, 2005), which is composed of eleven books, in the main courses. This set is accompanied by the Quartet computer software (formed of eleven levels) which is parallel with the main course book. A grammar book, Milestones of English Grammar-*Perfecting and practicing English structure* (Küçük et al., 2006), is also followed in the main courses. Additionally, instructors use two books, Password one and Password two (Butler, 2003), in reading courses, and they make use of a book called Writing to Communicate (Boardman & Frydenberg, 2002) in writing courses. The program has not yet adopted a regular speaking book. Therefore, instructors use activities such as role plays and games prepared by the speaking office in speaking courses. Lastly, videos with levels ranging from elementary to upper intermediate are employed in video courses. These videos are accompanied by video workbooks, Challenge and Real Lives (Simpson, 2004), Cutting Edge (Cunningham & Moor, 2003) and An Ocean Apart (McHugh, 2003), which contain comprehension questions.

The students are assessed on their performance on quizzes and midterms administered throughout the year, their participation in speaking lessons, writing assignments, and an end-of-course assessment. Table 1 below shows the weighting of the criteria for assessing students at Zonguldak Karaelmas University Preparatory School:

Table 1 - Weighting of the Students' Assessment Criteria

Assessment criteria	Percentage
First Midterm:	8%
Second Midterm:	8%
Third Midterm:	8%
Fourth Midterm:	8%
20 quizzes + 4 mock exams:	8%
Participation in speaking lessons:	5%
Writing assignments:	5%
End-of-course assessment:	50%
Total:	100%

The contents of speaking, writing, video and laboratory lessons are not incorporated into the exams. Speaking course instructors give their students a grade ranging from one to five at the end of the academic year, considering their presentation performance and participation in the lesson. Additionally, writing course instructors give a mark, again ranging from one to five, bearing in mind the assignments the students have completed throughout the year.

Participants

Two different groups of participants were included in this study. Twenty nine C-level instructors who were working at Zonguldak Karaelmas University Foreign Languages Compulsory Preparatory School in the 2005-2006 academic year formed the first group, and 52 C- level undergraduate students who were enrolled in the same institution during the same period formed the second group. In addition, the scores of 477 C- level undergraduate students enrolled in the school in the 2005-2006 academic year were examined in order to answer the third and fourth research questions.

Preparatory class instructors

There were 29 instructors at Zonguldak Karaelmas University Preparatory School, and all of them taught C-level classes in the 2005-2006 academic year. Therefore, the researcher aimed to involve all 29 instructors, some of whom were working at the testing office during the time period of this study. Since instructors work in rotation in the testing office in this institution, almost all of the instructors had already worked in the testing office. Consequently, it was not possible to exclude the testing office members from the study.

Question 1 in section IV of the instructors' questionnaire collected data about the educational backgrounds of the instructors. Table 2 below presents the data obtained.

Table 2 - Educational Backgrounds of the Instructors

Degree	Frequency	Percentage
B.A. degree	21	72%
M.A. degree	8	28%
Total	29	100%

Question 2 in section IV of the instructors' questionnaire collected data about the teaching experience of the instructors which can be seen in Table 3 below.

Table 3 - Teaching Experience of the Instructors

Years of Teaching Experience	Frequency	Percentage
1 to 4 years	12	41.3%
5 to 8 years	12	41.3%
9 to 12 years	3	10.3%
More than 13 years	2	6.8%
Total	29	99.7%

* Percentages do not equal 100% due to rounding.

Question 3 in section IV of the instructors' questionnaire collected data about the testing experience of the instructors, which can be seen in Table 4 below.

Table 4 - Testing Experience of the Instructors

Years of Testing Experience	Frequency	Percentage
No experience	4	13.7%
Less than one year	4	13.7%
1 to 3 years	13	44.8%
More than 3 years	8	27.5%
Total	29	99.7%

* Percentages do not equal 100% due to rounding.

Question 4 in section IV of the instructors' questionnaire gathered information about whether the instructors have taken any courses on testing. The relevant data can be found in Table 5 below.

Table 5 - Testing Courses Taken by the Instructors

Taken courses on testing	Frequency	Percentage
Yes	23	79%
No	6	21%
Total	29	100%

Former preparatory class students

C- Level (True Beginner) was chosen as the most appropriate level for this study for three reasons. First, this group has the highest number of students. Second, the success of C+ (Beginner) and B (Lower intermediate), students may be due to the basic English knowledge they acquired in other institutions (e.g., high school) in the past. However, C- level students have most probably learnt everything about English in the institution in question. Lastly, the end-of- course assessment is always constructed according to C- level students' proficiency level.

There were 645 students at the Preparatory School in the 2005-2006 academic year. If B level students, C+ level students and dropouts are excluded, 477 C- level

students are left. Preparatory school education is compulsory for 13 departments, and these departments are given in Table 6.

Table 6 - The Departments in Which the Students Are Enrolled

Faculties	Departments				
Economics and administrative sciences	Business administration	Economics			
Arts and science	Mathematics	Biology	Chemistry	Physics	
Engineering	Geodesy and photogrammetry	Electronical and electrics	Mechanical	Mining	Civil
Medicine	Medicine				
Fine Arts and design	Architecture				

52 C- level students, or more than 10% of the total number of students at this level, participated in this study. The researcher obtained from the students' affairs office last years' students list in which the students were categorized according to the departments they were enrolled in. Then, the first or last four students from each department were selected from this list. Sometimes it was impossible to involve the first four students in the study, since they had been enrolled in B or C+ levels rather than C- level, when they were in preparatory class. Therefore, the researcher sometimes had to choose the last four students from the list. Additionally, two students were chosen as substitute students by employing the same method. The reason for this is that the selected students might not be present in the class at the time of the study or might be reluctant to participate in the study.

The above mentioned selection technique was not employed with the architecture department, because the architecture department is not in the main campus. It is situated in Safranbolu (a far district of Zonguldak) which was difficult to visit at the time of the

study. Consequently, the researcher had to use the 'snowball sampling' method which involves detecting a few people who fulfill the criterion of a particular study and then asking these participants to find further members of the population (Dörnyei, 2003, p. 72). In other words, the researcher first identified one of her former students who was enrolled in the architecture department and suitable for participating in the study. Then, she contacted him and asked him to help her in finding three more participants who were also suitable for the study.

Question 1 in section IV of the students' questionnaire collected data about the educational backgrounds of the students. Table 7 below presents the data obtained.

Table 7 - Educational Backgrounds of the Students

Attendance at preparatory class in another institution before	Frequency	Percentage
Yes	14	27%
No	38	73%
Total	52	100%

Preparatory class students who fail do not repeat the preparatory class. They go to their departments, and they try to pass the proficiency test which is administered at the beginning of each academic year. They must pass the proficiency test before they graduate. Otherwise, they will not receive their graduate certificate. Therefore, first class students who will complete the questionnaire might have passed or failed the preparatory class, and Question 2 in section IV of the students' questionnaire collected data about whether students have passed or failed the preparatory class. Table 8 below presents the data obtained.

Table 8 - Success of the Students in Preparatory Class

Success	Frequency	Percentage
Passed	49	94%
Failed	3	6%
Total	52	100%

Instruments

The instruments employed in this study were two questionnaires and C- level students' test scores.

Questionnaires

Dörnyei (2003, p. 1) indicates that "questionnaires are uniquely capable of gathering a large amount of information quickly in a form that is readily processable." In this study the researcher aimed to collect data from a large population in a limited amount of time. Consequently, two questionnaires were selected as the main research tools.

Since no similar study had been done before, the questionnaires were created by the researcher. Some sources of invalidity and unreliability suggested by the literature provide a basis for evaluating achievement tests. The researcher turned these sources into questionnaire items. Some of the questionnaire items were prepared as negatively oriented questions so that the students would not think that the researcher wanted them to write only positive things about the testing practices in Z.K.U. Prep School. In other words, by this way, the students would not think that the researcher was biased.

The teachers' questionnaire was in English, and the students' questionnaire was in Turkish. In this way, it was felt that the participants would comprehend and respond to the questions better.

The researcher attached a letter to the beginning of both of the questionnaires to explain the aim and importance of the study and get the consent of the participants. (See Appendix A for the letter which was prepared to get the consent of the instructors, Appendix B for the Turkish version and Appendix C for the English version of the letter which was prepared to get the consent of the students.)

Instructors' questionnaire

A questionnaire formed of four sections was distributed to 29 English instructors (see Appendix D). The first section concerned the instructors' perceptions of the face validity of the achievement tests. This section was intended to contribute to the answer to the first research question. The next one was about the instructors' perceptions of whether the current testing system permits scorer reliability or not. This section was intended to help answer research question two. The way the researcher measured reliability in this study was not strictly "objective" because she relied on instructors' perceptions of reliability. However, because specific questions about test and testing situation characteristics were asked, rather than general questions about reliability, it was felt that this was a relatively objective way of examining reliability. These two sections involved Likert scale items in which the participants were asked to circle the alternative which reflects their opinions best. The alternatives were 'Strongly Disagree' (SD), 'Disagree' (D), 'Uncertain' (U), 'Agree' (A) and 'Strongly Agree' (SA). The third section was in the form of open-ended questions related to the issues mentioned in the

previous sections of the questionnaire. Lastly, the background section of the questionnaire was placed at the end of the questionnaire. The reason for this is that, as Oppenheim (1992) indicates, when the participants learn the aim of a study and decide to contribute, they expect interesting questions. Therefore, Oppenheim claims that in order not to distract the attention of participants, questions which require personal information should be put at the end of the questionnaires. Table 9 shows the distribution of the questions in the instructors' questionnaires.

Table 9 - Distribution of the Questions in the Instructors' Questionnaire

	Section I	Section II	Section III	Section IV
Focus	instructors' perceptions of the face validity of the achievement tests	instructors' perceptions of whether the current testing system permits scorer reliability or not.	open-ended questions related to the issues mentioned in the second and third sections	background information
Number of questions	12	13	4	4

Students' questionnaire

The students' questionnaire consisted of four sections (see Appendix E for the Turkish version and Appendix F for the English version). The first two sections were concerned with students' perceptions of the face validity of achievement tests and the reliability of tests in terms of their performance, respectively. The section about the perceptions of the face validity of achievement tests was in parallel with section one of the instructors' questionnaire. For the same reason mentioned above, the questions in the second section of the students' questionnaire, which were about specific test and testing

situation characteristics, were felt to be reasonably objective ways of examining reliability. Both the first and the second sections included Likert scale items in which the participants were asked to circle the alternative [Strongly Disagree' (SD), 'Disagree' (D), 'Uncertain' (U), 'Agree' (A) or 'Strongly Agree' (SA)] which reflected their opinions best, as in the instructors' questionnaire. The third section was in the form of open-ended questions related to the issues mentioned in the first and second sections. The last section was asking for information about students' background. Table 10 indicates the distribution of the items in the students' questionnaire.

Table 10 - Distribution of the Questions in the Students' Questionnaire

	Section I	Section II	Section III	Section IV
Focus	students' perceptions of the face validity of the achievement tests	students' perceptions of the reliability of the tests in terms of their performance	open-ended questions related to the issues mentioned in the second and third sections.	background information
Number of Questions	12	22	2	2

Test scores

Apart from the questionnaires, all C- level students' midterm and end-of-course assessment scores from the 2005-2006 academic year were examined. If the researcher had involved only some students' scores in the study, the selected samples may not have reflected the exact situation. In other words, the larger the sample, the more reliable the results. Therefore the scores of all C- level students were included in this study.

Data Collection Procedures

First of all, the instructors' questionnaire was piloted with three instructors on 20 January 2007. The number (three instructors) represents 10% of the actual number of the participant instructors. The aim of the piloting was to discover the problematic items in the questionnaire and to see whether it was hard to understand the questions. The researcher made the necessary changes to the questionnaire on 22 January.

On 23 January, the instructors' questionnaire was conducted by getting permission from the director of the Foreign Languages Preparatory School. The researcher handed in the questionnaires to the officer who was responsible for the circulation of administrative documents. The officer handed out the questionnaires to the instructors in return for signature, and on 29 January the teachers' questionnaires were collected by the same officer.

On 2 February 2007, the researcher got permission from the administrators to use last year's test scores (the 2005-2006 academic year) and to get last years' student lists. Then, the researcher obtained from the students' affairs office 477 C- level students' midterm and end-of-course assessment scores, and last years' student lists in which the students were categorized according to the departments in which they were enrolled. The researcher selected four students from each department by examining the lists.

On the same day, the researcher asked three instructors to proofread the Turkish translation of the students' questionnaire. After the necessary changes were made, the questionnaire was piloted with five students on 9 February. The number (five students) represents 10% of the actual number of the participant students. The researcher made some minor changes on the questionnaire after the piloting. The students' questionnaire

was administered on 23-26-27 February. This procedure took three days, since the researcher had to visit several faculties to collect data. The researcher used a one- to-one administration method. The reason for this is that since the students were from various departments, it was impossible to assemble them together. Furthermore, as Dörnyei (2003) states, "one- to-one administration is a much more personal form of administration than mail surveys, and therefore the chances for the questionnaires to be returned are significantly better" (p. 81). The researcher learned the weekly schedule of English language instructors who teach at the departments in question from the students' affairs office on 22 February. Learning their weekly schedules helped the researcher to define the dates of questionnaire administration for each department. On the defined dates, the researcher explained the significance of the study to the English instructors, got permission from them and took the students to another empty classroom at the beginning of the lesson in order not to interrupt the class. Furthermore, by this way, the students could easily contact the researcher, if questions arose. After they answered all the questions, the researcher collected the questionnaires.

Then, since the architecture department is situated in Safranbolu (a far district of Zonguldak), on 28 February the researcher contacted one of her former students who was enrolled in the architecture department and suitable for participating in the study and asked him to help her in finding three more participants who were suitable for the study. The researcher was able to obtain the e-mail addresses of these four architecture students on 2 March and sent the questionnaire through e-mail to them on the same day. The completed questionnaires were returned on 5 March.

Data Analysis

First, the overall means of each section in both the instructors' and students' questionnaire were calculated. Furthermore, overall means of the two subsections in section II of the students' questionnaire were computed. Then, independent samples t-tests were used to compare the means. All these analyses were conducted with the help of a statistician.

Next, the researcher analyzed the qualitative data gathered from open-ended questions in Section III of the two questionnaires. Participants' responses were categorized according to key words and common themes and entered in tables.

For the fourth sections of the two questionnaires, which ask for background information of the participants, the frequencies and percentages were calculated. The results were presented in tables.

Then, the correlations among C- level students' first term averages, second term averages, cumulative averages (consisting of averages of four midterms conducted throughout the year) and the end-of-course assessment scores were computed by means of Pearson Product Moment Correlation Coefficient, again with the help of a statistician, to determine the predictive validity. Lastly, the correlations between face validity and reliability and face validity and predictive validity were evaluated to answer the fourth research question.

Conclusion

In this chapter, the methodology used to carry out the study was described in terms of its setting, participants, instruments and data collection procedures. In chapter four, data analysis and the specific outcomes will be discussed in detail.

CHAPTER IV: DATA ANALYSIS

Introduction

This study investigates how well face validity reflects relatively more objective measures: reliability and predictive validity. The study also aims to examine the predictive validity, face validity and reliability of tests administered at Zonguldak Karaelmas University Preparatory School.

This study addresses the following questions:

1. To what extent do the achievement tests possess face validity?
    - To what extent do the achievement tests represent the course content in the eyes of the instructors?
    - To what extent do the achievement tests represent the course content in the eyes of the students?
    - Is there a difference between the two groups' perceptions of the achievement tests' representativeness of the course content?
2. To what extent do the achievement tests possess reliability?[4]
    - To what extent does the current testing system permit scorer reliability?
    - To what extent do the structure of the tests and the testing conditions permit students to accurately demonstrate their language knowledge and skills?
3. To what extent do the achievement tests possess predictive validity?

[4] In this thesis, scorers' reliability was determined by asking specific questions to the scorers about scoring practices, and reliability in terms of the test takers' performance was determined by asking specific questions to the students about the structure of the tests and the testing conditions.

- How well do the achievement tests conducted in the first term predict success in the second term?
- How well do the achievement tests conducted throughout the year predict success in the end-of-course assessment?

4. How closely does the face validity of the achievements tests reflect the reliability and predictive validity of these tests?

Data Analysis Procedures

In this study the researcher used both quantitative and qualitative data analysis procedures. Three sets of data were used in the data analysis procedures. The first set of data, which was gathered from Likert scale questions in both instructors' and students' questionnaires, was analyzed quantitatively. The second set of data, which was collected through six open-ended questions in two questionnaires, was analyzed qualitatively. The third set of data, which was composed of midterm and end-of-course assessment scores of C- (beginner) level students, was analyzed quantitatively.

The first step of the data analysis procedure was the analysis of the Likert scale question responses in two questionnaires. The data obtained from the 25 Likert scale questions in section I and section II of the instructors' questionnaire and the 34 Likert scale questions in section I and section II of the students' questionnaire were entered into SPSS. Means and standard deviations have been calculated for each question in the two questionnaires. Additionally, independent samples t-tests were used to compare the means.

The second step of the procedure was to analyze the data gathered from the six open-ended questions in section III of the two questionnaires. Both the instructors' and the students' responses were examined with the aim of finding key words and common themes. Then, the participants' responses were categorized according to these key words and common themes and entered in tables.

The last step of the procedure was to analyze the midterm and end-of-course assessment scores of the students. In order to analyze the relationship between tests, Pearson Product Moment Correlation Coefficient was used. First, students' averages in the first and second term were calculated. Next, the correlation between first term averages and second term averages was determined to discover the degree of the first term achievement tests' predictive validity. Thirdly, the correlation between first term averages and the end-of-course assessment scores was determined to again examine the predictive validity of the first term achievement tests. Then, the correlation between second term averages and the end-of-course assessment scores was determined to check the predictive validity of the second term achievement tests. Lastly, students' cumulative averages (including the averages of scores obtained from four midterms) just before they took the end-of-course assessment were calculated, and the correlation between students' cumulative averages and the end-of-course assessment scores was determined to examine the predictive validity of the tests conducted throughout the year.

The results of the analysis will be presented in three main parts. The first part discusses the first sections of both instructors' and students' questionnaires which investigate the instructors' and students' perceptions of the face validity of the achievement tests. This part will be used to answer research question one. In the second

part, an analysis of the questions in section II of the two questionnaires, which are concerned with the extent to which the achievement tests possess reliability, is presented. This part addresses research question two. The data about face validity and reliability will be supported by the participants' responses to the open-ended questions. The data gathered from the open-ended questions will be presented in Appendix G and Appendix H. Finally, the third part discusses the quantitative data gathered to determine the extent to which the achievement tests possess predictive validity, and it will be used to answer research question three.

The Extent to Which the Achievement Tests Possess Face Validity

The questions in section I of the instructors' and students' questionnaires sought to answer research question one, which investigates the instructors' and students' perceptions of the face validity of the achievement tests administered in Zonguldak Karaelmas University Prep School. In this part data are presented in the form of three subsections: instructors' perceptions of the face validity of the achievement tests, students' perceptions of the face validity of the achievement tests, and any difference between the two groups' perceptions of the face validity of the achievement tests.

Instructors' perceptions of the face validity of the achievement tests

All twelve questions in the first section of the instructors' questionnaire are concerned with the instructors' perceptions of the face validity of the achievement tests. Therefore, in order to determine the instructors' perceptions of the face validity of the achievement tests, a mean score was calculated for each question, and the means of all 12 questions were then averaged together, producing a "mean of the means." Table 11 below shows the mean of the means of these 12 questions.

Table 11 - Mean of the Means, Instructors' Perceptions of Face Validity

TOTAL (SECTION-1)	Group	N	Mean	Std. Deviation
	Instructors	29	3.431	.555

In interpreting the means of the Likert scale items, the following scale was used.

Figure 1 - Rating Scale for Interpreting Likert-Scale Responses

Mean	Degree	Opinion
4.5-5	Very high	Strongly agree
3.5-4.4	High	Agree
2.5-3.4	Moderate	Undecided
1.5-2.4	Low	Disagree
1.1-1.4	Very low	Strongly Disagree

As can be seen in Table 11, the mean is 3.431, which is closer to moderate than high. Consequently, it can be assumed that the achievement tests represent the course content to a moderate degree in the eyes of the instructors. Additionally, a further analysis was made to define the validity of this section of the instructors' questionnaire. The mean which represents the means of the first 11 questions was compared with the mean of the 12th question, which was a question about the teachers' perceptions of the overall representation of the course content in the exams, using an independent samples t-test. Table 12 shows the findings of this comparison.

Table 12 - Validity Analysis, Instructors' Perceptions of Face Validity

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-11 (total)	3.411	0.936	1.980	0.058
Q12	3.655	0.542		

It has been observed that there is a difference which approaches significance between the mean which represents the means of the first 11 questions and the mean of question 12, and the means do not fall into the same category. The mean which represents the means of the first 11 questions falls into the category of moderate, and the

other falls into the category of high. Therefore, this section of the instructors' questionnaire should be examined more closely. With the aim of gaining greater insights about the perceptions of the instructors, the eleven related questions were compared with the 12th question, and means, standard deviations, t-values and *p*-values have been calculated. The results of the independent samples t-test are given in Table 13.

Table 13 - Detailed Validity Analysis, Instructors' Perceptions of Face Validity

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1The content of the main course book 'Quartet' was represented in the exams sufficiently.	4.310	0.541	3.768	0.001*
Q2 The content of the grammar book 'Milestones' was represented in the exams sufficiently.	4.000	0.755	1.907	0.067
Q3 The content of the writing courses was represented in the exams sufficiently.	3.862	0.953	1.063	0.267
Q4 The content of the reading courses was represented in the exams sufficiently.	3.310	1.105	-1.625	0.115
Q5 The content of the speaking courses was represented in the exams sufficiently.	2.276	1.131	-7.320	0.000*
Q6 The content of the video courses was represented in the exams sufficiently.	2.138	0.789	-8.968	0.000*
Q7 Grammar taught in the courses was represented in the exams sufficiently.	4.483	0.634	4.446	0.000*
Q8 The vocabulary taught in the courses was represented in the exams sufficiently.	4.069	0.842	2.188	0.037*
Q9 The listening practices focused on in the courses were represented in the exams sufficiently.	2.517	1.271	-5.607	0.000*
Q10 The content of the laboratory courses was represented in the exams sufficiently.	2.586	1.150	-5.574	0.000*
Q11 The exercises made in the courses were represented in the exams sufficiently.	3.965	0.680	2.073	0.048*
Q12 In general, the contents of the courses were represented in the exams sufficiently.	**3.655**	**0.936**		

As can be seen in Table 13, while the instructors disagree with Q5 and Q6, they agree with Q1, Q2, Q3, Q7, Q8, Q11 and Q12. It can be further indicated that the instructors are undecided about Q4, Q9 and Q10. It has also been observed that there is a significant difference between Q12 and Q1, Q5, Q6, Q7, Q8, Q9, Q10 and Q11.

Q5 and Q6 are about the aspects of the curriculum that are known by the researcher to be not included in the tests, and the participants were expected to answer these questions negatively. Therefore, it made sense to exclude them from the validity

analysis. It has been found that when Q5 and Q6 are excluded from the analysis, there is no significant difference between the mean of question 12 and the mean which represents the means of nine questions, and they fall into the same category, high. Table 14 shows the findings of the independent samples t-test used to compare the means.

Table 14 - Validity Analysis, Instructors' Perceptions of Face Validity, 2 Questions Omitted

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-4, Q7-11 (total)	3.678	0.552	0.177	0.861
Q12	3.655	0.936		

Therefore, it can be concluded that the instructors perceive the achievement tests as possessing a high degree of face validity, but when asked specifically about speaking and video courses, they appear to believe that those courses are not well represented in the exams. Additionally, it seems that instructors are undecided about whether the contents of the reading courses, laboratory courses and the listening practices focused on in the courses are well represented in the exams. Three participants' (P3, P4 and P13) responses to the open-ended questions support these findings.

> (Participant 3) In our institution, like most of other institutions, speaking cannot be evaluated properly. **Since this skill exists in the curriculum we should also test it.**
>
> (Participant 4) **I believe that with listening and speaking exams we will have more reliable exams.**
>
> (Participant 13) The exams should include all skills. **They should not include only grammar, vocabulary and writing but also listening and speaking.**

The above mentioned quotes indicate that the instructors are obviously aware that there are some parts of the curriculum that are not tested in the exams.

Students' perceptions of the face validity of the achievement tests

All twelve questions in the first section of the students' questionnaire are concerned with the students' perceptions of the face validity of the achievement tests. Therefore, in order to determine the students' perceptions of the face validity of the achievement tests, a mean score was calculated for each question, and the means of all 12 questions were then averaged together, producing a "mean of the means." Table 15 below shows the mean of the means of these 12 questions.

Table 15 - Mean of the Means, Students' Perceptions of Face Validity

TOTAL (SECTION-1)	Group	N	Mean	Std. Deviation
	Students	52	3.320	.515

As can be seen in Table 15, the mean is 3.320, which is closer to moderate than high. Consequently, it can be assumed that the achievement tests represent the course content to a moderate degree in the eyes of the students. Additionally, a further analysis was made to define the validity of this section of the students' questionnaire. The mean which represents the means of the first 11 questions was compared with the mean of the 12th question, which was a question about the students' perceptions of the overall representation of the course content in the exams, using an independent samples t-test. Table 16 shows the findings of this comparison.

Table 16 - Validity Analysis, Students' Perceptions of Face Validity

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-11 (total)	3.247	0.338	7.709	0.000*
Q12	4.135	0.793		

There is a significant difference between the mean which represents the means of the first 11 questions and the mean of question 12, and the means do not fall into the same category. The mean which represents the means of the first 11 questions falls into the category of moderate, and the other falls into the category of high. Therefore, this section of the students' questionnaire should be examined more closely. With the aim of gaining greater insights about the perceptions of the students, the eleven related questions were compared with the 12th question, and means, standard deviations, t-values and *p*-values were calculated. The results of the independent samples t-test are given in Table 17.

Table 17 - Detailed Validity Analysis, Students' Perceptions of Face Validity

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1The content of the main course book 'Quartet' was represented in the exams sufficiently.	4.135	0.864	0.000	1.000
Q2 The content of the grammar book 'Milestones' was represented in the exams sufficiently.	4.039	0.989	-0.598	0.552
Q3 The content of the writing courses was represented in the exams sufficiently.	3.557	1.092	-3.307	0.002*
Q4 The content of the reading courses was represented in the exams sufficiently.	3.019	1.075	-6.533	0.000*
Q5 The content of the speaking courses was represented in the exams sufficiently.	2.000	0.990	-13.954	0.000*
Q6 The content of the video courses was represented in the exams sufficiently.	1.942	0.850	-14.566	0.000*
Q7 Grammar taught in the courses was represented in the exams sufficiently.	4.654	0.480	3.987	0.000*
Q8 The vocabulary taught in the courses was represented in the exams sufficiently.	4.365	0.658	1.571	0.122
Q9 The listening practices focused on in the courses were represented in the exams sufficiently.	2.077	0.968	-12.090	0.000*
Q10 The content of the laboratory courses was represented in the exams sufficiently.	2.328	1.043	-11.629	0.000*
Q11 The exercises made in the courses were represented in the exams sufficiently.	3.596	1.272	-3.009	0.004*
Q12 In general, the contents of the courses were represented in the exams sufficiently.	**4.135**	**0.793**	-	

As can be seen in Table 17, while the students disagree with Q5, Q6, Q9 and Q10, they agree with Q1, Q2, Q3, Q8, Q11 and Q12. Furthermore, the table shows that

the students strongly agree with Q7, and they are undecided about Q4. Lastly, it has been observed that there is a significant difference between Q12 and Q3, Q4, Q5, Q6, Q7, Q9, Q10, Q11.

As mentioned before in the previous section, Q5 and Q6 are about the aspects of the curriculum that are known by the researcher to be not included in the tests, and the participants were expected to answer these questions negatively. Therefore, it made sense to exclude them from the analysis. It has been found that when Q5 and Q6 are excluded from the analysis, there is still a significant difference between the mean of question 12 and the mean which represents the means of nine questions; however they fall into the same category, high. Therefore, it can be assumed that this section of the students' questionnaire is valid. Table 18 shows the findings of the independent samples t-test.

Table 18 - Validity Analysis, Students' Perceptions of Face Validity, 2 Questions Omitted

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-4, Q7-11 (total)	3.529	0.548	-5.165	0.000*
Q12	4.135	0.793		

Consequently, it can be deduced that the students perceive the achievement tests as possessing a high degree of face validity. However, when asked specifically about speaking, video and laboratory courses and listening practices focused on in the courses, students appear to believe that they are not well represented in the exams. Seven participants' (P19, P20, P21, P27, P35, P36 and P44) responses to the open-ended questions support these findings.

(Participant 19) **A listening section should definitely be added to the exams, and a separate speaking exam should be conducted.**

(Participant 20) In the exams we usually came across the contents of the grammar courses. **We were expecting the contents of the laboratory courses to be included in the exams as well, however they were not.**

(Participant 21) **In addition to the exams the students might be asked to answer the questions of an instructor in English,** and it might be a good idea to add the grades taken from this oral exam to the students' average.

(Participant 27) **The contents of all the courses should be included in the exams.** The students should not be responsible for studying only the contents of the main course book, "Quartet" and the grammar book "Milestones of English Grammar".

(Participant 35) **It would have been better, if the exams had included a speaking section.**

(Participant 36) The exams were instructive in terms of grammar. **However, they were insufficient in terms of speaking.**

(Participant 44) **The exams should have represented the contents of all courses.** By this way we could have learnt English better, and we would not have thought that speaking and laboratory courses were useless.

Additionally, it seems that students are undecided about whether the contents of the reading courses are well represented in the exams. One participant's (P40) response to the open-ended questions supports this finding.

(Participant 40) **I believe that the vocabulary taught in the reading courses are not well represented in the exams.**

The above mentioned quotes indicate that, with the exception of the last comment about vocabulary, the students seem to share the instructors' concern about the representation of the listening and speaking courses on the exams.

Difference between instructors' and students' perceptions of the face validity of the achievement tests

The questions in the first section of the instructors' questionnaire were parallel to the questions in the first section of the students' questionnaire, and all of these questions were concerned with the face validity of the achievement tests. Therefore, in order to find the difference between the two groups' perceptions of face validity, the means which represent the means of questions in the first sections of both questionnaires were compared, after excluding the aspects of the curriculum that are known by the researcher to be not included in the tests. Table 19 below shows the findings of the independent samples t-test.

Table 19 - Comparison of Instructors' and Students' Perceptions of Face Validity

	Group	N	Mean	Std. Deviation	t-value	*p*-value
Q1-4, Q7-11 (total)	**Instructors**	29	3.678	0.553	-1.164	0.248
	Students	52	3.529	0.548		

As a result of this comparison it has been found that these means are not significantly different from each other, and both instructors and students perceive the achievement tests as possessing a high degree of face validity. Furthermore, to have a better understanding, the twelve questions in the first section of the instructors'

questionnaire were compared with the corresponding questions in the students' questionnaire. Table 20 below shows the findings of this comparison.

Table 20 - Detailed Comparison of Instructors' and Students' Perceptions of Face Validity

Questions	Groups	Mean	Std. Deviation	t-value	*p*-value
Q1 The content of the main course book 'Quartet' was represented in the exams sufficiently.	Students	4.135	0.864	-0.991	0.325
	Instructors	4.310	0.542		
Q2 The content of the grammar book 'Milestones' was represented in the exams sufficiently.	Students	4.039	0.989	0.182	0.856
	Instructors	4.000	0.756		
Q3 The content of the writing courses was represented in the exams sufficiently.	Students	3.557	1.092	-1.257	0.213
	Instructors	3.862	0.953		
Q4 The content of the reading courses was represented in the exams sufficiently.	Students	3.019	1.075	-1.157	0.251
	Instructors	3.310	1.105		
Q5 The content of the speaking courses was represented in the exams sufficiently.	Students	2.000	0.990	-1.142	0.257
	Instructors	2.276	1.131		
Q6 The content of the video courses was represented in the exams sufficiently.	Students	1.942	0.850	-1.018	0.312
	Instructors	2.138	0.789		
Q7 Grammar taught in the courses was represented in the exams sufficiently.	Students	4.654	0.480	1.265	0.212
	Instructors	4.483	0.634		
Q8 The vocabulary taught in the courses was represented in the exams sufficiently.	Students	4.365	0.658	1.756	0.083
	Instructors	4.069	0.842		
Q9 The listening practices focused on in the courses were represented in the exams sufficiently.	Students	2.077	0.968	-1.622	0.112
	Instructors	2.517	1.271		
Q10 The content of the laboratory courses was represented in the exams sufficiently.	Students	2.328	1.043	-1.034	0.304
	Instructors	2.586	1.150		
Q11 The exercises made in the courses were represented in the exams sufficiently.	Students	3.596	1.272	-1.702	0.093
	Instructors	3.966	0.680		
Q12 In general, the contents of the courses were represented in the exams sufficiently.	Students	4.135	0.793	2.444	0.017*
	Instructors	3.656	0.936		

It has been observed that there is a significant difference between instructors and students perceptions only when Q12 is considered. However, both means fall into the same category, agree. Consequently, the significant difference between them simply suggests a slightly higher degree of agreement on the part of the students.

The Extent to Which the Achievement Tests Possess Reliability

The questions in section II of the instructors' and students' questionnaires sought to answer research question two, which investigates instructors' and students' perceptions of the reliability of the achievement tests administered in Zonguldak Karaelmas University Prep School. In this part data are presented in the form of four subsections: instructors' perceptions of scorers' reliability, students' perceptions of reliability in terms of the structure of the tests, students' perceptions of reliability in terms of testing conditions and students' perceptions of reliability in general.

Instructors' perceptions of scorers' reliability

All thirteen questions in the second section of instructors' questionnaire are concerned with the instructors' perceptions of scorers' reliability. In order to determine the instructors' perceptions of scorers' reliability, the mean which represents the means of these thirteen questions was calculated. Among these questions the 5^{th} and 10^{th} questions have negative orientations, and this was taken into consideration during the data analysis. In other words, the scoring system was reversed for these questions. There are also negative questions in the second section of the students' questionnaire. In order to emphasize such questions' negative orientation, the symbol $^{+}$ has been placed on the right upper side of these questions, and the following scale has been used during the data analysis of these negative questions:

Figure 2 - Reversed Rating Scale for Interpreting Negatively-Oriented Likert-Scale Responses

Mean	Degree	Opinion
4.5-5	Very low	Strongly Disagree
3.5-4.4	Low	Disagree
2.5-3.4	Moderate	Undecided
1.5-2.4	High	Agree
1.1-1.4	Very high	Strongly agree

Table 21 below shows the mean which represents the means of the thirteen questions.

Table 21 - Mean of the Means, Scorers' Reliability

TOTAL (SECTION-2)	Group	N	Mean	Std. deviation
	Instructors	29	3.851	0.487

As can be seen in Table 21, the mean is 3.851. Consequently, it can be assumed that the degree of scorers' reliability is high. Additionally, a further analysis was made to define the validity of this section of the instructors' questionnaire. The mean which represents the means of the first 12 questions was compared with the mean of the 13th question, which was a question about the instructors' overall perceptions of the scorers' reliability, using an independent samples t-test. Table 22 shows the findings of this comparison.

Table 22 - Validity Analysis, Scorers' Reliability

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-12 (total)	3.8736	0.480	-3.248	0.003*
Q13	3.586	0.733		

It has been observed that there is a significant difference between the mean which represents the means of the first 12 questions and the mean of question 13. However, both fall into the same category, agree. Therefore, it can be assumed that this section of the instructors' questionnaire is valid.

Furthermore, with the aim of gaining greater insights about the perceptions of the instructors of scorers' reliability, the related questions and their means and standard deviations are presented in Table 23.[5]

Table 23 - Detailed Analysis of Scorers' Reliability

Questions	Mean	Std. Deviation
Q1 The questions included in the exams permitted objective scoring	3.724	0.960
Q2 Testing office provided a detailed answer key	4.138	0.915
Q3 The scorers who marked the exam papers were trained	3.344	0.936
Q4 Students were identified by number, not name when scoring was subjective (e.g., in writing sections) to provide objectivity	2.448	1.088
Q5^{+} Only one instructor scored each exam paper when scoring was subjective	3.448^{+}	1.298^{+}
Q6 The rating scales included in the key helped me while I was scoring the exam papers	3.828	1.002
Q7 We had meetings to agree with acceptable answers after the exams	4.448	0.572
Q8 The class which I instructed as the main course teacher and the class which I invigilated during the exams were two different classes	4.670	0.541
Q9 The class which I instructed as the main course teacher and the class whose papers I scored were two different classes	4.670	0.541
Q10 $^{+}$ The deadline for scoring and returning the exam papers to the main course instructors affected my scoring practices negatively	3.757^{+}	1.123^{+}
Q11 I scored the exam papers in a reliable manner	4.621	0.561
Q12 All my colleagues scored the exam papers in a reliable manner	3.445	0.856
Q13 In general, the scoring system was reliable	**3.586**	**0.733**

As can be seen in Table 23, while the instructors disagree with Q4 and Q10 (i.e. the deadline did not affect their scoring practices negatively) they are undecided about Q3, Q5 and Q12. Furthermore, the table shows that the instructors agree with Q1, Q2, Q6, Q7 and Q13, and they strongly agree with Q8, Q9 and Q11.

It can be concluded that the instructors perceive the scorers' reliability as high. However, when asked specifically whether students were identified by number, not

[5] The questions which have the symbol $^{+}$ on the right upper side have negative orientations.

name when scoring was subjective (e.g., in writing sections) to provide objectivity, instructors appear to believe that students were not identified by number when scoring was subjective. One participant's (P17) response to the open-ended questions supports this finding.

> (Participant 17) **We identified students by their names, not numbers.** In other words, we knew whose paper we were scoring. It would have been better, if we had not seen the students' names while scoring.

Additionally, instructors appear to be undecided about whether only one instructor scored each exam paper when scoring was subjective, and whether the scorers who marked the exam papers were trained. Five participants' (P3, P4, P7, P12 and P19) responses to the open-ended questions support the latter finding.

> (Participant 3) Scoring is an important task. We should give more importance to it. **In order to promote scorer reliability in my institution instructors might be trained in scoring by someone who is well-equipped in scoring.**
>
> (Participant 4) **Not having any training in testing is what hinders scorer reliability in my institution.**
>
> (Participant 7) **For writing tests some standardization training can be given to the instructors.**
>
> (Participant 12) **Instructors' lack of knowledge in testing and scoring is what hinders scorer reliability.**
>
> (Participant 19) **As instructors are not trained, I do not believe that scoring practices are reliable in my institution.** Instructors must be trained.

Lastly, instructors appear to be undecided when asked specifically whether all their colleagues scored the exam papers in a reliable manner. Six participants' (P8, P10, P12, P14, P15 and P20) responses to the open-ended questions support this finding.

> (Participant 8) **Instructors must take marking serious in my institution.**
>
> (Participant 10) The sections requiring subjective scoring hindered the reliability of our exams, especially the writing sections. **Completely different grades were given to the same students by different scorers.**
>
> (Participant 12) Maybe the number of quizzes hinders scorer reliability. There are lots of quizzes during one academic year. **Therefore, instructors may not find sufficient time to evaluate the exam papers in a reliable manner.**
>
> (Participant 14) **Exams must be scored objectively by the scorers.**
>
> (Participant 15) **Instructors have different views about the writing sections of the exams.** This hinders scorer reliability in my institution.
>
> (Participant 20) Carelessness is an important factor. **The scorers mark the exam papers quickly, and this leads to mistakes.**
>
> (Participant 20) **Sometimes the scorers do not read the reading text themselves. They look for the answers given in the key. However, comprehension questions can be answered by the students in a number of ways.**

All the above mentioned concerns of the instructors seem to stem from a lack of training or a lack of time for scoring.

Students' perceptions of reliability in terms of the structure of the tests

Q1-6, Q8-12 and Q21 in the second section of students' questionnaire are concerned with reliability in terms of the structure of the tests. Among these questions Q1, Q2, Q3, Q8 and Q21 have negative orientations, and this was taken into consideration during the data analysis. In order to arrive at an estimation of reliability in terms of the structure of the tests, the mean which represents the means of these 12 questions were calculated. Table 24 below shows the mean which represents the means of these 12 questions.

Table 24 - Mean of the Means, Reliability of Test Structure

TOTAL (SECTION-2/ subsection 1)	Group	N	Mean	Std. Deviation
	Students	52	3.937	0.396

As can be seen in Table 24, the mean is 3. 937. Consequently, it can be assumed that the degree of reliability in terms of the structure of the tests is high. Additionally, a further analysis was made to define the validity of this subsection of the students' questionnaire. The mean which represents the means of Q1-6 and Q8-12 was compared with the mean of the 21st question, which was a question about the students' overall perceptions of reliability in terms of the test structure, using an independent samples t-test. Table 25 shows the findings of this comparison.

Table 25 - Validity Analysis, Reliability of Test Structure

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-6, Q8-12 (total)	3.969	0.373	-3.031	0.004*
Q21[6]	3.596+	1.034		

[6] This question is a negative one: In general, the structure of the tests hindered my ability to display my best performance in the exams. Since the question is negative, the reversed scale is used. According to the reversed scale, the mean 3.596 corresponds to disagree or low, and this means that the students agree that tests did not hinder their ability to display their best performance in the exams, and the reliability is high.

It has been observed that there is a significant difference between the mean which represents the means of the 11 related questions and the mean of question 21. However, both fall into the same category, high. Therefore, it can be assumed that this subsection of the students' questionnaire is valid. Furthermore, with the aim of gaining greater insights about reliability in terms of the structure of the tests, the related questions and their means and standard deviations are presented in Table 26.[7]

Table 26 - Detailed Analysis, Reliability of Test Structure

Questions	Mean	Std. Deviation
Q1^{+} Sometimes, two (or more) questions in the test seemed to be closely related, so that if I could not answer one question, I could not answer the other question either.	2.846^{+}	1.055^{+}
Q2^{+} The exams included too many questions.	3.385^{+}	1.105^{+}
Q3^{+} The exams included an insufficient number of questions.	3.904^{+}	0.747^{+}
Q4 The instructions explaining what to do in each section in the exams were explicit and clear.	3.769	1.096
Q5 The points allotted for each section of the exam were always stated in the exam papers.	4.692	0.643
Q6 Time given to the students to complete the exam was always stated in the exam papers.	4.556	0.698
Q8^{+} All the questions in the exams had the same difficulty level.	3.654^{+}	0.947^{+}
Q9 The exam questions were explicit and clear.	3.692	0.919
Q10 The lay out of the exam papers was fine.	4.365	0.687
Q11 The exam papers were legible.	4.556	0.574
Q12 The tables which were employed in the exams were clear and easy to interpret.	4.231	0.757
Q21^{+} **In general, the structure of the tests hindered my ability to display my best performance in the exams.**	**3.596**$^{+}$	**1.034**$^{+}$

As can be seen in Table 26, while the students are undecided about Q1 and Q2 they disagree with Q3, Q8, and Q21 (i.e., they felt that the number of the questions was sufficient, the questions were of varying difficulty level, and the structure of the tests did not hinder their performance). Furthermore, it has been observed that the students agree

[7] The questions which have the symbol $^{+}$ on the right upper side have negative orientations.

with Q4, Q9, Q10 and Q12. Lastly, it has been found that the students strongly agree with Q5, Q6 and Q11.

It has been found that the students perceive reliability in terms of the structure of the tests as high. However, they appear to be undecided when asked specifically whether two (or more) questions in the test seemed to be closely related, so that if they could not answer one question, they could not answer the other question either. One participant's (P1) response to the open-ended questions supports this finding.

> (Participant 1) Questions were asked in groups in the grammar book, "Milestones of English Grammar". Consequently, questions were asked in groups in the exams as well. This means that **if one question was wrong, there was a possibility that other questions within the group were wrong too.**

Furthermore, students appear to be undecided when asked specifically whether the exams included too many questions, and one participant's (P26) response to the open-ended questions supports this finding.

> (Participant 26) **More questions could have been asked, in order to test all of what we learned.**

This student's remark might have been prompted by a concern that the entire curriculum is not represented on the tests.

Students' perceptions of reliability in terms of testing conditions

Q7, Q13-20, and Q22 in the second section of the students' questionnaire are concerned with reliability in terms of testing conditions. Therefore, in order to determine the students' perceptions of the tests' reliability in terms of testing conditions, the mean which represents the means of these 10 questions was calculated. Among these questions

Q14-15, Q17-20 and Q22 have negative orientations, and this was taken into consideration during the data analysis. Table 27 below shows the mean which represents the means of these 10 questions.

Table 27 - Mean of the Means, Reliability of Testing Conditions

TOTAL (SECTION-2/ subsection 2)	Group	N	Mean	Std. Deviation
	Students	52	3.777	0.530

As can be seen in Table 27, the mean is 3.777. Consequently, it can be assumed that the degree of reliability in terms of testing conditions is high. Additionally, a further analysis was made to define the validity of this subsection of the students' questionnaire. The mean which represents the means of Q7 and Q13-20 was compared with the mean of the 22nd question, which was a question about the students' overall perceptions of the tests' reliability in terms of testing conditions, using an independent samples t-test. Table 28 shows the findings of this comparison.

Table 28 - Validity Analysis, Reliability of Testing Conditions

Questions	Mean	Std. Deviation	t-value	*p*-value
Q7, Q13-20 (total)	3.761	0.523	1.255	0.215
Q22[8]	3.923+	1.064		

It has been observed that there is no significant difference between the mean which represents the means of the 9 related questions and the mean of question 22, and both means fall into the same category, high. Therefore, it can be assumed that this subsection of the students' questionnaire is valid. Furthermore, with the aim of gaining

[8] This question is a negative one: In general, the bad environmental conditions hindered my ability to display my best performance in the exams. Since the question is negative, the reversed scale is used. According to the reversed scale, the mean 3.923 corresponds to disagree or low, and this means that the students agree that the bad environmental conditions did not hinder their ability to display their best performance in the exams, and the reliability is high.

greater insights about reliability in terms of the structure of the tests, the related questions and their means and standard deviations are presented in Table 29.[9]

Table 29 - Detailed Analysis, Reliability of Testing Conditions

Questions	Mean	Std. Deviation
Q7 Information about how much the given tests would affect the final grade was always announced.	3.519	1.393
Q13 The instructors helped us to get used to the format of the exams.	4.039	0.949
Q14^{+} The time given to complete the exams was too short.	3.865^{+}	1.048^{+}
Q15^{+} The time given to complete the exams was too long.	3.308^{+}	1.001^{+}
Q16 Equal timing was given to all classes which took the same test.	4.558	0.669
Q17^{+} Distracting sounds and noises lowered my performance in the exams.	3.077^{+}	1.266^{+}
Q18^{+} The little amount of light in the classrooms lowered my performance in the exams.	4.096^{+}	1.071^{+}
Q19^{+} The degree of the temperature in the classrooms lowered my performance in the exams.	3.750^{+}	1.135^{+}
Q20^{+} The little amount of air in the classrooms lowered my performance in the exams.	3.635^{+}	1.048^{+}
Q22^{+} In general, the bad environmental conditions hindered my ability to display my best performance in the exams.	**3.923**$^{+}$	**1.064**$^{+}$

It has been found that while the students agree with Q7 and Q13, they are undecided about Q15 and Q17. Furthermore, the table shows that the students strongly agree with Q16, and they disagree with Q14, Q18, Q19, Q20 and Q22. In other words, they disagreed with the negatively-oriented questions, indicating a high degree of reliability for these testing conditions.

Lastly, the analysis revealed that the students perceive the reliability in terms of testing conditions as high, but they appear to be undecided about whether the time given to complete the exams was too long and whether distracting sounds and noises lowered their performance in the exams. Seven participants (P3, P16, P20, P21, P22, P32 and P46) responses to the open-ended questions support the latter finding.

[9] The questions which have the symbol $^{+}$ on the right upper side have negative orientations.

(Participant 3) **One of the students was coughing continuously.** He/she should have taken the exam in another classroom.

(Participant 16) **Sometimes students who were late for the exams made noise.** The teachers tried to prevent such noises; however this was students' responsibility.

(Participant 20) **Noise of the students who completed the exam and left the classroom distracted my attention a lot.** In short, the exams did not use to end as silent as they had started.

(Participant 21) **The invigilators' chat among themselves and with the students distracted my attention a lot during the exams.**

(Participant 22) **Noise caused by the instructors' high-heeled shoes distracted my attention.**

(Participant 32) **Noise caused by the instructors who were wandering around the classroom distracted my attention.**

(Participant 46) **Noise caused by the students who were trying to cheat distracted my attention.**

Clearly, the students are concerned about the level of noise during the testing situation, caused by both the instructors and the students.

Students' perceptions of reliability in general

All 22 questions in the second section of students' questionnaire were concerned with the students' perceptions of the reliability of the tests. Therefore, in order to determine students' general perceptions of the reliability of tests, the mean which represents the means of these 22 questions was calculated. Among these questions Q1-3,

Q8, Q14-15 and Q17-22 have negative orientations, and this has been taken into consideration during the data analysis. Table 30 below shows the mean which represents the means of these 22 questions.

Table 30 - Mean of the Means, Students' General Perceptions of Reliability

TOTAL (SECTION-2)	Group	N	Mean	Std. Deviation
	Students	52	3.864	0.420

As can be seen in Table 30, the mean is 3.864. Consequently, it can be assumed that the degree of reliability in general is high. Additionally, a further analysis was made to define the validity of this section of the students' questionnaire. The mean which represents the means of the first 20 questions in this section was compared with the mean which represents the means of Q21 and Q22, the questions about the students' overall perceptions of the tests' reliability. Table 31 shows the findings of this comparison.

Table 31 - Validity Analysis, Students' General Perceptions of Reliability

Questions	Mean	Std. Deviation	t-value	*p*-value
Q1-20 (total)	3.875	0.402	1.166	0.249
$Q21^{+}$- 22^{+} (total)	3.759	0.888		

It has been observed that there is no significant difference between the mean which represents the means of the 20 related questions and the mean which represents the means of Q21 and Q22, and they fall into the same category, high. Therefore, it can be concluded that this section of the students' questionnaire is valid, and the degree of reliability in general is high.

The Extent to Which the Achievement Tests Possess Predictive Validity

The midterm and the end-of-course assessment scores of the students were used to answer research question three, which addresses the extent to which the achievement tests administered in Zonguldak Karaelmas University Prep School possess predictive validity. In this part data are presented in the form of four subsections: the correlation between students' first term averages and second term averages, the correlation between students' first term averages and end-of-course assessment scores, the correlation between students' second term averages and end-of-course assessment scores, and the correlation between students' cumulative averages and end-of-course assessment scores.

The correlation between first term and second term averages

First, the correlation between first term and second term averages was investigated in order to determine the predictive validity of the first term achievement tests. The results are given in Table 32.

Table 32 - Correlation, First and Second Term Averages

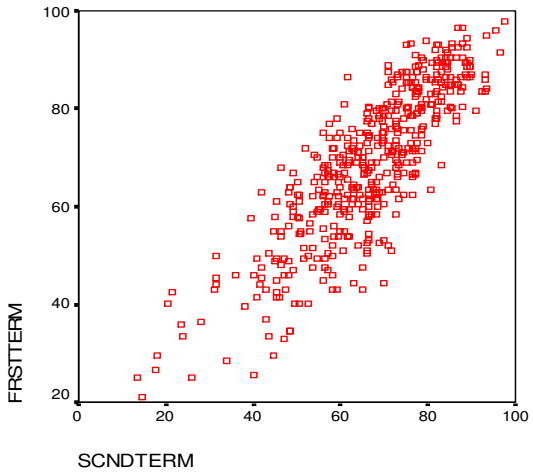
		FRSTTERM	SECONDTERM
FRSTTERM	Pearson Correlation	1	.844(**)
	Sig. (2-tailed)	.	.000
	N	477	477

****p*=<.000**

It has been observed that there is a significant positive correlation (.844, $p<.000$) between the first term and second term averages. Furthermore, in order to see the strength of the correlation graphically, each student's first term and second term averages were placed on a scatter plot diagram (see Figure 3). At the end of the analysis it has been found that the achievement tests conducted in the first term have a high level

of predictive validity. In other words, students' performances on the first term achievement tests can predict their performances on the second term achievement tests.

Figure 3 - The Correlation between First Term and Second Term Averages



The correlation between first term averages and the end-of-course assessment scores

The correlation between first term averages and the end-of-course assessment scores was examined in order to determine the predictive validity of the first term achievement tests. The results are given in Table 33.

Table 33 - Correlation, First Term Averages and the End-of-Course Assessment Scores

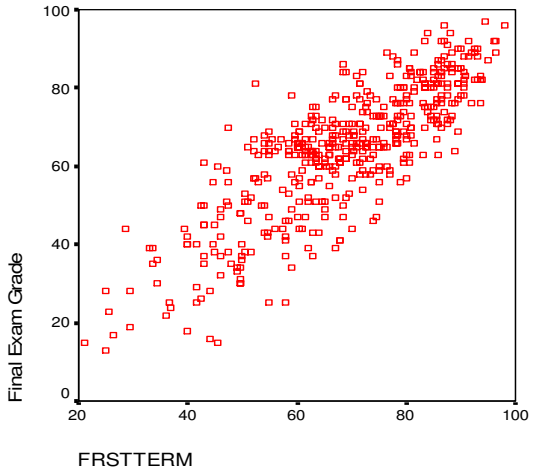
		Final Exam Grade
FRSTTERM	Pearson Correlation	.824(**)
	Sig. (2-tailed)	.000
	N	477

****p*=<.000**

It has been observed that there is a significant positive correlation (.824, $p<.000$) between the first term averages and the end-of-course assessment scores. In order to see the strength of the correlation graphically, students' first term averages and their end-of-course assessment scores were placed on a scatter plot diagram (see Figure 4). The analysis once again revealed that the achievement tests conducted in the first term have a

high level of predictive validity. In other words, students who do well on the first term midterm tests tend to do well on the end-of-course assessment.

Figure 4 - The Correlation between First Term Averages and the End-of-Course Assessment Scores



The correlation between second term averages and the end-of-course assessment scores

The correlation between second term averages and the end-of-course assessment scores was investigated in order to determine the predictive validity of the second term achievement tests. The results are given in Table 34.

Table 34 - Correlation, Second Term Averages and the End-of-Course Assessment Scores

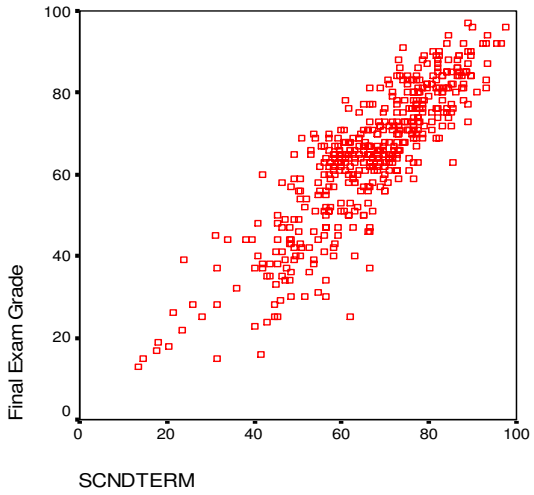
		Final Exam Grade
SCNDTERM	Pearson Correlation	.870(**)
	Sig. (2-tailed)	.000
	N	477

****p*=<.000**

It has been observed that there is a significant positive correlation (.870, $p<.000$) between the second term averages and the end-of-course assessment scores. In order to see the strength of the correlation graphically, students' second term averages and their end-of-course assessment scores were placed on a scatter plot diagram (see Figure 5).

The analysis revealed that the achievement tests conducted in the second term have a high level of predictive validity.

Figure 5 - The Correlation between Second Term Averages and the End-of-Course Assessment Scores



The correlation between students' cumulative averages and the end-of-course assessment scores

The correlation between students' cumulative averages (including four midterms conducted throughout the 2005-2006 academic year) and the end-of-course assessment scores was examined in order to determine the predictive validity of the achievement tests conducted throughout the year. The results are given in Table 35.

Table 35 - Correlation, Cumulative Averages and the End-of-Course Assessment Scores

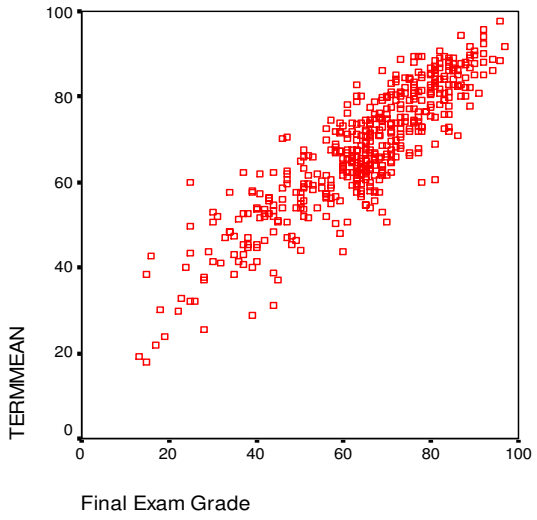
		Final Exam Grade	Cumulative Average
Final Exam Grade	Pearson Correlation	1	.881(**)
	Sig. (2-tailed)	.	.000
	N	477	477

****p*=<.000**

There is a significant positive correlation (.881, $p<.000$) between students' cumulative averages and end-of-course assessment scores. Additionally, in order to see

the strength of the correlation graphically, students' cumulative averages and the end-of-course assessment scores were placed on a scatter plot diagram (see Figure 6). This shows that the achievement tests conducted throughout the year, when considered together, have a high level of predictive validity. In other words, the students' performances throughout the year are a good predictor of their final achievement scores.

Figure 6 - The Correlation between Cumulative Averages and the End-of-Course Assessment Scores



Conclusion

In this chapter, the data obtained from the questionnaires and assessment scores were analyzed and presented in four parts. In the first part, the data consisted of Likert scale questions in section I of both instructors' and students' questionnaire which answered research question 1, regarding face validity. In the second part, the data consisted of Likert scale questions in section II of both instructors' and students' questionnaire which answered research question 2, regarding reliability. In the last part, the data gathered from assessment scores were presented quantitatively in order to address research question 3, regarding predictive validity.

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

CHAPTER V: CONCLUSION

Introduction

This study has been conducted to investigate the relationship between face validity and relatively more objective measures: reliability and predictive validity. It has also been aimed to examine the predictive validity, face validity and reliability of tests administered at Zonguldak Karaelmas University Preparatory School. The research questions posed for the study are as follows:

1. To what extent do the achievement tests possess face validity?
    - To what extent do the achievement tests represent the course content in the eyes of the instructors?
    - To what extent do the achievement tests represent the course content in the eyes of the students?
    - Is there a difference between the two groups' perceptions of the achievement tests' representativeness of the course content?
2. To what extent do the achievement tests possess reliability?
    - To what extent does the current testing system permit scorer reliability?
    - To what extent do the structure of the tests and the testing conditions permit students to accurately demonstrate their language knowledge and skills?
3. To what extent do the achievement tests possess predictive validity?
    - How well do the achievement tests conducted in the first term predict success in the second term?

- How well do the achievement tests conducted throughout the year predict success in the end-of-course assessment?

4. How closely does the face validity of the achievements tests reflect the reliability and predictive validity of these tests?

Overview of the Study

Two different groups of participants were included in this study. Twenty nine C-level instructors who were working at Zonguldak Karaelmas University Foreign Languages Compulsory Preparatory School in the 2005-2006 academic year formed the first group, and 52 C- level undergraduate students who were enrolled in the same institution during the same period formed the second group.

The instruments employed in this study were two questionnaires (one for the instructors and one for the students) and test scores of 477 C- level students who were enrolled in the institution in the 2005-2006 academic year. The instructors' questionnaire was composed of four sections (see Appendix D). The first two sections involved Likert scale items. In the first section, there were 12 questions which aimed to investigate the instructors' perceptions of the face validity of the achievement tests. The next section was composed of 13 questions, and these questions were about the instructors' perceptions of whether the current testing system permits scorer reliability or not. Section III consisted of four open-ended questions which were designed to obtain instructors' additional comments on the reliability and validity of achievement tests. In the last section, there were four questions which aimed to gather background information about the instructors.

The students' questionnaire was also composed of four sections (see Appendix E for the Turkish version and Appendix F for the English version of the questionnaire). The first two sections of this questionnaire involved Likert scale items. In the first section, there were 12 questions which aimed to investigate the students' perceptions of the face validity of the achievement tests. The next section was composed of 22 questions, and these questions were about the students' perceptions of the reliability of tests in terms of their performance. Section III consisted of two open-ended questions which were designed to obtain students' additional comments on the reliability and validity of achievement tests. In the fourth section, there were two questions which aimed to gather background information about the students.

Apart from the questionnaires, the correlations between students' first term averages, second term averages, cumulative averages (consisting of averages of four midterms conducted throughout the year) and end-of-course assessment scores were computed to establish the degree of predictive validity.

Discussion of Findings

The findings of this study will be presented in four different sections: the extent to which the achievement tests possess face validity, the extent to which the achievement tests possess reliability, the extent to which the achievement tests possess predictive validity and the extent to which the face validity of the achievements tests reflects the reliability and predictive validity of these tests. These sections correspond to the four research questions.

The Extent to Which the Achievement Tests Possess Face Validity

With the purpose of determining the extent to which the achievement tests possess face validity, both instructors' and students' were asked their opinions about the achievement tests' representativeness of the course content, and their opinions were compared with one another. Analysis of the results revealed that the achievement tests represent the course content to a high degree both in the eyes of the instructors and the students. In other words, there is no difference between the two groups' perceptions of the achievement tests' representativeness of the course content. Therefore, it can be concluded that the achievement tests possess face validity to a high degree.

However, data gathered from the first section of the questionnaires pointed to such weaknesses of the current testing practices as the lack of listening and speaking sections in the exams and not incorporating the contents of the video and laboratory courses into the exams. The data gathered from the open-ended questions supported the above-mentioned findings. Three instructors suggested that a listening and a speaking section should be included in the exams. Additionally, nine students stated that a speaking section should be included in the exams, and seven students emphasized the necessity of adding a listening section to the exams. Lastly, four students indicated that the content of the laboratory courses should be incorporated into the exams.

The Extent to Which the Achievement Tests Possess Reliability

With the aim of determining the extent to which the achievement tests possess reliability, instructors were asked their opinions about scorers' reliability. The findings show that scorers' reliability is high. Additionally, the students were questioned about reliability in terms of the structure of the tests and reliability in terms of their

performance. The findings indicate that reliability, in terms of both the structure of the tests and students' performance, is high. Lastly, the findings suggest that the reliability of the achievement tests in general is high.

On the other hand, the data gathered from the second section of the instructors' questionnaire pointed to a specific weakness of the current testing practices: that of not identifying the students by number, instead of names, when scoring was subjective. The data gathered from the open-ended questions also supported this finding. One instructor said that identifying students by name, not number, hindered scorers' reliability.

The Extent to Which the Achievement Tests Possess Predictive Validity

The results of the Pearson Product Moment Correlation coefficient revealed that there was a significant positive correlation (.844, $p<.000$) between the first term and second term averages. Additionally, it was found that there was a significant positive correlation (.824, $p<.000$) between the first term averages and the end-of-course assessment scores. It was also revealed that there was a significant positive correlation (.870, $p<.000$) between the students' second term averages and the end-of-course assessment scores. Lastly, the analysis indicated that there was a significant positive correlation (.881, $p<.000$) between the students' cumulative averages and the end-of-course assessment scores.

These findings suggest that the predictive validity of the achievement tests conducted in Z.K.U. is high. Consequently, the test scores can be employed to make inferences concerning students' achievement on the following tests administered in Prep School and to diagnose and treat the weaknesses of the students. In this way, students might learn from their mistakes and their success might increase. Additionally, the test

scores might be used to make inferences concerning students' future achievement in their department English courses. In other words, students who have been successful in Prep School might also be successful in the English courses which they will take in their departments, and those who have been unsuccessful in Prep School might also be unsuccessful in their departments. Consequently, some measures might be taken by the Prep School administrators with the aim of preventing the unsuccessful students' future failure. For instance, summer courses can be opened with the aim of treating these students' weaknesses. In fact, there is a summer course for the students who failed in Prep School, but it aims to prepare the students for the proficiency exam which is a multiple choice test. In this course instructors teach multiple choice test techniques rather than treating the real weaknesses of the students. In other words, this course does not meet the future academic needs of the students. For that reason, in addition to the summer courses which prepare the students for the proficiency exam, academic English summer courses which really address the needs of the students might be opened.

The Extent to Which Face Validity Reflects Reliability and Predictive Validity

As mentioned above, the face validity and reliability of the achievement tests are high in the eyes of both students and instructors. The data gathered from the tests scores show that the predictive validity of the achievement tests is also high. These findings indicate that the face validity of the achievements tests reflects the reliability and predictive validity of these tests well.

As mentioned in the literature review, other researchers have looked at face validity along with more objective measures. Nakamura (2006) examined face validity through an informal questionnaire and discussions with 809 freshman university

students. Most of the students agreed that the test in question had face validity. Content validity was established through a discussion about the test items. The instructors discussed how well the test items reflect the content of the text book they were using and the content of their teaching. All the English instructors involved in the test construction process agreed that the pilot placement test possessed content validity.

Ösken (1999) examined the face validity and a more objective measure, content validity, of the end-of-course assessment administered at Hacettepe University, Department of Basic English (DBE) in the 1997-1998 academic year. The findings indicated that the end-of-course assessment represented the course contents in the eyes of the instructors; however there was a limited representation of the course contents when the proportions of language items in the course books were compared with the test items in the end-of-course assessment. According to Ösken, the mismatch between face validity and content validity might have been due to the lack of test objectives. Furthermore, according to Ösken, the number of course objectives was high in terms of grammar, and it was impossible to test all aspects of grammar. Therefore, the testers might have chosen the main structures to test while ignoring the others.

The other researcher who investigated both face validity and an objective measure, the content validity of tests, is Serpil (2000). He looked at the face validity and content validity of midterm achievement tests administered at Anadolu University School of Foreign Languages. His findings indicated that the instructors in general thought that the midterm tests' representativeness of the courses' content was moderate to high. However, it was found that the degree of the tests' representativeness of the course material was low. In other words, in this study, face validity did not appear to

predict content validity. Serpil speculated that the lack of clearly defined testing criteria and course objectives was the main factor causing such a conflict among the results. The present study is different from the above mentioned studies in that face validity is compared not with content validity, but with reliability and predictive validity. It is possible that if content validity had been explored in the present study, it might have revealed a similar mismatch between content and face validity, in spite of the apparent correlation among face validity, reliability, and predictive validity.

Pedagogical Implications

According to the findings, the face validity and reliability of the achievement tests are high in the eyes of both students and instructors. The predictive validity of the achievement tests is also high. These findings show that face validity does not contradict with relatively more objective measures of tests such as reliability and predictive validity. However, face validity and reliability analyses revealed some important weaknesses in the testing system. These weaknesses would not have been revealed, if the researcher had looked at only face validity, or only reliability, or only predictive validity. Therefore, it is very important to look at tests from multiple perspectives, and get information from a variety of sources. In other words, using only one way of looking at tests might hinder seeing the whole picture.

Next, the questionnaires employed in this study might serve as checklists. In other words, other institutions might use these questionnaires by making some or no changes on them to check the face validity and reliability of their own achievement tests.

Furthermore, people from other institutions might read the instructors' and students' additional comments on testing practices carried out in Z.K.U. Then, they

might make use of Z.K.U. instructors' and students' suggestions and comments to promote the quality of tests within their institutions.

This study was conducted within a particular institution, Zonguldak Karaelmas University Prep School. Therefore, some of the pedagogical implications drawn from the study mainly concern the curriculum unit, testing office and the administrators of the institution in particular.

To start with, the results of this study show that there is a gap between the contents of some courses and the tests in the eyes of the stakeholders. Therefore, a speaking and a listening section should be included in the exams. Additionally, the contents of the laboratory, video and reading courses should be incorporated into the exams.

According to Hughes (2003) having clear, well-defined objectives helps teachers to teach and test their students better. The reason for this is that clear objectives provide criteria for the instructors who have to decide which language points to weight on the test over the others. It is obvious that the instructors who have to decide which language points to weight on the test over the others in Z.K.U. have failed to test the contents of some courses, and this failure has revealed that the institution does not have well-defined objectives. Consequently, it can be concluded that the members of the curriculum unit, which has been opened very recently, should be encouraged to define the goals and objectives of the program.

Although the degree of scorers' reliability has been found to be high, the instructors' responses to the Likert scale questions and the open-ended questions provided a basis for a number of suggestions about the scoring procedures, and these

suggestions should be considered in order to improve scorers' reliability. Some of these suggestions are as follows:

- Instructors should be trained in testing and scoring.
- Students should be identified by number, not name when scoring is subjective.
- More than one instructor should score each exam paper when scoring is subjective.
- The importance of scoring should be frequently emphasized by the administrators.
- Testing office members should prepare a second key soon after the exams after reviewing the answers of some students. The reason for this is that some answers are unpredictable and hard to score. By this way, testing office members can prepare a more detailed key including the scores suitable for the debated answers given by the students.
- Instructors may be led to study on their own on testing and scoring. They can be encouraged to read relevant articles.[10]

Similar to what is mentioned above, although the degree of reliability is high in the eyes of the students, their responses to the Likert scale questions and the open-ended questions provided a basis for a number of suggestions These suggestions should be considered in order to promote reliability in terms of the test takers' performance. Some of these suggestions are as follows:

[10] The complete list of instructors' suggestions about scoring procedures can be found in the last table presented in Appendix G.

- Two or more questions in a test should not be closely related, so that even if a student cannot answer one question she/he can still answer the other questions.
- Testing office members should make sure that the number of questions is appropriate to adequately measure the desired objectives.
- Testing office members should make sure that the duration of the test is appropriate for the number of items and the abilities of the students.
- The school administration should make sure that the testing environment is as quiet as possible, so as to reduce distraction.
- Exams must not be administered in the corridors.
- The instructors should be trained.[11]

Additionally, since the predictive validity of the achievement tests is high, the test scores can be employed to make inferences concerning students' achievement on the following tests administered in Prep School and to diagnose and treat the weaknesses of the students. In this way, students' might not make the same mistakes and their success might increase. Furthermore, the test scores might be used to determine the students who are likely to be unsuccessful in the English courses they will take in their departments, and these students may be encouraged to participate in the summer courses which might be opened with the aim of treating the real weaknesses and addressing the academic needs of the students.

[11] The complete list of students' suggestions about reliability in terms of the test takers' performance can be found in the last table presented in Appendix H.

Limitations of the Study

The reliability of the achievement tests was measured by looking at teachers' and students' perceptions of specific test characteristics and testing practices, rather than by direct measurement. This was felt to be not an objective, but a relatively objective way of measuring reliability.

Additionally, questionnaires for instructors and students were selected as the main research instruments in this study. The reason for this is that as Brown and Rodgers (2002) indicate, "if large scale information is needed from a great many people, questionnaires are typically a more efficient way of gathering that information" (p. 142). However, some of the instructors or students who want to insult or do not want to insult the institution might have answered the questions accordingly. Furthermore, the personal opinions of the instructors and the students about the testing office members might have affected the ratings. In other words, the participants might have behaved emotionally while they were filling in the questionnaires.

Lastly, it was felt that examining content and concurrent validity and internal reliability were beyond the scope of this thesis. However, the study would have been stronger if it had also included content and concurrent validity and internal reliability analyses. In this way, the findings obtained from these analyses could also be compared with face validity.

Implications for Further Studies

This study explored the predictive validity, face validity and reliability of tests used at Zonguldak Karaelmas University Prep School. Since this study was a local one, it might be replicated by researchers from other universities. In this way, other

institutions will also have the opportunity to assess the tests conducted in their language programs, and the quality of tests might increase within these language programs.

Furthermore, a research study examining the content and concurrent validity and internal reliability or consistency, in addition to what has been looked at (face validity, reliability, and predictive validity) in the current study, might be conducted with the aim of gaining greater insights about how well face validity reflects more objective measures of tests.

Conclusion

This research study investigated the validity and reliability of the achievement tests conducted at Zonguldak Karaelmas University Prep School. It also investigated how closely face validity reflects relatively more objective measures of tests such as reliability. The data were collected through two questionnaires and test scores.

The findings of the questionnaires revealed that the degree of face validity and reliability of the achievement tests conducted at Z.K.U. is high in the eyes of both the instructors and the students. The data gathered from the tests scores indicated that the predictive validity of the achievement tests is also high. These findings indicate that face validity reflects relatively more objective measures such as reliability and predictive validity well. However, face validity and reliability analyses revealed some important weaknesses in the testing system. These weaknesses would not have been revealed, if the researcher had looked at only face validity, or only reliability, or only predictive validity. Therefore, it is very important to look at tests from multiple perspectives, and get information from a variety of sources.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and assessment.* Cambridge: Cambridge University Press.

Axman, R. (1989). *Constructing classroom achievement tests.* (ERIC Document Reproduction Service No. ED315426)

Aydın, E. (2004). *Testers' perceptions of the test development process and teachers' and testers' attitudes towards the resulting achievement tests.* Unpublished master's thesis, Bilkent University, Ankara.

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

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

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

Boardman, C.A., & Frydenberg, J. (2002). Writing to communicate: *Paragraphs and essays* (2nd ed.). New York: Pearson Education, Inc.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1-25.

Brown, A., Davies, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (Eds.). (1999). Studies in language testing 7: *Dictionary of language testing.* Cambridge: Cambridge University Press.

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

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

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*, 653-675.

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

Brown, J. D., & Rodgers, T.S. (2002). *Doing second language research.* Oxford: Oxford University Press.

Brualdi, A. (1999). *Traditional and modern concepts of validity.* (ERIC Document Reproduction Service No. ED435714)

Butler, L. (2003). Password: *A reading and vocabulary text.* New York: Pearson Education, Inc.

Cardoso, R. M. F. (1998). *Authentic foreign language testing in a Brazilian university entrance exam.* (ERIC Document Reproduction Service No. ED423675)

Cohen, A.D. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle.

Cumming, J., & Maxwell, G. (1999). Contextualizing Authentic Assessment. *Assessment in Education, 6,* 177-196.

Cunningham, S., & Moor, P. (2003). *Cutting edge.* New York: Pearson Education, Inc.

Davies, A. (1990). *Principles of language testing.* Oxford: Basil Blackwell Ltd.

Davies, P., & Pearse, E. (2000). *Success in language teaching*. New York: Oxford University Press.

Dörnyei, Z. (2003). Questionnaires in second language research: *Construction, administration and processing.* Manwah, NJ: Lawrence Erlbaum Associates, Inc.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Halleck, G. B., & Moder, C. L. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly*, *29*, 733-758.

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

Hoekje, B., & Linnel, K. (1994). "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, *28*, 103-123.

Hughes, A. (2003). *Testing for language instructors* (2nd ed.). Cambridge: Cambridge University Press.

Kenyon, D., & Van Duzer, C. (2003). *Valid, reliable, and appropriate assessments for adult English language learners.* (ERIC Document Reproduction Service No. ED482742)

Kunnan, A, J, (2000). *Fairness and validation in language assessment.* Cambridge: Cambridge University Press.

Kuntasal, I. (2001). *Perceptions of teachers and testers of achievement tests prepared by testers in the Department of Basic English at Middle East Technical University*. Unpublished master's thesis, Bilkent University, Ankara.

Kuroki, K. (1996). *Achievement testing: A final achievement test model for Japanese junior high school students.* (ERIC Document Reproduction Service No. ED395449)

Küçük, F., Okumuş, N., Tekin, İ., Yaman, C. & Yorgancı, N. et al. (2006). Milestones of English grammar- *Perfecting and practicing English structure.* Istanbul: New Life Elt.

Lewkowicz, J. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, *17*, 43-64.

Manola, J. R., & Wolfe, E. W. (2000). *The impact of composition medium on essay raters in foreign language testing.* (ERIC Document Reproduction Service No. ED443836)

Marvin, L., & Simner, C. (1999). Postscript to the Canadian Psychological Association's Position Statement on the TOEFL. Retrieved June 6, 2007, from http://cpa.ca/documents/TOEFL. html

McHugh, A. (2003). *An ocean apart.* England: Pearson Education, Inc.

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

Nakamura, Y. (2006). Analysis of a placement test: an interim report of a pilot version. Retrieved June 6, 2007, from http://review.keio-up.co.jp/

Norris, J. M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum Magazine, 38*, 18.

Oppenheim, A.N. (1992). *Questionnaire design, interviewing and attitude measurement.* London: Pinter.

Ösken, H. (1999). *An assessment of the validity of the midterm and the end of course assessment tests administered at Hacettepe University Department of Basic English.* Unpublished master's thesis, Bilkent University, Ankara.

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

Q Group. (2005). *Quartet.* U.S.A.: Q Group Ltd.

Rudner, L. M., & Schafer, W. D. (2001). *Reliability.* (ERIC Document Reproduction Service No. ED458213)

Rudner, L. M. (1994). *Questions to ask when evaluating tests.* (ERIC Document Reproduction Service No. ED385607)

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, *23*, 1-34.

Serpil, H. (2000). *An assessment of the content validity of the midterm achievement tests administered at Anadolu University Foreign Languages Department.* Unpublished master's thesis, Bilkent University, Ankara.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, *14*, 341-349.

Simpson, M. (2004). *The challenge and real lives.* England: Pearson Education, Inc.

Spence-Brown, R. (2006). The real world and the language tester: considerations of authenticity and interactiveness in the design and assessment of language tests. Retrieved June 6, 2007, from jweb.kokken.go.jp/kenshu/Rs/robyn.htm

Taylor, L. (2005). Washback and impact. *ELT Journal, 39,* 154-155.

Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, UK: Prentice Hall International (UK) Ltd.

Weir, C. (1995). *Understanding and developing language tests.* New York: Phoenix ELT.

Yeğin, O.P. (2003). *The predictive validity of Başkent University proficiency exam (BUEPE) through the use of the three-parameter IRT model's ability estimates*. Unpublished master's thesis, Middle East Technical University, Ankara.

APPENDIX A

CONSENT LETTER FOR THE INSTRUCTORS

Dear Colleagues,

I am currently enrolled in 2007 MA TEFL Program at Bilkent University. I am carrying out a research study on instructors' and students' perceptions of the validity and reliability of achievement tests. This study, whose main instruments are two questionnaires, is expected to contribute to the testing system of Zonguldak Karaelmas University English Prep School, the literature and my research.

Therefore, I ask you to answer the questionnaire questions as honestly and efficiently as possible. Please, keep in mind that your responses will be kept confidential, and your completion of the questionnaire will be regarded as consent for my using the data obtained in my research study.

You should not transcribe your name on the questionnaire. However, some background information is needed to classify your answers and to make statistical comparisons. You will find the relevant section on page five. Finally, if you would like to receive feedback on the results of this research study, please transcribe your mail address on the blank provided at the end of the questionnaire. Thank you very much for devoting your time and contributions.

Funda Küçük
MA TEFL Program
Bilkent University, ANKARA
fundak79@yahoo.com

APPENDIX B

CONSENT LETTER FOR THE FORMER STUDENTS (TURKISH VERSION)

Sevgili Öğrenciler,

Ben halen Bilkent Üniversitesi 2007 MA TEFL programına kayıtlı bir yüksek lisans öğrencisiyim. Öğretmenlerin ve öğrencilerin yıl içinde yapılan sınavların geçerliliği ve güvenilirliği konusundaki görüşlerine ilişkin bir araştırma yapmaktayım. Temel araçları iki anket olan bu çalışmanın, Zonguldak Karaelmas Üniversitesi İngilizce Hazırlık Okulunun sınav sistemine, literatüre ve benim araştırmama katkıda bulunacağı umulmaktadır.

Bu nedenle, sizlerden anket sorularını mümkün olduğunca dürüst ve uygun şekilde cevaplamanızı rica ediyorum. Vermiş olduğunuz yanıtlar gizli tutulacaktır, ve bu anketi doldurmanız elde edilen verileri çalışmamda kullanmam için izin niteliği taşımaktadır.

Anketin üzerine isminizi yazmamalısınız. Fakat, cevaplarınızı sınıflandırmak ve istatistiksel karşılaştırmalar yapmak amacıyla özgeçmişinize dair bazı bilgilere ihtiyaç duyulmaktadır. İlgili bölümü beşinci sayfada bulabilirsiniz. Zamanınızı ayırdığınız için ve katkılarınızdan dolayı çok teşekkür ederim.

Funda Küçük
MA TEFL Programı
Bilkent Üniversitesi, ANKARA
fundak79@yahoo.com

APPENDIX C

CONSENT LETTER FOR THE FORMER STUDENTS (ENGLISH VERSION)

Dear Students,

I am a master's student who is currently enrolled in 2007 MA TEFL Program at Bilkent University. I am carrying out a research study on instructors' and students' perceptions of the validity and reliability of the exams conducted throughout the year. This study, whose main instruments are two questionnaires, is expected to contribute to the testing system of Zonguldak Karaelmas University English Prep School, the literature and my research.

Therefore, I ask you to answer the questionnaire questions as honestly and efficiently as possible. Your responses will be kept confidential, and your completion of the questionnaire will be regarded as consent for my using the data obtained in my research study.

You should not transcribe your name on the questionnaire. However, some background information is needed to classify your answers and to make statistical comparisons. You can find the relevant section on page five. Thank you very much for devoting your time and contributions.

Funda Küçük
MA TEFL Program
Bilkent University, ANKARA
fundak79@yahoo.com

APPENDIX D

INSTRUCTORS' QUESTIONNAIRE

Please, answer the following questions considering the exams administered last year (in the 2005-2006 academic year).

Section I ***Please put a (√) in the box which reflects your point of view best, and please choose only one answer for each statement.***

No	Questions about instructors' perceptions of the face validity of achievement tests.	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
		SA	A	U	DA	SD
1	The content of the main course book 'Quartet' was represented in the exams sufficiently.					
2	The content of the grammar book 'Milestones of English Grammar-*Perfecting and Practicing English Structure*' was represented in the exams sufficiently.					
3	The content of the writing courses was represented in the exams sufficiently.					
4	The content of the reading courses was represented in the exams sufficiently.					
5	The content of the speaking courses was represented in the exams sufficiently.					
6	The content of the video courses was represented in the exams sufficiently.					
7	Grammar taught in the courses was represented in the exams sufficiently.					
8	The vocabulary taught in the courses was represented in the exams sufficiently.					
9	The listening practices focused on in the courses were represented in the exams sufficiently.					
10	The content of the laboratory courses was represented in the exams sufficiently.					
11	The exercises made in the courses were represented in the exams sufficiently.					
12	In general, the contents of the courses were represented in the exams sufficiently.					

Section II

Scorers' reliability: refers to the degree to which test scores are free from instructors' measurement errors. In other words, the grades which are given by these instructors are as objective and trustworthy as possible. (This definition might help you while you are interpreting the questions).

Please put a (√) in the box which reflects your point of view best, and please choose only one answer for each statement.

No	Questions about instructors' perceptions of whether the testing system permitted scorer reliability or not.	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
		SA	A	U	DA	SD
1	The questions included in the exams permitted objective scoring.					
2	Testing office provided a detailed answer key.					
3	The scorers who marked the exam papers were trained.					
4	Students were identified by number, not name when scoring was subjective (e.g., in writing sections) to provide objectivity.					
5	Only one instructor scored each exam paper when scoring was subjective.					
6	The rating scales included in the key helped me while I was scoring the exam papers.					
7	We had meetings to agree on acceptable answers after the exams.					
8	The class which I instructed as the main course teacher and the class which I invigilated during the exams were two different classes.					
9	The class which I instructed as the main course teacher and the class whose papers I scored were two different classes.					
10	The deadline for scoring and returning the exam papers to the main course instructors affected my scoring practices negatively.					
11	I scored the exam papers in a reliable manner.					
12	All my colleagues scored the exam papers in a reliable manner.					
13	In general, the scoring system was reliable.					

Section III

1. Do you have other comments or suggestions about the content of the exams? If yes, please transcribe in the blanks provided.

……
……
……
……
……
……
……
……
……
……

2. What promoted scorer reliability in our institution in your opinion? You can list more than one item.

……
……
……
……
……
……
……
……
……
……

3. What hindered scorer reliability in our institution in your opinion? You can list more than one item.

……
……
……
……
……
……
……
……

4. Do you have any suggestions to promote scorer reliability within our institution? If yes, please transcribe in the blanks provided.

……
……

……
……
……
……
……
……
……
……

Section IV – Background Information

Please, tick (√) the suitable answer for you.

1. Which program did you last graduate from?
B.A. degree ( ) M.A. degree ( )

2. How long have you been teaching totally?
1 to 4 years ( ) 9 to 12 years ( )
5 to 8 years ( ) more than 13 years ( )

3.How long have you worked as a testing office member totally?
no experience ( ) 1 to 3 years ( )
less than one year ( ) more than 3 years ( )

4. Have you taken any courses on testing?
Yes ( ) No ( )

Would you like to receive feedback on the results of this research study? If yes, please transcribe your mail address on the blank provided.

Your mail address: ……

Thank you very much for your cooperation!

APPENDIX E

FORMER STUDENTS' QUESTIONNAIRE (TURKISH VERSION)

Lütfen aşağıdaki soruları geçen yıl (2005-2006 öğretim yılında) Hazırlık Okulunda yapılmış olan sınavları göz önünde bulundurarak cevaplayınız.

I.Bölüm ***Lütfen, sizin görüşünüzü en iyi biçimde yansıtan kutuyu (√) şeklinde işaretleyiniz, ve lütfen her bir ifade için yalnızca bir cevap seçiniz..***

No	Öğrencilerin ders içerikleri ve sınav içerikleri arasındaki uyum konusundaki görüşlerine ilişkin sorular	Kesinlikle Katılıyorum	Katılıyorum	Kararsızım	Katılmıyorum	Kesinlikle Katılmıyorum
1	Ana ders kitabı 'Quartet'in' içeriğine sınavlarda yeterince yer verildi.					
2	Dilbilgisi kitabı 'Milestones of English Grammar-Perfecting and Practicing English Structure'ın' içeriğine sınavlarda yeterince yer verildi.					
3	Yazı (writing) derslerinin içeriğine sınavlarda yeterince yer verildi.					
4	Okuma (reading) derslerinin içeriğine sınavlarda yeterince yer verildi.					
5	Konuşma (speaking) derslerinin içeriğine sınavlarda yeterince yer verildi.					
6	Video derslerinin içeriğine sınavlarda yeterince yer verildi.					
7	Derslerde işlenen dilbilgisi (grammar) konularına sınavlarda yeterince yer verildi.					
8	Derslerde öğretilen kelimelere sınavlarda yeterince yer verildi.					
9	Derslerde yapılan dinleme (listening) çalışmalarına sınavlarda yeterince yer verildi.					
10	Laboratuar derslerinin içeriğine sınavlarda yeterince yer verildi.					
11	Derslerde yapılan alıştırmalara sınavlarda yeterince yer verildi.					
12	Genelde, derslerin içeriklerine sınavlarda yeterince yer verildi.					

2.Bölüm ***Lütfen, sizin görüşünüzü en iyi biçimde yansıtan kutuyu (√) şeklinde işaretleyiniz, ve lütfen her bir ifade için yalnızca bir cevap seçiniz..***

No	Öğrencilerin, kendi performansları bakımından sınavların güvenilirliği konusundaki görüşlerine ilişkin sorular	Kesinlikle Katılıyorum	Katılıyorum	Kararsızım	Katılmıyorum	Kesinlikle Katılmıyorum
1	Bazen, sınavdaki iki (ya da daha fazla) soru birbiriyle yakından alakalı görünüyordu. Bu nedenle, bir soruyu yapamadıysam diğerini de yapamadım.					
2	Sınavlar çok fazla soru içeriyordu.					
3	Sınavlar yetersiz sayıda soru içeriyordu.					
4	Sınavlarda her bir bölümde ne yapılması gerektiğini açıklayan talimatlar açık ve netti.					
5	Sınavın her bir bölümüne ayrılan puan miktarı sınav kağıtlarında her zaman belirtiliyordu.					
6	Öğrencilere sınavı tamamlamaları için verilen süre sınav kağıtlarında her zaman belirtiliyordu.					
7	Yapılan sınavların öğrencinin nihai (en son) notunu ne derece etkileyeceği her zaman duyuruldu.					
8	Sınavlardaki tüm sorular aynı zorluk derecesindeydi.					
9	Sınav soruları açık ve netti.					
10	Sınav kağıtlarının sayfa düzeni güzeldi.					
11	Sınav kağıtları okunaklıydı.					
12	Sınavlarda kullanılan tablolar açık ve anlaşılırdı.					
13	Öğretmenler bizim sınav formatına alışmamıza yardımcı oldu.					
14	Sınavı tamamlamamız için verilen süre çok kısaydı.					
15	Sınavı tamamlamamız için verilen süre çok uzundu.					
16	Aynı sınava giren tüm sınıflara aynı süre tanındı.					
17	Dikkat dağıtıcı sesler ve gürültüler sınavlardaki performansımı düşürdü.					
18	Sınıflardaki düşük düzeydeki ışık miktarı sınavlardaki performansımı düşürdü.					
19	Sınıflardaki sıcaklık derecesi sınavlardaki performansımı düşürdü.					
20	Sınıflardaki düşük düzeydeki hava miktarı sınavlardaki performansımı düşürdü.					
21	Genelde, sınavların yapısı sınavlarda en iyi performansımı sergilememi engelledi.					
22	Genelde, kötü ortam koşulları sınavlarda en iyi performansımı sergilememi engelledi.					

III. Bölüm

1. Sınavların içeriği hakkında başka yorum ya da önerileriniz var mı? Eğer varsa, lütfen aşağıdaki boşluklara yazınız.

……

……

……

……

……

……

……

……

……

……

2. Yukarıda bahsedilenler dışında, daha önce hiç sınav esnasında performansınızı düşüren bir durumla karşılaştınız mı? Karşılaştıysanız, lütfen durumu anlatınız, ve sizce bu durumu düzeltmek için ne yapılabilir?

……

……

……

……

……

……

……

……

……

……

……

……

IV. Bölüm - Özgeçmiş Bilgileri

Lütfen, sizin için uygun olan seçeneği (√) şeklinde işaretleyiniz.

1. **Daha önce Zonguldak Karaelmas Üniversitesi dışında herhangi bir kurumda İngilizce hazırlık okudunuz mu?**
   Evet ( ) Hayır ( )

2. **Geçen yıl Hazırlık Okulunu başarıyla mı tamamladınız?**
   Evet ( ) Hayır ( )

Yardımlarınızdan dolayı çok teşekkür ederim!!!

APPENDIX F

FORMER STUDENTS' QUESTIONNAIRE (ENGLISH VERSION)

Please answer the following questions considering the exams administered last year (in the 2005-2006 academic year) in Prep School.

Section I

Please put a (√) in the box which reflects your point of view best, and please choose only one answer for each statement.

No	Questions about students' perceptions of the match between course contents and exam contents	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
		SA	A	U	DA	SD
1	The content of the main course book 'Quartet' was represented in the exams sufficiently.					
2	The content of the grammar book 'Milestones of English Grammar-Perfecting and Practicing English Structure' was represented in the exams sufficiently.					
3	The content of the writing courses was represented in the exams sufficiently.					
4	The content of the reading courses was represented in the exams sufficiently.					
5	The content of the speaking courses was represented in the exams sufficiently.					
6	The content of the video courses was represented in the exams sufficiently.					
7	Grammar taught in the courses was represented in the exams sufficiently.					
8	The vocabulary taught in the courses was represented in the exams sufficiently.					
9	The listening practices focused on in the courses were represented in the exams sufficiently.					
10	The content of the laboratory courses was represented in the exams sufficiently.					
11	The exercises made in the courses were represented in the exams sufficiently.					
12	In general, the contents of the courses were represented in the exams sufficiently.					

Section II ***Please put a (√) in the box which reflects your point of view best, and please choose only one answer for each statement.***

No	Questions about students' perceptions of the reliability of tests in terms of their performance	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
		SA	A	U	DA	SD
1	Sometimes, two (or more) questions in the test seemed to be closely related, so that if I couldn't answer one question, I couldn't answer the other question either.					
2	The exams included too many questions.					
3	The exams included an insufficient number of questions.					
4	The instructions explaining what to do in each section in the exams were explicit and clear.					
5	The points allotted for each section of the exam were always stated in the exam papers.					
6	Time given to the students to complete the exam was always stated in the exam papers.					
7	Information about how much the given tests would affect the final grade was always announced.					
8	All the questions in the exams had the same difficulty level.					
9	The exam questions were explicit and clear.					
10	The lay out of the exam papers was fine.					
11	The exam papers were legible.					
12	The tables which were employed in the exams were clear and easy to interpret.					
13	The instructors helped us to get used to the format of the exams.					
14	The time given to complete the exams was too short.					
15	The time given to complete the exams was too long.					
16	Equal timing was given to all classes which took the same test.					
17	Distracting sounds and noises lowered my performance in the exams.					
18	The little amount of light in the classrooms lowered my performance in the exams.					
19	The degree of the temperature in the classrooms lowered my performance in the exams.					
20	The little amount of air in the classrooms lowered my performance in the exams.					
21	In general, the structure of the tests hindered my ability to display my best performance in the exams.					
22	In general, the bad environmental conditions hindered my ability to display my best performance in the exams.					

Section III

1. Do you have other comments or suggestions about the content of the exams? If yes, please transcribe in the blanks given below.

……

……

……

……

……

……

……

……

……

……

2. Have you ever come across a situation which lowered your performance during the exams other than the ones mentioned above? If yes, please describe the situation, and what can be done to improve this situation in your opinion?

……

……

……

……

……

……

……

……

……

……

……

……

Section IV - Background Information

Please, tick (√) the suitable answer for you.

1. **Did you attend English prep class in an institution other than Zonguldak Karaelmas University before?**
   Yes ( ) No ( )

2. **Did you complete prep class successfully last year?**
   Yes ( ) No ( )

Thank you very much for your help!!!

APPENDIX G

INSTRUCTORS' RESPONSES TO THE OPEN-ENDED QUESTIONS

Focus: Q1-Do you have other comments or suggestions about the content of the exams? If yes, please transcribe in the blanks provided.		
Suggestions	**Count**	**%**
Authenticity should be promoted.	4	13.7
A listening and a speaking section should be included in the exams.	3	10.3
Various question types should be included in the exams.	2	6.8
There should be a closer cooperation between the instructors and the testing office members.	1	3.4
The content of the exams should be parallel with the goals of the language teaching program.	1	3.4
Synonyms and antonyms can be asked in the vocabulary section of the exams.	1	3.4
Active (commonly used) verbs should be included in the questions.	1	3.4
Speaking skills should be tested both when the students are prepared and unprepared. Students should be asked to make presentations (prepared). They should also be interviewed by the instructors (unprepared).	1	3.4

Focus: Q2-What promoted scorer reliability in our institution in your opinion? You can list more than one item.		
Comments	**Count**	**%**
Marking the papers twice by two different instructors	12	41.3
Detailed answer keys	10	34.4
Qualified questions which promote objective scoring	8	27.5
Standardization meetings	8	27.5
Each instructor's scoring the papers of the classes other than the classes which they instruct.	3	10.3
Announcements made by the testing office about the changes in the answer keys.	2	6.8
Forming a group of experienced instructors to score the writing section of the exams	1	3.4
Clear rating scales	1	3.4
Enough time to check the exam papers	1	3.4
Experience	1	3.4

Focus: Q3-What hindered scorer reliability in our institution in your opinion? You can list more than one item.		
Comments	**Count**	**%**
The subjective nature of scoring writing skills	7	24.1
Careless and quick marking	4	13.7
Limited time to check the exams	3	10.3
Not having any training in testing or scoring	3	10.3
The high number of quizzes	2	6.8
Not having a detailed scale for writing sections of the exams	1	3.4
Writing office's not making it clear how to evaluate the writing section	1	3.4
The effect of instructors' different educational backgrounds on their scoring writing practices.	1	3.4
Identifying students by name not number	1	3.4
Inadequate answer keys	1	3.4
Questions with more than one answer	1	3.4
Purely sticking to the key	1	3.4

Focus: Q4-Do you have any suggestions to promote scorer reliability within our institution? If yes, please transcribe in the blanks provided.		
Suggestions	**Count**	**%**
Instructors should be trained in testing and scoring.	3	10.3
Importance of scoring should be frequently emphasized by the administrators.	3	10.3
Testing office members should prepare a second key soon after the exams after reviewing the answers of some students. The reason for this is that some answers are unpredictable and hard to score. By this way, testing office members can prepare a more detailed key including the scores suitable for the debated answers given by the students.	2	6.8
Instructors may be led to study on their own on testing and scoring. They can be encouraged to read relevant articles.	2	6.8
More time should be given to the instructors for scoring.	1	3.4
Teaching load should be decreased.	1	3.4
More objective questions should be included in the exams.	1	3.4
More down to earth criteria should be specified for scoring.	1	3.4
Instructors should specialize in certain courses, and they should be assigned the task of scoring the related sections in the exams. For instance: if an instructor has specialized in writing, she or he should be assigned the task of scoring writing section.	1	3.4

APPENDIX H

STUDENTS' RESPONSES TO THE OPEN-ENDED QUESTIONS

Focus: Q1-Do you have other comments or suggestions about the content of the exams? If yes, please transcribe in the blanks given below.		
Comments/Suggestions	**Count**	**%**
A speaking section should be included in the exams.	9	17.3
A listening section should be included in the exams.	7	13.4
The content of the laboratory courses should be incorporated into the exams.	4	7.6
Writing skills should be assessed more frequently.	4	7.6
The exams should be harder.	4	7.6
Sometimes, we came across questions which we had not been instructed about. However, other classes had been instructed about these questions. Such happenings should be prevented.	2	3.8
The exams should be easier.	2	3.8
Distracters in the vocabulary section could be more distractive.	1	1.9
Vocabulary taught in reading courses should be incorporated into the exams.	1	1.9
Questions which are exactly the same as the ones in "Quartet" course book should not be incorporated into the exams.	1	1.9
The questions could be clearer.	1	1.9
More questions should be asked, in order to test all of what students learn.	1	1.9
Reading section should be easier.	1	1.9
The questions used to be asked in groups. Therefore if one of the answers was wrong, others within the group might be wrong as well.	1	1.9

Focus: Q2-Have you ever come across a situation which lowered your performance during the exams other than the ones mentioned above? If yes, please describe the situation.					
Category	**Subcategory**	**Count**	**%**	**General count**	**General %**
Noise	Noise of the students who completed the exam and leaving the classroom	7	13.4	18	34.6
	Instructors' chat among themselves	7	13.4		
	Noise caused by the instructors' who are wandering around	3	5.7		
	Chat between instructors and students	2	3.8		
	Noise caused by the students who are trying to cheat	1	1.9		
	Coughing students	1	1.9		
	High-heeled shoes of female instructors	1	1.9		
	Noise of students who are late for the exams	1	1.9		
Temperature	Having a seat near the window when the weather is sunny	2	3.8	3	5.7
	Cold corridors	1	1.9		
Seating	Having a seat in the corridor	4	7.6	7	13.4
	Having a seat in crowded classes	3	5.7		
	Uncomfortable chairs	2	3.8		
	Chairs with partially broken legs	1	1.9		
	Chairs with low legs	1	1.9		
Psychology	Instructors' standing still very close to the students	2	3.8	4	7.6
	Instructors' giving information about the content of the exams before they are conducted	2	3.8		
	Instructors who are staring at students' exam papers	1	1.9		
	Stress	1	1.9		

Focus: Q2-If you come across a situation which lowered your performance during the exams, what can be done to improve this situation in your opinion?			
Problem	**Solution**	**Count**	**%**
Noise	The instructors should be trained.	2	3.8
	The instructors should be more careful about the students who are trying to cheat.	1	1.9
	The noise in the corridors can be prevented by the instructors who are on duty.	1	1.9
Temperature	The classes must not be too hot or too cold. The instructors should be more interested in and careful about this issue.	1	1.9
Seating	Exams must not be administered in the corridors.	4	7.6
	The physical conditions of the classes in which the exams will be administered should be examined beforehand by the invigilators.	1	1.9
	The classes should be enlarged.	1	1.9
Psychology	Instructors should either inform all the classes about the content of the exams, or none of them. Instructors should not be anxious about the grades of the students who they themselves teach.	1	1.9
	The stress of the students should be decreased with the help of an advisor (instructor) or a psychologist.	1	1.9