REPUBLIC OF TURKEY
ÇAĞ UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

THE ENGLISH PROFICIENCY EXAM IN EFL CONTEXT: A VALIDATION STUDY

THESIS BY

Semiha GÜRSOY

SUPERVISOR

Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

MASTER OF ARTS

MERSİN, April 2013

REPUCLIC OF TURKEY

ÇAĞ UNIVERSITY

DIRECTORSHIP OF THE INSTITUTE OF SOCIAL SCIENCES

We **certify** that thesis under the title of **"THE ENGLISH PROFICIENCY EXAM IN EFL CONTEXT: A VALIDATION STUDY"** is satisfactory for the award of the degree of **Master of Arts** in the Department of **English Language Teaching**.

.....

Supervisor- Head of Examining Committee: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

.....

Member of Examining Committee: Assist. Prof. Dr. Hülya YUMRU

.....

Member of Examining Committee: Assist. Prof. Dr. Erol KAHRAMAN

I certify that this thesis conforms to formal standards of the Institute of Social Sciences.

.....

19 / 04 / 2013

Assoc. Prof. Dr. Haluk KORKMAZYÜREK
Director of Institute of Social Sciences

Note: The uncited usage of the reports, charts, figures and photographs in this thesis, whether original or quoted for mother sources is subject to the Law of Works of Arts and Thought. No: 5846.

ii

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ, whose invaluable scientific guidance, and encouragement made this study possible. With her positive attitude, her inspirations and her great efforts in spending a lot of time to explain things clearly, simply and patiently, she helped to make research become more pleasurable. Has it not been for her faith in me, it wouldn't have been possible for me to finish this study.

Besides my thesis supervisor, I am also grateful for the invaluable insights and recommendations put forth by the other jury members, Assist. Prof. Dr. Hülya YUMRU, Assist. Prof. Dr. Erol KAHRAMAN and Assist. Prof. Dr. Kim Raymond HUMISTON.

I feel grateful to the Director of Preparatory School of Çağ University, Hamdi ÖNAL, for the permission I was given to carry out my study.

I wish to give a heartfelt thanks to Senem ZAİMOĞLU. This thesis would have remained a dream had it not been for Senem. Without her motivation, it would be impossible for this study to come into existence. I will never forget her sincerity and encouragement. She has been my inspiration as I hurdle all the obstacles in the completion of this research.

My sincere and special thanks go to my colleagues and great friends, Laura Ellen ŞAKIRGİL and Can ŞAKIRGİL who participated in the study and spent their valuable time to help me in this study.

I want to take the opportunity to thank Burcu ÖZER for sharing my responsibilities on the job and her constant support and good fellowship.

I cannot finish without saying how grateful I am for my family. My parents Nilgün and Necati KAHYALAR have raised me to always strive for knowledge, progress continuously and become a better person. They have always offered their endless support and constant prayers. They have sacrificed so much for me and my siblings to have a good life, receive good education and to have a loving and safe home. My family has been a great blessing and support for me throughout my life and especially during my hardest times.

I would like to extent my deepest thanks to my lovely twin sisters, Selin and Pelin KAHYALAR who have always been a great support to me all through my life. They were always with me whenever I needed them and were willing to help with any kind of matter.

Lastly, and most importantly, I cannot find any suitable words to express enough gratitude for my husband Emre GÜRSOY. I love him endlessly. He always helped and trusted me. He encouraged me with his constant support and companionship when I felt so desperate. This academic journey would not have been possible without his love, patience, and sacrifices along the way. Thank you very much.

April 19, 2013

Semiha GÜRSOY

ÖZET

YABANCI DİL İNGİLİZCE YETERLİLİK SINAVI: BİR GEÇERLİLİK ÇALIŞMASI

Semiha GÜRSOY

Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı
Tez Danışmanı: Doç. Dr. Şehnaz ŞAHİNKARAKAŞ
Nisan 2013, 62 Sayfa

İngilizce yeterlliik sınavının amacı öğrencilerin lisans programındaki öğretimi ve akademik çalışmaları başarıyla takip edecek düzeyde dil ve akademik becerilere sahip olup olmadıklarını ölçmektir. Bu çalışmada Çağ Üniversitesi Hazırlık Okulu İngilizce Yeterlilik Sınavı'nın geçerliliğinin ve güvenirliliğinin değerlendirilip ölçülmesi amaçlanmıştır. Bu hedefle, Bachman (1990) tarafından düzenlenilen Dil Becerileri İçerik Yapı Modeli kullanılmıştır. Araştırmanın örneklemini 2011-2012 akademik yılında hazırlık okulunda öğrenim gören 133 öğrencinin yeterlilik sınav sonuçları oluşturmaktadır.

Veriler öğrencilerin 2011-2012 akademik yılı başında girdikleri Yeterlik Sınavı sonuçları yoluyla toplanmıştır. Toplanan veriler tanımlı istatistik yöntemleri ve içerik analiz modeli kullanılarak analiz edilmiştir. İngilizce Yeterlik Sınavı bölümleri arasında istatiksel olarak önemli farklar olup olmadığını ve testin güvenilirliğini incelemek için Cronbach Alfa katsayısı hesaplanmış ve her bir bölümün istatiksel verileri için tanımlı istatistik verileri toplanmıştır.

Çalışmanın sonuçları Çağ Üniversitesi Hazırlık Okulu İngilizce Yeterlik Sınavının amacına kısmen ulaştığını göstermiştir. Çalışmada elde edilen bulgular sınavın daha geçerli olması için içerik, kaynak ve değerlendirme boyutlarında geliştirmeye gerek olduğunu ortaya koymuştur.

Anahtar Kelimeler: Geçerlilik, Güvenilirlik, Dil becerileri, İngilizce Yeterlilik Sınavı, Yeterlilik

ABSTRACT

THE ENGLISH PROFICIENCY EXAM IN EFL CONTEXT: A VALIDATION STUDY

Semiha GÜRSOY

MA Thesis, Department of English Language Teaching
Supervisor: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ
April 2013, 62 Pages

The aim of the English Proficiency Exam at Çağ University is to assess whether the students have the ability and skills to follow the education program and academic studies in the undergraduate program.

The purpose of this study was to investigate and study the validity and the reliability of the English Proficiency Exam conducted at Çağ University Preparatory School. To this end, the Model for language ability by Bachman (1990) was utilized. 133 students' proficiency exam results in the 2011-2012 academic year were taken as samples in the study.

The data were gathered through the exam results of the students who had taken the exam English Proficiency Exam in the related term. The data were analyzed through descriptive statistics and content analysis. Cronbach's Alpha co efficiency was calculated to investigate the reliability of the exam and descriptive statistics were calculated for each components of the English Proficiency Exam.

Results of the study indicated that the English Proficiency Exam partially served for its purpose. The findings revealed that some improvements in the content, materials and assessment dimensions of the program were required to make the Exam more valid and reliable.

Key Words: Validity, Reliability, Language Skills, English Proficiency Exam, Proficiency

LIST OF TABLES

LIST OF ABBREVIATIONS

EPE : English Proficiency Exam

FCE : First Certificate in English Examination

CPE : Certificate of Proficiency Exam

TOEFL : Test of English as a Foreign Language

SPSS : Statistical Package for Social Sciences

TABLE OF CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 1

1. INTRODUCTION

In this chapter, the importance of and need for a validity study of the English Proficiency Exam (EPE) of the Preparatory School at Çağ University was outlined. The need for gathering validity-related evidence to help vindicate the effectiveness of a given test was explained. The background of the study, statement of the problem, purpose, significance, and some important definitions used in the study were mentioned for the purpose of giving a general idea about the structure of the thesis.

1.1. Background of the Study

Foreign language testing has been a very important issue for many hundred years as it is a complementary part of the teaching-learning process. Language tests can be a valuable tool for providing relevant information regarding several concerns in language teaching and they also can provide evidence of the results of learning and instruction and furthermore give feedback on the effectiveness of the teaching program itself (Bachman & Palmer, 1996). Because of this fact, even today we still critically discuss the question of what and how we are testing in the field of education. Any language test should consider validity and reliability as the basic criteria. Validity is a relatively well-researched topic in the study of language testing, attracting considerable attention by language test designers and researchers in recent decades (Walt & Steyn, 2008). Reliability is also the focus of much research and is described as the consistency of measurement. This means a reliable test score will be consistent among different characteristics of the testing situation.

1.1.1. Bachman's Model of Communicative Language Ability

Every test starts with some abstract belief theories like what language is, what proficiency level consists of, what language learning involves and what language users do with languages (Alderson, Clapham and Wall, 1995). Each of these theories has constructs which are its principal components and the relationship between these components. Construct validation is used to assess how well a test measures the construct. Then, for the purpose of validation, Alderson, Clapham and Wall (1995) state that "test specifications need to make the theoretical framework which underlies the test explicit and to spell out relationships among its constructs, as well as the relationship between the theory and the purpose for which the test is designed" (p. 17).

Bachman's Model is one of these theoretical frameworks which were developed for the purpose of test analysis and it was an extension of earlier models "in that it attempts to characterize the processes by which the various components interact with each other and with the context in which language use occurs" (Bachman 1990:81).

Although it is important to be aware of the full range of components of language ability as language tests are designed and developed and language test scores are interpreted by us, as teachers, many of the language tests we develop will focus on only one or a few of these areas of language knowledge. "We believe, therefore, that the design of every language test, no matter how narrow its focus, should be informed by a broad view of language ability" (Bachman & Palmer, 1996, p.67). For that reason, it can be used as a part of the validation study.

Bachman's Model includes two main sections which are *organizational knowledge* and *pragmatic knowledge*. Each of these sections also consists of a number of components such as, *grammatical knowledge* and *textual knowledge* which are two areas of organizational knowledge, and *functional* and *sociolinguistic knowledge* which are two areas of pragmatic knowledge.

Grammatical competence is the knowledge of grammar and vocabulary at a sentence level. It enables the building and recognition of well-formed, grammatically accurate utterances, according to the rules of syntax, semantics, morphology, and phonology /graphology.

Textual competence is the knowledge and application of cohesion and coherence rules and devices in building larger texts/discourse. It enables the connection of utterances and sentences into cohesive, logical and functionally coherent texts and/or discourse.

Functional competence is competence to convey and interpret communicative intent or function behind a sentence, utterance or text. It encompasses macro-functions of language use such as, transmission of information, social interaction and getting things done/persuading others, learning and thinking, creation and enjoyment and micro-functions, or speech acts like requests, threats, warnings, pleas, etc., and the conventions of use.

Socio-cultural competence focuses on appropriateness in producing and understanding utterances. These include rules of politeness; sensitivity to register, dialect or variety; norms of stylistic appropriateness; sensitivity to "naturalness"; knowledge of idioms and figurative language; knowledge of culture, custom and institutions; knowledge of cultural references; and uses of language through interactional skills to establish and maintain social relationships.

As it was mentioned before, the medium of instruction at Çağ University is English.

For that reason the selection of competencies and tasks in the English Proficiency Exam has to do with how useful and important they are in real communication situations and tasks so that the students could encounter in educational contexts, moreover, in the community and on their future jobs.

1.2. Statement of the Problem

Çağ University strains at consistently improving the learning of its students. It was observed that a serious institutional commitment to lifelong learning has extensive implications for how we teach our students. According to this view, we focused more on what our students learn than what we teach. This can be a challenging paradigm-shifting innovation for most people, who sometimes relish merely the *sage-on-the-stage* model of teaching. However, it is a paradigm shift for over a decade that has radically recast the nature of higher education. To fulfill this paradigm shift, Çağ University has requested that Preparatory School (1) identify and publish expected learning outcomes which involves to prepare the students for studies in their departments, providing them the necessary English skills required for their higher education at international standards and also to make a significant contribution to our students in their journeys in this global world; (2) demonstrate that the students who complete their programs have achieved the stated outcomes; and (3) provide evidence consistently across its programs that its assessment activities lead to improvement of teaching and learning.

The Preparatory School of Çağ University followed this request and has identified and published its expected learning outcomes. Furthermore, in order to demonstrate that students have achieved these stated outcomes, the Preparatory School has established direct and indirect measurement methods, or assessment activities, to show evidence of learning. One of the direct measurement tools is the English Proficiency Exam (EPE), which nearly all students are required to take, as the medium of instruction at Çağ University is English. The purpose of the English Proficiency Exam is to determine how well students perform in the skill areas of listening, reading, writing, grammar and vocabulary. Then there is a need for the validation study of the English Proficiency Exam at Çağ University to search for the effectiveness of the teaching program.

1.3. Purpose of the Study

This study gathers validity-related evidence to help answer some questions concerning the validity of the EPE at Çağ University. The Preparatory School is interested in improving

the quality of the EPE, and this study collects qualitative evidence and quantitative data from test scores, analyzes them, and attempts to interpret the quantitative analysis to make suggestions for improvement, which will be later used to positively influence the teaching and assessment process for future students of Çağ University.

1.4. Significance/Need for the Study

This research study outlines an initial investigation into the validity of the EPE as the university and the staffs of the department want to know whether the EPE measures what it is supposed to measure. In particular, the aim was to examine the validity and reliability of the EPE. It is one of the missions of Çağ University to lead in social developments and to raise individuals who are equipped with modern knowledge, and can follow national and international event. To realize this mission, Çağ University provides education at world's standards and has innovative academic programs. Scientific freedom and responsibility are its most important values, as well as fairness and equity (Çağ University, 2010: 1). The purpose of this research is to answer the question: To what extent is the English Proficiency Exam valid and reliable? It might be open to question whether the test scores are a straight representation of a student's level of language knowledge or skills if a test is not valid. Moreover, the decisions that are made on the basis of these scores are founded on unreliable grounds.

1.5. Research Questions

Based upon the aim of this study, the following questions are established to provide guidance for collecting applicable evidence:

1-How valid is the English Proficiency Exam of Çağ University's Preparotory School?

2- How reliable is each component of the English Proficiency Exam of Çağ University's Proficiency Exam?

3- Is there sufficient variation in test scores in each component of the English Proficiency Exam of Çağ University Preparotory School?

4- Are there any potential item problems in any components of the English Proficiency Exam concerning (excluding writing);

a) item difficulty
b) item discrimination

1.6. Limitations

This study focuses on answering specific questions regarding the validity of the EPE scores. Due to the limited time frame, an extensive and overall validation study lies outside the scope of this research study. Thus the limitations of this study are as follows:

This research does not analyze whether the content covers all aspects of each skill area though it would be helpful in this sort of validation study. The aim of this study is not to investigate whether the tasks or items of each skill area cover the full construct of each skill area. The researcher will not consider the conditions under which the EPE is administered to students. This study does not address the administration of the exam.

1.7. Definitions of the Terms

Language proficiency is a term which has always been used in the language testing field and it is used in two different ways here. Firstly, it is related directly to ability and is defined as the degree of competence or capability in a given language demonstrated by an individual in a given point of time independent of a specific course or textbooks or teaching methods. Secondly, it relates to the extent and adequacy of an individual's control or mastery of target language in all kinds of social interactive or situations including work settings as demonstrated in tests. The former meaning of proficiency denotes competence whereas the later one specifies performance (Giri, 2002).

*Rating scale (*also proficiency scale): A scale consisting of several categories used for making judgements of performance. The levels of rating (proficiency) scale are usually explained by what subjects can do with the language and their ability in the various language skills and features (Starr, 2008).

Validity is the degree to which a test measures what it claims to be measuring.

Construct Validity is the extent to which a test measures the concept or construct that it is intended to measure.

CHAPTER 2

2. REVIEW OF LITERATURE

The aim of Chapter Two is to give a theoretical basis to carry out such a research study. This chapter represents and describes the role of testing in educational programs and different test designs and their functions are also discussed. The need for the validity study and the details for it are described. Lastly, the connection between validity and reliability is appointed, reliability is explained and the importance of it is defined.

2.1. The Role and Use of Testing in Language Program

In chapter one, it was explained that Çağ University requests in its policy that Preparatory School (1) identify and publish expected learning outcomes; (2) demonstrate that the students who complete their programs have achieved the stated outcomes; and (3) provide evidence consistently across its programs that its assessment activities lead to improvement of teaching and learning, not only to make an explicit judgement about the worth of program to determine if the standards have been met but especially to achieve the improvement of learning at Çağ University. The third item obviously requests evidence for the validity of tests. As we are in the era of accountability the demand for language program is on the rise (Suvedi, 2002). For this reason, we need to understand the process of language program and the necessity of it before explaining the meaning of validity.

We should be aware of the importance of testing in language programs and various test designs which can be used in this period. Additionally, testing should be considered to be tied up to all the other parts of the program rather than isolated part in the process of teaching and learning (Starr, 2008).

As it was stated in Çağ University's accreditation committee report (Çağ, 2010:1), it was its mission to provide education at world's standards. Scientific freedom and responsibility are also its most important values as well as fairness and equity. It is vital for any program to have a continuous evaluation no matter how large or small to assure that their teaching and learning tasks have been achieved. While the main purpose of evaluation is to identify the strengths and weaknesses of a program or a project to improve the quality of it, this is not the only one. It is also providing evidences to show that there is an effective learning process. Suvedi (2002) states that "when we evaluate we collect information about a program's actual inputs and/or outcomes and then compare that info to some preset standards or expectations and a judgement is made about the program or activity" (p.2).

Evaluation has been defined with various different definitions. One typical dictionary definition of evaluation is "to find or state amount or value of; appraise, asses, to examine and judge" (Swannel, 1988). Brown (1989) defines evaluation as "the systematic collection and analysis of all relevant information necessary to promote the improvement of curriculum and asses its effectiveness and efficiency as well as participants' attitudes within a context of particular institutions involved" (p.223). This definition shows that collecting relevant information is not enough alone. It must be done systematically and analyzed throughly to make decisions and/or determine the effectiveness and efficiency of programs and projects or any element of curriculum in general to improve them.

Since curriculum evaluation is important in the education process, different approaches to language program have emerged which, according to Brown (1989), generally falls into four categories, which are *product oriented approaches, static characteristic approaches, process oriented approaches* and *decision facilitation approaches.*

In *product oriented approaches* the focus is on whether the instructional objectives and goals of a program have been achieved or not (Brown, 1989). For this reason there should be clearly defined goals and measurable behavioral characteristics of program. Starr (2008) exemplifies them as "students, the subject matter, societal consideration, philosophy of education and learning philosophy.

Statistic characteristic approach is called as "professional judgement" evaluation according to Worthen and Sanders (1973, cited in Brown, 1989). The aim of this approach is to determine the effectiveness of a particular program which is conducted by outside experts. The institutions should provide all the records related to the effectiveness of the program and also demonstrate the adequacy of the physical learning facilities for such an evaluation. The expert group conducting the evaluation formulates a report according to their observations after assessing the quality of the program in detail, based on the information described above.

Process- oriented approach began with the realization of the importance of changing and improving the curriculum by the help of evaluation process as well as the importance of achieving program objectives (Brown, 1989). Brown (1989) claims some of the most important foci of program evaluations as:

> (1) the distinction between formative and summative evaluation; (2) the importance of evaluating not only whether the goals have been met but also whether the goals themselves are worthwhile; (3) goal free evaluation". For example the evaluators should not only limit themselves to studying the expected goals of the program, but

also consider the possibility that there were unexpected outcomes which should be recognized and studied (Brown, 1989:226).

In *decision facilitation approach* program evaluation serves for those who make judgements and decisions for the program. These are usually the administrators. Starr (2008) says that, "information is collected to help make decisions about the state of the overall system, program planning, program implementation, program improvement, and the overall value of the program" (p:10). According to Brown (1989), evaluation is a continuing process and it should maintain information useful to decision makers.

Language programs carrying out an evaluation generally make use of most or all of these approaches. Assessment is the process of a program evaluation in which students are evaluated whether they are learning or not. It is the systematic collection, review and use of information to enhance students' learning and improvement. By the help of different kind of measures, students are assessed to identify whether or not they are achieving learning goals that have been determined by their faculties for their courses and programs. On the other hand, assessment results provide qualitative information which helps the faculties to decide on how to improve courses or programs through changes in curriculum, teaching materials and things like that. If it is integrated in the planning cycle for curriculum development and review, assessment results can supply a powerful rationale for securing support for curricular and other changes. Meanwhile, it may provide comparative data which can give important and useful information about the students' performance on how well they are meeting the learning outcomes of the program or course, or it may also show their actual performance compared with those at other similar institutions. The evaluation occurs in different dimensions (Brown, 1989) according to the circumstances of the program and the type of decisions which are needed to be made. These dimensions are so connected with each other and they are comprised of two perspectives. Each of these perspectives should be considered in an evaluation as they may have important information.

Formative and *Summative Perspective* is the first dimension and it has an impact on information and on the types of decisions that will finally blossom out from each purpose. The aim of a formative evaluation is "evaluating students in the process of forming their competencies and skills with the goal of helping them to continue that growth process" (Brown, 2004:6). The purpose here is to enhance the teaching and learning process of a program and the gathered information gives insights in the results of the program, its strengths and weaknesses.

Summative evaluation typically occurs at the end of a course or unit of instruction and

the aim is to measure, or summarize what a student has learnt (Brown, 2004). The purpose of this type of evaluation is to decide if the program is effective and successful and whether a new curriculum is needed. The distinction between *product* and *process*, as the second dimension of the two views, is based on differences in what information might be considered.

The focus of the product evaluation is to find out whether the goals (product) of the program are achieved or not. "Product and summative evaluations both tend to focus on product because the purpose is to make decisions about whether or not the goals of the program have been achieved" (Muşlu, 2007:11).In *process-oriented evaluation*, the focus is more on how the program (process), which helps to arrive at these goals (product), keeps on. Formative evaluations generally deal with process as the aim is to find whether or not the goals have been met and to study and improve those processes that were involved (Brown, 1995, cited in Muşlu, 2007).

2.2. Test Designs

The purpose of the test should be determined before designing it. Defining the purpose of the test will help to choose the right kind of test, and it will also help to focus on the specific objectives of the test (Brown, 2004). Tests can be classified according to the types of information they provide. "This categorization will prove useful in deciding whether an existing test is suitable for a particular purpose or not" (Hughes, 2003:11). The four types of test are: *achievement tests, diagnostic tests, placement tests and proficiency tests.*

Achievement tests are related to the classroom lesson, units or the syllabus and "the primary role of an achievement test is to determine whether course objectives have been met" (Brown, 2004:48). Hughes (2003) says that, some testers have the view of basing the content of achievement tests on a detailed course syllabus or on the books or other materials used. He adds that although it has an obvious appeal and can be considered a fair test, the disadvantage is that "if the syllabus is badly designed, or the books and other materials are badly chosen, the results of a test can be very misleading" (2003:13). The alternative idea for that is to base the content of achievement tests on course objectives instead of the detailed content of a course. Then, it will have many advantages one of which is that the test can reveal how far the students have achieved those objectives and in a roundabout way, this will oblige course designers to be clear about objectives (Hughes, 2003).

Diagnostic tests are used, as it can be understood from the name, to diagnose learners' strengths and weaknesses. Such tests intend to show what learning style needs to take place and also gives ideas about learners' language ability. In other words, they show what they

know or don't know about a language and/or whether they can master the language skills. Such tests should give information on what students need to study on in the future (Brown, 2004).

Placement tests have the purpose of placing students correctly at the stage of the teaching program which is ideal for their abilities. Hughes (2003:16) states that "placement tests are typically used to assign students to classes at different levels". According to Brown (2004) a placement test generally involves sampling of the material which is needed to be covered in the various courses in a syllabus. He also adds that the result of the test should reveal that the material is challenging enough for the students which is, neither too easy nor too difficult.

Proficiency tests on the other hand, are used to measure people's ability in language regardless of any training they could have had in the target language. According to Hughes (2003), the content of a proficiency test is based on a specification of what candidates have to be able to do in the language so as to be considered proficient rather than on the content or objectives of language courses. Brown (2004) defines that, "a proficiency test is not limited to any one course, curriculum, or single skill in the language: rather, it tests overall ability" (p: 44). For example, a test can be designed to see whether a student's English is good enough to have a course of study at a British university. Another example might be to find out if someone has a good performance of speaking the language in a business setting. However, there are other proficiency tests which have nothing to do with any occupation or course of study in mind. These kinds of tests can be based on a more general concept of proficiency. Cambridge First Certificate in English Examination (FCE) and the Cambridge Certificate of Proficiency in English Examination (CPE) are British examples and the aim of such tests is to show if candidates have reached a certain standard with respect to a set of specified abilities (Hughes, 2003). Brown makes further explanation by saying;

> A key issue in testing proficiency is how the constructs of language ability are specified. The tasks that test-takers are required to perform must be legitimate samples of English language use in a defined context. Creating these tasks and validating them with research is a time-consuming and costly process. Language teachers would be wise not to create an overall proficiency test on their own. A far more practical method is to choose one of a number of commercially available proficiency tests. (2004:45)

2.3. Norm-referenced vs. Criterion-referenced Testing

Norm-referenced and criterion-referenced tests are frame of reference which is

necessary to clarify test scores. They differ in their purposes, the way that the content selected and the scoring which defines interpretation of test results.

In norm-referenced tests, the aim is to classify students. For this reason, the scores are interpreted through mean, median, standard deviation and percentile rank to "place test-takers along a mathematical continuum in rank order" (Brown, 2004:7). The scores of such kind of tests are reported back by numerical score (like 230 out of 300) and percentile rank (such as 80 per cent). To put it in another way, if we administer a reading test to an individual student and want to know about his/her performance on the test, we can get two different answers, one of which might be that the student got a score which placed him/her in the top 10 per cent of students who have taken the test. The second answer might be that he/she did better than 60 per cent of the students who took the test (Hughes, 2003). Those kinds of tests which are designed to give such information are called norm-referenced tests and they would help teachers to choose students for different ability level, for example reading or mathematics instructional groups (Bond, 1996). Brown (2004) says that standardized tests are typical of norm-referenced tests, such as Test of English as a Foreign Language (TOEFL) and they are prepared to be administered to large groups and their results are reported back efficiently (2004). He also adds that these kinds of tests should have pre-determined, fixed responses which will make scoring easier and quicker at a very low expense. Hughes (2003) further explains that norm-referenced tests show the relation between one candidate's performance with the others rather than showing what the student is able to do in the language directly.

The information interpreted through a norm-referenced test might be useful and important to make a decision about whether the test takers need more assistance or is a candidate of a gifted program. However, this creates a major weakness of norm-referenced tests as the scores don't give much or enough information about students' actual knowledge or what they can do. Bond (1996) clarifies this in her article saying, "the validity of the score in these decision processes depends on whether or not the content of the norm-referenced test matches the knowledge and skills expected of the students in that particular school system" (p.3).

While norm-referenced tests reveal the rank of students, criterion –referenced tests on the other hand determines "what test takers can do and what they know, not how they compare to others" (Anastasi, 1988 as quoted in Bond, 1996:2). The aim of such kind of tests is to report how well a student has learned the pre-determined knowledge and skills and whether they can perform a task or set of tasks satisfactorily (Hughes, 2003). In contrast to norm-referenced tests, the principle of criterion-referenced tests is that the students, who

perform the pre-determined tasks satisfactorily, pass and those who don't, fail. These kinds of tests do not have anything to do with whether all of these students are successful or none of them is successful. The National Center for Fair and Open Testing (2007) gives an example to criterion-referenced tests saying; multiple choice tests and on-the-road driving tests to get a driver's license. Everyone can pass these exams if they can drive well enough and have the knowledge about driving rules. The advantage of these tests, then that "students are encouraged to measure their progress in relation to meaningful criteria without feeling that, because they are less able than most of their fellows" (Hughes, 2003:21).

One of the two benefits of criterion-referenced tests is that the motivation of the students to reach the standards, and to set meaningful standards in the way of what people are able to do, which doesn't change with different groups. Bond (1996:3) further explains that; "as long as the content of the test matches the content that is considered important to learn, the criterion-referenced test gives the student, the teacher, and the parent more information about how much of the valued content has been learned than a norm-referenced test".

The main difference between tests and other components, such as teaching materials and learning activities of an instructional program, is in their purpose, as the primary purpose of other components is to promote learning while the primary purpose of tests is to measure (Bachman & Palmer, 1996). The intended use of a test is the most important concern in designing and developing a language test. When we make a decision about the most appropriate test method for the specific test situation, knowing the different types of tests, the information which scores provide and understanding how we can use them for different circumstances can be really helpful for us. For this reason, it can be said that the usefulness of a test is the most significant quality. Bachman and Palmer (1996) believe that "test usefulness provides a kind of metric by which we can evaluate not only the tests that we develop and use, but also all aspects of test development and use" (p.17). So, the idea of ensuring the quality of information tests give should also be kept in mind as well as being aware of the kind of information they provide. Decisions which are made in the educational field affect people's life in one way or another. For this reason, validity and reliability of these decisions and of course of these tests come forward. Reliability and validity are two of the critical and essential measurement qualities for tests. By the help of these qualities, the major justification for using test scores to make inferences or decisions are provided.

2.4. Validity and Reliability of Test Scores

2.4.1. Test Score Validity

Validity is the main concept of testing and assessment. In our everyday lives, we usually observe the things, behaviors, actions before reflecting any validity decisions and make inferences which lead to action or beliefs. However, we don't ask formal questions or make a list of evidences on validity while doing this. In language testing, on the other hand, this should be our priority to do so, so that we can produce a chain of reasoning and evidence from what we think a test score means, and "the actions we intend to take on the basis of that inference, back to the skills, abilities or knowledge that any given test taker may have" (Fulcher and Davidson, 2007:3).

To avoid making serious mistakes in language testing, we should dispel doubts by removing as much uncertain things as possible so that the scores can reveal the points we want to measure and have a meaning. Validity is the extent to which a test represents or assesses the specific concept that the researcher aims to measure. A test should be valid to apply and interpret the results precisely. Thus we can describe the test validity as the characteristics of a test when it is applied to a particular population.

The definition of validity affects all language test users because accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation (Chapelle, 1999). Lado (1961:321) asked the question "Does a test measure what it is supposed to measure?" to define the validity and said, "If it does, it is valid." Kelly (1927:14) explained the issue of validity as, "the problem of validity is that of whether a test really measures what it purports to measure" (quoted in Weir, 2005:12). However, proving that a test is valid is not enough alone. It should also be proved that "we are measuring what we think we are measuring" (Sireci, 2007:477). For this reason, presenting relevant evidence to defend such use is involved in supporting the use of a test for a particular purpose and it should be kept in mind that, what is to be validated is the use of a test for a particular purpose, not the test itself (Sireci, 2007). More recent writings on validity theory emphasize the importance of viewing validity as a unitary concept. Messick (1989b) gave the definition of validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). As it was mentioned before, decisions which are made in the educational field affect people's life in one way or another. In her thesis, Starr (2008) gave an example to show the importance of decision making using the test scores as:

> It would be very tragic if a hospital hired a surgeon on the basis of high scores received on a set of required exams, but the surgeon turned out to be incapable of performing safe surgeries. The social consequences might include the death of a patient, which would have far-reaching consequences for the patient's family, the surgeon's family and the hospital. Or, for example, if a school district hires a foreign language teacher on the basis of high scores on a language proficiency test, the school district expects the teacher to be proficient in the language skills. If the teacher couldn't speak, understand, read or write the language well, he or she would not be an effective teacher. (p.22)

The intended measurement of the validity between the test and the behavior is detected by a number of research rather than a single statistics. Sireci (2007) stated that;

- Validity is not a property of a test. Rather, it refers to the use of a test for a particular purpose.
- To evaluate the utility and appropriateness of a test for a particular purpose requires multiple source of evidence.
- If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose.
- Evaluating test validity is not a static, one-time event; it is a continuous process (p.478).

Validity doesn't exist naturally in a test. Validity must be established for each particular use of a test. Every test use involves inferences or interpretation; therefore, all validation requires the combination of logical argument and empirical evidence needed to support those inferences (Shepard, 1993). The evidence that might be collected to support the interpretation of the test score includes construct validity, content validity, criterion-related validity, and reliability.

The construct validity of a test is undermined when it is claimed that the test can be used for a variety of purposes: by individuals to assess their English Proficiency; for schools to conduct entrance, placement, or graduation tests; and conglomerates to recruit elites and promote employees (Shih, 2008). There is no test which can definitely meet all of these ends, as Bachman and Palmer (1996) noted, "misconception about testing is, there is one 'best' test for any given situation" (p.7). Construct validity evidence is the extent to which the test measures the right psychological traits such as intelligence, self-esteem and creativity

(Brualdi, 1999). In the assessment field, construct validity tries to answer the question "does this test actually tap into the theoretical construct as it has been defined?" (Brown, 2004:25).

Content validity evidence is about the content of the test and it focuses on whether the test questions stand for the skills in the specified subject area. It can be said that, if a test measures knowledge of the content domain of which was prepared to measure knowledge, then it has content validity. In other words, content validity deals with whether the test items sample the content area to be measured representatively and adequately. For example, a grammar test must be designed to measure the knowledge or control of grammar. Otherwise, it would lack content validity if good scores depended primarily on knowledge of history or maths, or if it only had questions about one aspect of grammar (e.g., tenses) as "areas that are not tested are likely to become areas ignored in teaching and learning" (Hughes, 2003:27). For this reason, we should try to find the best answer to the question, what is important to test instead of what is easy to test. To achieve this Hughes (2003) states that the study of content validation should be done at the same time with the development of the test itself; it is not useful and a good idea to wait until the test is already being used.

Anastasi (1988:132, in Weir, 2005:19) outlined the following items to establish content validity:

1- The behavior domain to be tested must be systematically analyzed to make certain that all major aspects are covered by the test items, and in the correct proportions;
2- The domain under consideration should be fully described in advance, rather than being defined after the test has been prepared;
3- Content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content.

Criterion validity on the other hand, looks to reveal that test scores are systematically related to one or more outcome criteria such as grades, class rank, other tests and teacher ratings compared to performance on the test (Brualdi, 1999). Criterion validity tries to measure how well a person has learned a specific body of knowledge and skills. Most of the classroom based-assessments with teacher designed tests are example for the criterion-referenced assessment. Multiple choice tests designed to get a driving license can be another example to criterion-referenced testing as everybody who knows enough about driving rules can pass the test. In the education field, criterion-referenced tests are prepared to find out how much a student has learned the material taught in a specific course. These kinds of tests don't include, for example, phonetic questions if the aim of the test is to measure the grammatical

knowledge of the students in a specific area. The students who have taken this grammar class could pass the test if they have been taught well enough and if they studied properly and of course, if the test was designed well.

2.4.2. Messick's Validity Model

Although validity is a unified concept and it was traditionally seen as being composed of three separate types which are content, criterion, and construct validity, "the content of a test should be a representation of the construct interpretation and cannot carry on any test purpose on its own" (Starr, 2008:24). Construct validity is a unified concept as criterion-related validity is also needed to be based on construct-related evidence. These different types of validity are different complementary components for the validity evidence which together makes up to what extent the test is valid rather than three separate types of construct (Mesick, 1989a, 1989b). Mesick's model for validity has been a widely cited "progressive matrix" which explained validity and the process of validation (Chapelle, 1999). Messick (1989a) explains this framework, shown in Table 1, by distinguishing two interconnected facets of the unitary concept. He further said that:

> One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or the outcome of the testing, being either interpretation or use. If the facet for source of justification (that is either an evidential basis or consequential basis) is crossed with the function or outcome of the testing (that is, either test interpretation or test use), we obtain fourfold classification. (p.20)

Table 1. ***Facets of Validity***

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance and Utility(R\U)
Consequential Basis	CV + Value Implications	CV+R\U+VI+Social Consequences

Taken from "Validity" by Messick, 1989a, p.20. In R.L. Linn *Educational Measurement*.

Messick (1989) presented his unified but multifaceted validity framework in Table 1, which included the evidential and consequential bases of test interpretation and use. The evidential basis for validity includes both test score interpretation and test score use. The evidential basis for interpreting tests includes the empirical study of construct validity. The evidential basis for using tests includes both construct validity and relevance/utility, which are defined as the theoretical contexts of implied applicability and usefulness.

The consequential basis of validity involves both test score interpretation and test score use too. It's required to make judgements of the value implications in the consequential basis to interpret tests, "which are defined as the contexts of implied relationships to good/bad, desirable/undesirable, etc. score interpretations" (Brown, 2000:9). The consequential basis for using tests includes construct validity, relevance/utility, value implications and making judgments of social consequences, "which are defined as the value contexts of implied consequences of test use and the tangible effects of actually applying that test" (Brown, 2000:9).

As we can clearly understand from the figure, construct validity should be kept in mind as a super ordinate concept containing each forms of validity. On the whole, Messick (1995) puts forward the importance of construct validity as follows:

> The entire progressive matrix represents construct validity, which is another way of saying that validity is a unified concept. One implication of this progressive-matrix formulation is that both meaning and values as well as both test interpretation and test use, is intertwined in the validation process. Thus, validity and values are one imperative, not two, and test validation implicates both the science and the ethics of assessment, which is why validity has force as a social value. (p.749)

2.4.3. Test Score Reliability

Reliability and validity are often discomposed. Jones (2001, as quoted in Weir, 2005:22) thinks that everyday meaning of reliability adds powerful positive connotations to its technical meaning in testing. He further explains reliability as "a highly desirable quality in a friend, a car or a railway system. Reliability in testing also denotes dependability in the sense that, a reliable test can be depended on to produce very similar results in repeated uses" (2001:1). In general, reliability has to do with the consistency of test scores. Reliability analysis is usually considered initiative in the test validation process, as there is no need to spend time investigating the validity of a test if it isn't reliable. A test can be reliable but not valid, whereas a test cannot be valid yet unreliable (Shuttleworth, 2009, Alderson, Clapham

and Wall, 1995). Reliability refers to the consistency of measurement of scores across different evaluators in different time periods. Bachman and Palmer (1996) say that reliability is clearly an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure. When we administer two tests covering similar material, we prefer students' scores be similar as the more comparable the scores are, the more reliable the test scores are (Wells & Wollack, 2003).

To make it more clear, given that one Tuesday afternoon at two o'clock, one hundred students take a 100-item test which is neither impossibly difficult nor too easy for them, so they don't all get zero or a full, 100. Now, imagine that these students had taken this same test on the previous day, Monday, early in the morning. We can't expect them to have the exact same scores on the Tuesday as they got on the Monday even if we assume that the test was perfect, the administration conditions were almost the same, corrected by the teacher without any bias, and no forgetting or learning has occurred among students in the meantime. It is inevitable that we can never have complete trust in any set of scores. For this reason, our duty is to construct, administer and score tests in such a way that students would get similar scores even if the test was conducted on different day and time (Hughes, 2003). "The more similar the scores would have been, the more reliable the test is said to be" (Hughes, 2003:37).

2.4.4. Estimating Reliability

A reliability coefficient can be calculated to estimate the degree to which test scores are reliable. The ideal reliability coefficient is +1. If a test has the reliability coefficient of +1, it means that, no matter when the test was administered, this test definitely gives the same results for a particular set of students. If a test has a coefficient of zero, then it means that there is no reliability in this test. For example, if the reliability coefficient is .89 on a test score, this would mean that the scores are 89% reliable, with 11% measurement error. Lado (1961, as cited in Hughes, 2003) says that if the reliability coefficient is between .90 and .99 for a vocabulary structure and reading tests, then we can consider these tests as good. For the listening comprehension tests, the range is better if it is between .80 and .89 and oral production tests may range between .70 and .79.

Test-retest, parallel forms, internal consistency strategies and marker reliability are four basic strategies to estimate the reliability of tests. With the test-retest strategy, the same groups of students are given the same test twice. They should be conducted on different times, far enough apart time-wise for the reason that the students they won't likely to remember the

items on the test but it should also be close enough for them not to change influentially like learning more about the topic. The scores of the same student between the two tests are correlated to give a reliability coefficient. "This number can range from -1 to +1 on a continuous scale and 0 indicates total lack of reliability or complete inconsistency while 1 is the ideal value and indicates perfect reliability or complete consistency" (Weir, 2005: 25). The parallel-forms reliability is similar to the test-retest reliability and is an alternate form which is administered in two different sessions with two different but similar tests. The same language skills should be tested on the same breadth of items and any input should be the same in length, degree of topic familiarity and difficulty level (Weir, 2005). A correlation coefficient is calculated later by the tester using the two sets of scores.

Internal consistency is another and the most frequently used form of reliability which focuses on the consistency with each other of a test's internal elements. This is generally measured as Cronbach's Alpha. The questions on same topic are asked in multiple ways, helping in dispensing greater consistency in which Alpha of .90 reflects high reliability, .80 is of moderate reliability and .70 is of low reliability (Vohra, 2007). In many tests, two raters are used, so inter-rater reliability is established via correlation, perfect agreement being indicated by a correlation of 1.0 (Weir, 2005). Two or more independent markers of the same test establish greater consistency and thus reliability to the scores achieved by the student (Vohra, 2007).

2.4.5. Ways to Make Tests More Reliable

The reliability of the scoring is one of the components of test reliability. But there are many other situations which can also make scores unreliable. It is impossible to prevent all the sources that causes the unreliability, such as personal attributes, but there are some ways to increase consisitent performances of the students and hence, the reliability of tests. Hughes (2003:44) suggested some number of useful ways to increase the test reliability in the book, Testing for Language Teachers. In the next section, some of his ideas, which can contribute to the reliability of the English Proficiency Exam, are discussed.

1. Take enough samples of behaviour. Having more items on a test will make the test more reliable. It would be very difficult to determine "how good an archer someone was" (Hughes, 2003 p.36) if we relied on a single shot at the target. This is just the same on a testing situation. If we want to be sure of the performance of our students, we need to have enough evidences, which means, there should be enough independent items on the test so as

not to build an information on an answer through the previous one. Hughes (2003) explains this by a goog example:

Imagine a readig test that asks the question: “Where did the thief hide the jewels?” ıf an additional item following that took the form, “What was unusual about the hiding place?”, it would not make a full contribution to an increase in the reliability of the test. Why not? Because it is hardly possible for someone who got the original question wrong to get the supplementary question right. Such candidiates are effectively prevented from answering the additional question. (p.23)

2. Allow students to have a fresh start. To make the test results more reliable, it will help us to gain more information about all of the candidates. In an oral exam to test the speaking ability, for example, the students need to have as many fresh start as possible. However, this does not mean that the test needs to be too long. It should just be long enough to achieve a satisfactory reliability, not too long for the students to become bored or tired.

3. Exclude items which do not discriminate well between weaker and stronger students. Statistical analysis of items is used to determine the items which do not discriminate well. And for individual items, the item facility and item discrimination values reveal the necessary information.

By item discrimination coefficient, we get the information about the discrimination of the item between the stronger and weaker students. If the coefficient is higher, it means that the item discriminates well. The coefficient can range between the minimum number 0 and maximum 1. The more the items on the test discriminate well, the more the scores reliable will be. Items with a low discrimination coefficient value means that they are in need of improvement. These items should be reviewed or even taken out from the test.

The facility value , on the other hand, gives us information about the difficulty of an item. It reveals the perentage of students who got the answer right. By the help of this value, we can easily decide on the easy and difficult items. If the value is higher, it means that the item is very easy. Item facility value serves us when we decide about our purpose of the test. If we need to develop a proficiency test to” identify the top 10% of students, items on the test must be sufficiently difficult” (Starr, 2008). This kind of a test will have a high proportion of items that have a low facility value.

It will be better for a test to leave the extremely easy and difficult items. That test can discriminate better between weaker and stronger students then. A test with a very high or very low discrimination value means that the test is too easy or too difficult and does not discriminate well. However, “a small number of easy, non-discriminating items may be kept

at the beginning of a test to give candidates confidence and reduce the stress they feel" (Hughes, 2003 p.45).

4. Do not allow candidates too much freedom. In some kind of test, students are generally given several choices to choose their own questions to answer. This situation also give them a great deal of freedom to answer the easier questions. This is especially very common in writing tasks. The students are usually given a number of titles and asked to choose the one that they want to write about. This situation may have a negative effect on the reliability of the test because "the questions in themselves can vary in difficulty and require different emphasis in skill in order to perform the task" (Starr, 2008). This also may cause a problem in the scoring procedure. For these reasons, developing a test on one topic will be more reliable as it will also allow us to make a comparison between the students directly.

5. Write unambigious items and provide clear and explicit items. It is very important in a test to have a clear and explicit instructions. It also should not have more than one possible, correct answer. The items and the instructions should be developed in a way that the students could only give the answers that the examiner expects. To achieve this, it is important to give enough or may be a lot of information about how to perform the task. In a vocabulary test, for example, " the test developer should be aware of all the meanings of a word asked for. The item should then be worded so that either all or any of the meanings of the word are acceptable answers" (Starr, 2008, p.39).

6. Use items that permit scoring which is as objective as possible. When objectivity is considered, one may easily think that multiple choice items are the best as it allows completely objective scoring. Although they are a good and easy way of assessing, they are not appropriate to serve for all purposes. Moreover, it is really difficult to develop a good multiple choice test as it always require extensive pre-testing. Fill-in–the-blank items, open-ended items or essay questions which has a unique, possibly one-word, correct response, and the candidates produce themselves may be the alternatives for the multiple choice items. This also should provide clear and explicit instructions and the expected responses should be guided so as not to leave too much freedom.

7. Provide a detailed scoring key. This should be based on a clearly stated proficiency scales. A detailed scoring key is a fundemental tool which makes the scoring more objective. Such a scoring key which is as detailed as possible in its assignment of points, also provides high scorer reliability. It gives us information and provide a guideline about the performance of the students as well.

8. Train scorers. The most subjective part of testing is the scoring procedure. This makes trained scorers important. The scoring should be conducted by the ones who are familiar with the proficiency levels and trained on the scoring procedures. After the administration of each test and the scoring, patterns of scoring should be analyzed. The rating scale should be applied in a wide range to identify the true levels of proficiency of the students. "Individuals whose scoring deviates markedly and inconsistently from the norm should not be used again" (Hughes, 2003, p.49).

2.4.6. Standard Error of Measurement

The standard error of measurement is used to determine how far it is worth taking the reported score at face value (Weir, 2005). Hughes (2003) says "while the reliability coefficient allows us to compare the reliability of tests, it does not tell us directly how close an individual's actual score is to what he or she might have scored on another occasion" (p.40). We can determine "a range around a student's actual score within which that student's score would probably fall if he or she were to take the same language test over and over again, without the effect of remembering the items or learning more of the language" (Starr, 2008:33). Hughes (2004) illustrated an example with different statements and said:

> Suppose that a test has a standard error of measurement of 5. An individual scores 56 on the test. We are then in a position to make the following statements: 1-we can be about 68 per cent certain that the person's true score lies in the range of 51-61 (i.e. within one standard error of measurement of the score actually obtained on this occasion). 2-we can be about 95 per cent certain that their true score is in the range 46-66 (i.e. within two standard errors of measurement of the score actually obtained). 3-we can be 99.7 per cent certain that their true score is in the range 41-71 (i.e. within three standard errors of measurement of the score actually obtained). (p.41)

We now know and have seen the importance of reliability. For this reason, we should be very careful when we need to make important decisions on the basis of the test scores of students "whose actual scores place them close to the cut-off point (the point that divides 'passes' from 'fails')" (Hughes, 2004:42).

CHAPTER 3

3. METHODOLOGY

3.1. Introduction

Chapter Two provided a theoretical basis for the validation study of the English Proficiency Exam (EPE) for the Preparatory School at Çağ University. Chapter Three applies the theory to an investigation into the validity of the EPE. The method of this study includes these parts: research design, research questions, participants, instrumentations, data collection and, data analysis.

3.2. Design of the Study

This study was a validation study of the English Proficiency Exam conducted at Çağ University Preparatory School. The test, with its components, was given to Turkish students at Çağ University in their freshman semester beginning in fall semester 2011. Qualitative and quantitative data were collected and analyzed for these tests. The validity of the English Proficiency Exam at Çağ University was investigated. The exam, given at the beginning of the Fall Semester 2011-2012, was collected and each component of the exam was described. Finally, quantitative data was collected and analyzed for these tests.

3.3. Participants

The participants of the study are 133 students at Çağ University registered between the Fall Semester 2011-2012. They are between 18 and 25 years old. Among these 133 students, 45 of them registered for the Law Department, 13 for the International Relations, 16 for Management, 22 for the English Language Teaching, 14 for the International Finance, 1 for Public Relations, 16 for the International Trade, and 6 for Mathematics and Computer Scince. These students are required to take the English Proficiency Exam during the first year. Students whose level of English is insufficient are required to enroll in the English Preparatory School, as except for the Faculty of Law, the language of instruction is English. The students should get 70 points out of 100 to be successful and pass the exam.

3.4. Instrumentation

The purpose of the English Proficiency Exam is to assess the level of English proficiency of students at Çağ University. It provides information on how fluent the students are and how well they perform in the different language areas of listening, reading, writing,

vocabulary, dialogue and grammar. For that reason, the data of the study were collected through six components of the English Proficiency Exam which was conducted at the beginning of the Fall Semester 2011-2012.

3.4.1. Description of the English Proficiency Exam

The aim of the English Proficiency Exam is to identify the acquisition of English language skills (fluency and grammatical knowledge) in most areas of competence—reading, writing, and listening comprehension—including an understanding of the structure of the English language. All students in Çağ University can take the English Proficiency Exam to get the required points so as to start their university education in their departments, as the language of instruction, except for Faculty of Law, as mentioned before, is English. The students should get 70 points out of 100 to be successful and pass the exam. It can be understood that the main purpose in here, is to identify between the level of students as the ones who can start doing their degree in their departments and the others who are insufficient and need further education in English Language and so required to enroll in the English Preparatory School. There are three levels at Çağ University Preparatory School: beginner, elementary and pre-intermediate.

The beginner level of students has 28 hours of English lessons every week. 18 hours are dedicated to course book, 5 hours to listening & speaking and 5 hours to reading & writing. In the beginner level, the students take 14 pop quizes each term, 6 of which are course book quizes, 4 of them are reading & writing and 4 of them are listening & speaking. They also have 3 monthly exams each term and they take the final exam at the end of their academic year.

The elementary level of students, on the other hand, has 26 hours of English lessons while the pre-intermediate ones have 24. They have the same amount of listening & speaking and reading & writing classes as the beginner level of students. The number of the pop quizes, the monthly exams and the final exam they take is the same as the beginners as well. The contributions of the quizzes and exams to be successful in preparatory school are shown in Table 2.

For that reason there are six components to the English Proficiency Exam: listening, reading, grammar, vocabulary, dialogue completion, and writing. In order to answer the research questions, each instrument component, with its procedures and data analysis, will be described in the following section.

Table 2. ***Contributions of the Quizzes and Exams to Be Successful in Preparatory School***

Type of Exam	**Number of Exams**		**Percentage (%)**				
	1st Term	**2nd Term**	**1st Term**		**2nd Term**		**Total**
Pop Quizzes	14	14	20%	+	25%	=	45%
Monthly Exams	3	3	25%	+	30%	=	55%
					Annual Total	=	100%

Note. Passing Grade: Final exam result must be a minimum of 50%

Annual Total Grade 40% + Final Exam Result 60% = Passing Grade

3.4.2. Listening Comprehension

The aim of the listening comprehension part is to find how well the students understand what they hear in English. This part is paper based and is conducted separately from the other five components, as the first part of the EPE Exam, at the very beginning of the testing time. The test consists of ten short essay question items. Each item is a question related to the audio passage the students listen to. First, the students take the paper test with the ten item questions and are given a few minutes to read the questions, so that they have the idea of what to look for. After that, the listening material is selected from the CD player and played once, so as to notes can be taken. The students are given some time to answer the questions on their papers. At last, the audio passage is played again for the aim of giving students the second chance to make their final decisions about the correct answers.

The tests are collected by the teaching assistant administering the test. After the end of the testing procedure, the graders score each question by the scoring key provided. The maximum score that can be given to each answer is 1 point. The maximum score that can be given to the listening comprehension component is 10 points. In order to facilitate data analysis, the scores of this EPE component were transferred through SPSS.

3.4.3. Reading

The aim of the reading component of the English Proficiency Exam is to find how well students understand what they read in English. This part is a paper-based test and is

conducted separately from the other five components. The students are asked to analyze the ideas presented in seven brief passages with three questions each. The students read each passage and answer the three questions underneath each of the passage. The aim of this part is to decide whether or not the students are able to read critically in order to identify important ideas, understand direct statements, draw influences and conclusions, detect underlying assumptions and recognize word meanings in context. There are also kind of questions that students are asked to select the answer choice that best summarizes a passage, explains a word in a context, compares or contrasts two aspects of a passage, explains the implications or suggestions made in a passage, and identifies casual relationships.

The reading component is corrected by the graders using the scoring key provided. The maximum score that can be given for each task is 24 as there are 24 questions in this part. In order to facilitate data analysis, the scores of this EPE component were transferred through SPSS.

3.4.4. Grammar

The aim of the grammar part of the English Proficiency Exam is to find how well students can analyze sentences and identify the required structural forms in a given sentence. This part also is a paper-based test and administered separately from the other five components. The questions in grammar part consist of 24 different sentences with blanks for grammatical structer parts. For the blanks, five different items are provided under each question and students are supposed to identify the correct answer among them.

The grammar part is corrected by the graders using the scoring key provided. The maximum score that can be given for each question is 2 points, and the maximum total score is 48. In order to facilitate data analysis, the scores of this EPE component were transferred through SPSS.

3.4.5. Vocabulary

The aim of the vocabulary part of the English Proficiency Exam is to determine the vocabulary knowledge of the students, and to determine whether or not the students are able to understand the meaning of a particular word or phrase in the context of a sentence. Students are asked to consider grammatically similar words and choose the one that fits most logically into each sentence in place of a synonymous word. The component is a paper-based test and administered separately from the other five components. This part consists of 12 multiple-choice items with five distracters each. The students are given sentences with a target

vocabulary written in bold in each of them and they are asked to choose the synonyms of these words among the given five multiple-choice items, which are a), b), c), d), or e).

The vocabulary part is corrected by the graders using the scoring key provided. The maximum score that can be given for each question is 2 points and the maximum total score is 24. In order to facilitate data analysis, the scores of this EPE component were transferred through SPSS.

3.4.6. Dialogue Completion

The aim of the vocabulary part of the English Proficiency Exam is to determine whether the students can construct meaning from texts and monitor their reading to ensure that they in fact understand what they read. The component is a paper-based test and administered separately from the other five components. This part consists of 10 multiple-choice items with five distracters each. The students are given structures in dialogue type questions and one sentence or structure is given in blank for the aim of having the students identify the most appropriate structure from the given multiple-choice items.

The dialogue completion part is corrected by the graders using the scoring key provided. The maximum score that can be given for each question is 2 points and the maximum total score is 20. In order to facilitate data analysis, the scores of this EPE component were transferred through SPSS.

3.4.7. Writing

The aim of the writing part of the English Proficiency Exam is to determine how correctly and fluently the students are able to write a composition. The component is a paper-based essay question test and administered separately from the other five components. For the writing test, the students are given three topics. They choose one of these topics and write an essay of 150 words on it. For the first topic choice, they are given a discussion question to comment on the good and bad sides of the topic. The second topic asks the students write the effects of the given topic for which they give their own opinion and for the third topic option, the students are given an imagination situation and are asked to write an essay about what they would do in that situation.

The grading procedure is conducted by the native-speaker of English Professors, who checks the essays for correctness in areas including; word endings, word order, verb forms, idiomatic phrases, spelling and punctuation, and content. The maximum score given for the writing component is 20 points.

3.5. Data Collection

To conduct the study of validity and reliability of the English Proficiency Exam at Çağ University, the necessary permission from the Head of Çağ University Preparatory School was taken. Later on, each component of 133 students' test scores, conducted between 2011-2012 Fall Semester, was collected to analyze.

3.6. Data-Analysis

All test content, including all the test components, were compared to the major language ability construct categories derived from the model for language ability by Bachman (1990), in order to estimate how well the English Proficiency Exam represents a general language construct. Table 3 shows the categories of Bachman's model. The components of the English Proficiency Exam were listed first, and then, the language ability construct areas from Bachman's model of language ability, which can function as the basis for defining the language construct of a test, were adapted. The corresponding categories, which are grammatical knowledge, textual knowledge, functional knowledge, and socio-linguistic knowledge, are listed under each of these language ability areas. Finally, the content of each English Proficiency Exam component was analyzed and the categories, which are covered by the content of each component, were checked. The analyzing procedure was conducted by three teachers from Çağ University. Two of them were from the Preparatory School Department and one was from the English Language Teaching Department. Bachman's model and the directions from the book called *Language Testing in Practice* by Bachman and Palmer (1996) was followed to check the content of the English Proficiency Exam.

Table 3 was adapted from Bachman (1996, p. 68) and is a fuller visual metaphor of language competence, organizational and pragmatic language knowledge.

Table 3. ***Bachman's Model of language Ability***

<table>
<tr><td colspan="2">ORGANIZATIONAL KNOWLEDGE
(how utterances or sentences and texts are organized)</td></tr>
<tr><td>GRAMMATICAL KNOWLEDGE
(how individual utterances or sentences are organized)</td><td>TEXTUAL KNOWLEDGE
(how utterances or sentences are organized to form texts)</td></tr>
<tr><td>• Knowledge of vocabulary
• Knowledge of syntax
• Knowledge of phonology/graphology</td><td>• Knowledge of cohesion
• Knowledge of rhetorical or conversational organization</td></tr>
<tr><td colspan="2">PRAGMATIC KNOWLEDGE
(how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language use setting)</td></tr>
<tr><td>FUNCTIONAL KNOWLEDGE
(how utterances or sentences and texts are related to the communicative goals of language users)</td><td>SOCIOLINGUISTIC KNOWLEDGE
(how utterances or sentences and texts are related to features of the language use setting)</td></tr>
<tr><td>• Knowledge of ideational functions
• Knowledge of manipulative functions
• Knowledge of heuristic functions
• Knowledge of imaginative functions</td><td>• Knowledge of dialects/varieties
• Knowledge of registers
• Knowledge of natural or idiomatic expressions
• Knowledge of cultural references and figures of speech</td></tr>
</table>

Taken from "Language Testing and Practice" by Bachman & Palmer, 1996, p.68.

Cronbach's Alpha was calculated for the listening comprehension exam, the reading exam, the grammar exam, vocabulary exam, and the dialogue completion exam to estimate the reliability of the test scores. Moreover, overall descriptive statistics including mean and standard deviation of total scores were examined for each of these exams.

CHAPTER 4

4. RESULTS

In chapter four, the results of the test analysis are presented to answer the research questions about the validity of the English Proficiency Exam. The content of each test component is analyzed first, and then, Cronbach's alpha reliability coefficient of the listening, reading, grammar, vocabulary and dialogue completion components is presented. Finally, the results of the data analysis for each component of the English Proficiency Exam are presented.

4.1. Content Analysis

Bachman's (1990) model for language ability is used to analyze the content of the English Proficiency Exam. The aim of the content analysis is to examine how broadly the content of the test components takes in the major language ability construct categories. In this model, language knowledge includes "*organizational knowledge*" and "*pragmatic knowledge*". The areas of organizational knowledge are; grammatical knowledge, which also includes the knowledge of vocabulary, syntax, phonology and graphology, and textual knowledge which includes the knowledge of cohesion and rhetorical or conversational organization. Two areas of pragmatic knowledge are functional knowledge and sociolinguistic knowledge. Functional knowledge includes four categories which are ideational, manipulative, instrumental, and imaginative. The categories of the sociolinguistic knowledge are as follows: registers, natural or idiomatic expressions and cultural references or figures of speech.

The organizational knowledge areas covered by the English Proficiency Exam are presented in Table 4. This Table presents an overview of the construct areas of organizational knowledge covered by the English Proficiency Exam. The English Proficiency Exam components are listed in the columns and the construct areas are listed in the rows. Grammatical knowledge is the ability to use phonological, morphological, syntactic and semantic structures of the language. Textual knowledge involves the speaker's knowledge and ability in the areas of organizing information in a coherent manner and making effective use of cohesive devices like, however, therefore, and, to help the listener follow the organization of the response. It can be seen that the English Proficiency Exam components cover most of the general construct areas. For example, in the listening component, there are questions which require the students to listen to and hear some numbers, times and dates from the

listening task and select the correct item from the test. These kinds of questions were checked in the phonology or graphology column. And in the reading part, the questions which require the students to find the paraphrased sentences or words were checked in the rhetorical or conversational organization column and most of the questions require the knowledge of syntax and cohesion in most of the English Proficiency Exam components.

Table 4. ***Overview of language Constructs Area Coverage- Organizational Knowledge***

	listening	reading	grammar	vocabulary	dialogue completion	writing
Grammatical knowledge						
Vocabulary		✓	✓	✓		✓
Syntax	✓	✓	✓			✓
Phonology graphology	✓					✓
Textual knowledge						
Cohesion		✓	✓		✓	✓
Rhetorical or conversational organization		✓				✓

Table 5 presents an overview of the construct areas of pragmatic knowledge covered by the English Proficiency Exam. Functional knowledge is the speaker's ability to select language functions to reasonably address the test task. Bachman and Palmer (1996) call this as "illocutionary competence, enables us to interpret relationships between utterances or sentences and texts and the intentions of language users" (p. 69). The questions that require this kind of knowledge were checked in this part of the model. For example the questions, which include jokes and the use of the figurative language or poetry, were checked in the imaginative functions column.

Table 5. ***Overview of language Constructs Area Coverage- Pragmatic Knowledge***

	listening	reading	grammar	vocabulary	dialogue completion	Writing
Functional knowledge						
Ideational functions						✓
Manipulative functions					✓	✓
Heuristic functions						✓
Imaginative functions						✓
Sociolinguistic functions						
Registers					✓	
Natural or idiomatic expressions			✓		✓	✓
Cultural references/figures of speech					✓	✓

Sociolinguistic knowledge involves the speaker's ability to demonstrate an awareness of audience and setting by selecting socially and culturally appropriate language and register. Bachman (1996) states that "when we use cultural references, such as 'beyond the pale', or figures of speech, such as 'Do not push my buttons', to convey our intended meaning appropriately, we are using sociolinguistic knowledge" (p: 70). So, following this information, the questions which require use of dialects or varieties, registers, natural or idiomatic expressions, cultural references and figures of speech, were checked in the sociolinguistic functions column.

In terms of the English Proficiency Exam test content used to measure the above four competencies holistically, it can be concluded that it matches the theory of communicative language ability behind it.

4.2. Test Score Reliability

In order to estimate the reliability of the listening comprehension exam, the reading exam, the grammar exam, the vocabulary, and the dialogue completion exam of the EPE, Cronbach's alpha reliability coefficient was calculated.

According to Lado (1961, cited in Hughes, 2003: 39), "good vocabulary, structure and reading tests are usually in the .90 to .99 range, while auditory comprehension tests are more often in the .80 to .89 range; and oral production tests may range from .70 to .79". Table 6 shows that, the reliability coefficient of listening, reading, grammar, vocabulary, and dialogue completion are so close to each other but the most striking ones are listening and reading. The reliability coefficient of the listening component is .82 and the reading component is .91, which means they are at the intended level. Starr (2008) explains that, the reliability coefficient of a test is usually lower when the test has fewer items. We can assume that the general ability level is quite high since the focus of EPE is to get the required points to start the university education in the departments in English language. However, it is very crucial to have a deeper look at the descriptive statistics and the item analysis in order to determine any variance.

Table 6. ***Cronbach's Alpha***

	Listening	Reading	Grammar	Vocabulary	Dialogue Completion
Cronbach's Alpha	0.82	0.91	0.87	0.81	0.82

4.2.1. Listening

Table 7 represents descriptive statistics of the listening comprehension component. In this table, the descriptive statistics include the mean, standard deviation, and the standard error of measurement and variance. There are 10 questions in the listening component and 20 points total is given to this part. We can see that the mean of the listening component is 14, 24 and the standard deviation is 3, 34 while the variance is 11,21. The mean of the listening component is quite high but the standard deviation is low. The item analysis provides more information. When we look at the item analysis in Table 8, it shows the mean of each item which is around 1 point out of 2 total points possible for each item. Students who gave incorrect answers were given 1 point and the ones who gave correct answers were given 2

points. According to the Table, items 1, 3, and item 9 have the highest and exactly the same mean value. This suggests that these items might be too easy. On the other hand, the mean of item 5 and 6 show the lowest value among the other items. The discrimination of the other items are quite well. For this reason, the items that seem too easy might need to be revised. On the other hand, keeping some of the difficult items might be useful for the stronger students.

Table 7. ***Descriptive Statistics of the Listening Component***

	Listening
Mean	14.24
Standard Deviation	3.34
Standard Error of Measurement	.210
Variance	11.21

Table 8. ***Item Analysis of the Listening Component Items***

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Mean	1,7	1,5	1,7	1,3	1,1	1,1	1,6	1,6	1,7	1,2
Std. Deviation	0,5	0,6	0,6	0,6	0,4	0,5	0,6	0,6	0,5	0,4
Variance	0,3	0,4	0,3	0,3	0,3	0,3	0,3	0,3	0,2	0,2
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

4.2.2. Reading

Cronbach's alpha for the reading test (0.91) is just at the recommended level. The mean of the reading component shown in Table 9 is 29.76 and it gives relatively good standard deviation which is 7.04. The variance value of this component is 49.6. There are 24 questions in this section of the English Proficiency Exam and the total point given for this section is 48.

Table 10 shows the item analysis of this component and it reveals that item 9, 12, 15, 16, and 21 have the lowest mean value. This suggests that these items are the most difficult ones and need to be revised or taken out of the test. When we look at the other components, we can also see that the value of most of the other items is similar. So, revising those similar items also might help to increase the validity of the reading component. Item 17, on the other

hand, has the highest mean value and it is the easiest item on the test. Considering the low mean value of the other items, keeping item 17 on the test might be useful for the weaker students.

Table 9. ***Descriptive Statistics of the Reading Component***

	Reading
Mean	29.76
Standard Deviation	7.04
Standard Error of Measurement	.210
Variance	49.6

Table 10. ***Item Analysis of the Reading Component Items***

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
Mean	1,4	1,4	1,3	1,2	1,2	1,2	1,2	1,3	1,0	1,1	1,3	1,0
Std. Deviation	0,6	0,6	0,6	0,6	0,6	0,5	0,6	0,6	0,4	0,5	0,6	0,4
Variance	0,4	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,2	0,2	0,3	0,2
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24
Mean	1,2	1,2	1,0	1,0	1,6	1,1	1,1	1,2	1,0	1,2	1,2	1,1
Std. Deviation	0,6	0,5	0,5	0,5	0,6	0,5	0,6	0,6	0,5	0,6	0,7	0,6
Variance	0,4	0,3	0,2	0,2	0,3	0,2	0,3	0,3	0,2	0,3	0,4	0,3
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

4.2.3. Grammar

The descriptive statistics in Table 11 show that the standard deviation for grammar component is 6.58 and the mean value is 29.59. There are 24 items in this section and the overall point given to this section is 48. Considering this, the grammar component has a relatively good standard deviation and mean value. Nevertheless, the item analysis, as shown in Table 12 gives more accurate information as the Cronbach's alpha value of this component

is slightly lower (0.87) and thus a source of lower reliability. Examining some of these items might help to increase the reliability of the grammar component to the recommended level.

Table 11. ***Descriptive Statistics of the Grammar Component***

	Grammar
Mean	29.59
Standard Deviation	6.58
Standard Error of Measurement	.210
Variance	43.3

Items, which seem very difficult with a very low mean value, need to be revised or taken out of the item pool. Table 12 shows the item analysis of the grammar component. Very difficult items with a very low mean value need to be examined which are items 4, 9, 10, 11, 13, 14, 15, 19, and 23. Even though items 1, 2, 3, 12, 18 and 20 have the highest mean with a value of 1.3, among the others, this is obviously not quite a high value. Item 14 have the lowest mean value between the other items.

Table 12. ***Item Analysis of the Grammar Component Items***

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
Mean	1,3	1,3	1,3	1,1	1,2	1,2	1,2	1,2	1,1	1,1	1,1	1,3
Std. Deviation	0,5	0,5	0,6	0,4	0,5	0,5	0,6	0,5	0,5	0,5	0,6	0,6
Variance	0,3	0,3	0,3	0,2	0,3	3,0	0,3	0,3	0,2	0,3	0,3	0,3
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24
Mean	1,1	1,0	1,1	1,2	1,2	1,3	1,1	1,3	1,2	1,2	1,1	1,2
Std. Deviation	0,5	0,5	0,5	0,5	0,6	0,6	0,6	0,5	0,5	0,6	0,6	0,6
Variance	0,2	0,3	0,2	0,2	0,3	0,3	0,3	0,3	0,3	0,4	0,3	0,3
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

4.2.4. Vocabulary

The descriptive statistics presented in Table 13 show a very low standard deviation of 3.7. The mean value of this component is 14.28 with the variance value of 14.14. The total point given to this section is 24 out of 12 questions. The students are asked to find the synonyms of the highlighted words, which are asked in sentences, among five distracters.

Table 13. ***Descriptive Statistics of the Vocabulary Component***

	Vocabulary
Mean	14.28
Standard Deviation	3.7
Standard Error of Measurement	.417
Variance	14.14

Table 14 shows the item analysis of the vocabulary component. Items 2, 5, and 7 have the lowest mean with a value of 1.0. This low value can be interpreted as the most difficult items of this component and need to be revised or rejected. Although items 2, 5, and 7 have the lowest values and are in need of improvement, the other ones also seem to be in need of improvement too, as their value range is not fairly higher than items 2, 5, and 7. Item 8 has the highest mean with the value of 1.4, which also seems to be quite low. For this reason, apart from the lowest items which are 2, 5,7 and the highest item 8, the other items which are below the mean value of 1.2, such as items 1, 3, 9, and 10, need to be examined.

Table 14. ***Item Analysis of the Vocabulary Component Item***

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
Mean	1,1	1,0	1,1	1,2	1,0	1,2	1,0	1,4	1,1	1,1	1,2	1,2
Std. Deviation	0,5	0,5	0,5	0,6	0,5	0,5	0,5	0,6	0,5	0,5	0,6	0,6
Variance	0,3	0,2	0,3	0,3	0,3	0,3	0,3	0,4	0,2	0,3	0,3	0,4
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

4.2.5. Dialogue Completion

The descriptive statistics represented in table 15 show a slightly low mean, with also low standard deviation. The mean value of the dialogue component is 12.40, and the standard deviation is 3.33 with the variance value of 11.10. Considering 20 overall points given to this section out of ten questions, these values are quite low. In the dialogue completion part, students are given ten different dialogues with one missing line in each of them and are asked to find the appropriate respond among the five given items.

Table 15. ***Descriptive Statistics of the Dialogue Completion Component***

	Dialogue Completion
Mean	12.40
Standard Deviation	3.33
Standard Error of Measurement	.210
Variance	11.10

When we have look at the item analysis Table 16, we can see that items 2, 4, 6, 8, and 9 have the lowest mean value. Item 7 on the other hand has the highest mean value of 1.5. It is clear from the Table 16 that, except from the items 5 and 1, the rest of the value of the items are below 1.2, which suggests that they also need to be revised and improved.

Table 16. ***Item Analysis of the Vocabulary Component Item***

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Mean	1,3	1,1	1,2	1,1	1,3	1,1	1,5	1,1	1,1	1,2
Std. Deviation	0,5	0,5	0,6	0,5	0,6	0,5	0,6	0,5	0,5	0,5
Variance	0,3	0,3	0,3	0,3	0,3	0,3	0,4	0,2	0,2	0,3
Range	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0

The reason for the low values here might be that, this part requires the sociolinguistic knowledge of the language ability construct. If we refer back to the Table 5 on page 32, we can see that the dialogue completion is mainly constructed on the sociolinguistic functions area of the language ability. So, we can assume also that the students might be lack of the knowledge of cultural references, figures of speech, registers and natural or idiomatic expressions, which are needed to be able to answer the related questions here correctly.

4.3. Summary

This chapter has presented the results of the data analysis that collected the relevant evidence to answer the research questions in Chapter Five. Bachman's model for the language ability is used for the content analysis of the English Proficiency Exam and it showed that the exam covers most of the language construct areas in different components. Cronbach's alpha co efficiency was calculated for the reliability of the English Proficiency Exam. It revealed that the listening and the reading components have the highest Cronbach's alpha value although each of the other components, which are grammar, vocabulary and dialogue completion have relatively high values. The analysis of the descriptive findings was done through standard deviation, mean, variance and standard error of measurement. The analysis showed that some items in each of the components need to be revised or even rejected.

Chapter Five will provide a discussion of these findings with suggestions for improving the English Proficiency Exam and limitations of this study for the direction of future research.

CHAPTER 5

5. DISCUSSION, CONCLUSION AND RECOMMENDATIONS

The aim of this chapter is to find the answers to the research questions in terms of that evidence, and later, to discuss implications, limitations, and suggestions concerning this study. First, the research questions are answered through the data analysis to yield a better understanding of the results. Second, suggestions to improve the English Proficiency Exam and each of its components are presented. After that, a discussion of teaching implications of this study followed the suggestions. And finally, after the presentation of the final conclusion, suggestions for further research study are offered.

5.1. Discussion on the Content of the English Proficiency Exam

It is very important for a general proficiency test to cover the most important components of language ability. The content of the English Proficiency Exam, with its six components, covers all the language ability construct areas, which were suggested by Bachman (1996) in his model of general language ability. The study conducted to find the content of the English Proficiency Exam showed that it matches the theory of communicative language ability. The analysis represented very positive results. Nevertheless, this definition of language ability does not supply the specific definitions about the construct of independent language skill areas like, listening, reading, grammar, writing, speaking, vocabulary and dialogue completion. It only features the general language construct. Concerning each components of the English Proficiency Exam, suggestions to increase the validity are as follows:

- The construct of each EPE component could be defined and the content of the test could be based on the specific areas of the defined language skill constructs.
- The items on the EPE could be improved so as to be sure that examinees provide a language sample which is extensive enough to judge all the language competence areas selected for the test.
- The existing proficiency level definitions could be developed or provided on the basis of the defined language skill construct.
- A detailed scoring procedure and a scoring key could be developed or adapted
- The scoring procedure could be re-arranged and more scorers training could be included concerning the proficiency levels.
- A scoring guide that defines levels of performance could be provided

5.2. Discussion on the Reliability of Each Component of the English Proficiency Exam

The Cronbach's alpha reliability coefficient is used to estimate the reliability of the listening comprehension exam, the reading exam, the grammar exam, the vocabulary exam, and the dialogue completion exam. The reliability values of each of these components are 0.82, 0.91, 0.87, 0.81, and 0.82 respectively. As it is known, the closer Cronbach's alpha coefficient is to 1.0 the greater the internal consistency of the items in the scale. The reliability coefficient of listening and reading component are very high when they are compared to the other components. Although the grammar, the vocabulary, and the dialogue completion components have relatively high internal consistency as a reliability coefficient of 0.70 or higher is generally considered acceptable in most social science research situations, their reliability coefficients are slightly lower than the range suggested by Lado (1961). The reliability level of test scores is usually affected by the spread of proficiency levels and the number of items on a test. More number of items in a test can raise the value of alpha. Even though the test scores seem reliable, we need to have a closer look at the items of each of the components, because the high reliability does not mean that the test scores are valid.

It was not possible to estimate the reliability coefficient of the writing component, as the students obtained a single score for this component. Moreover, it was not also possible to calculate inter-rater reliability because the components were scored by one scorer. It is important to have information obtained through inter-rater reliability estimates to be able to draw conclusions about the reliability of test scores to improve such tests. In his book, *Testing for Language Teachers,* Hughes (2003) provides some practical ways that can help to increase test reliability. A few of these strategies that could increase reliability of the English Proficiency Exam are given as suggestions in the following section.

Suggestions;

- Take enough samples of behavior, as the more items that a test has the more reliable it will be.
- Each additional item should as far as possible represent a fresh start for the students. For example, in an interview used to test oral ability, the student should be given as many fresh starts as possible and the more independent passages reading or writing items have, the more reliable they will be.
- The more important the decisions based on a test, the longer the test should be

- Exclude items which do not discriminate well between weaker and stronger students as items on which strong and weak students perform with similar degrees of success contribute little to the reliability of a test.
- Do not allow students too much freedom. Students should not be given a choice, and the range over which possible answers might vary should be restricted.
- Write unambiguous items as it is essential that candidates should not be presented with items whose meaning is not clear or to which there is an acceptable answer which the test writer has not anticipated.
- Provide clear and explicit instructions.
- Ensure that tests are well laid out and perfectly legible.
- Make students familiar with format and testing techniques. If any aspect of a test is unfamiliar to candidates, they are likely to perform less well than they would do otherwise.
- Provide uniform and non-distracting conditions of administration.
- Use items that permit scoring which is as objective as possible. This does not mean that the multiple choice items are the most appropriate ones to do so. Some other alternative items such as open-ended items which has a unique response could be included in tests.
- Provide a detailed scoring key. This should specify acceptable answers and assign points for acceptable partially correct response. For high scorer reliability the key should be as detailed as possible in its assignment of points.
- Train scorers. This is especially important where scoring is most subjective. The scoring of compositions, for example, should not be assigned to anyone who has not learned to score accurately compositions from past administrations. After each administration, patterns of scoring should be analyzed. Individuals whose scoring deviates markedly and inconsistently from the norm should not be used again.
- Agree acceptable responses and appropriate scores at outset of scoring. A sample of scripts should be taken immediately after the administration of the test.
- Employ multiple, independent scoring. As a general rule, and certainly where testing is subjective, all scripts should be scored at least two independent scorers. Neither scorer should know how the other has scored a test paper. Scorers should be recorded on separate score sheets and passed to a third, senior, colleague, who compares the two sets of scores and investigates discrepancies.

5.3. Discussion on the Variation in Test Scores and Potential Item Problems of Each Component of the English Proficiency Exam.

In this part, each component of the English Proficiency Exam is discussed concerning the variation in test scores, item difficulty and item discrimination values.

5.3.1. Listening Comprehension

The item analysis of the listening comprehension indicated that there is variation in test scores. Even though the range of scores for each item is very wide, the scores of items 1, 3, and 9 concentrate around the score 2, which has an effect on the discrimination and facility of the item. For these items there is not sufficient variation in test scores.

Even though the listening component only has ten items and the variation of test scores for two items is limited, the component overall shows a relatively good reliability coefficient and can be considered a good assessment tool for the language skill of listening. Nevertheless, some of the questions still need to be improved in terms of the low item analysis value and also the easy items.

Suggestions:

- More listening passages of medium length and higher difficulty could be added and several related questions for each of the listening passages could be asked. This will increase the number of items in this component, improve the item facility and discrimination, and hence positively influence the reliability.
- Items that permit scoring as objective as possible could be used. An alternative to multiple choices is the open-ended item which has a unique, possibly one-word, correct response which the students produce themselves.
- A detailed scoring key could be provided as the key should be as detailed as possible in its assignment of points for the high scorer reliability.

5.3.2. Reading

The item facility of each item showed that items 9, 12, 15, 16, and 21 are the most difficult items as the students receive a very low score for each of these items. The other 19 items out of 24 indicate that there is sufficient variation in test scores. The reading component overall does show a good reliability coefficient and therefore can be considered a good assessment tool for the language skill of reading.

Suggestions;

- Items 9, 12, 15, 16, and 21 could be reviewed to determine whether the difficulty of

these items should be decreased, revised or totally taken out of the test as items on which strong students and weak students perform with similar degrees of success contribute little to the reliability of a test.

- Some of the other items, which have high difficulty value, might be revised to make them easier as a small number of easy, non-discriminating items may be kept at the beginning of a test and may also be useful to give students confidence and reduce the stress they feel.
- More specific questions for each of the reading passages could be asked.

5.3.3. Grammar

About two thirds of the items of the grammar component can be considered difficult. The results of the item analyses showed that there are many difficult items in general, which seem lower the discrimination ability of the items. Even though the easiest items with the highest mean value (1.3) among the others did not show relatively high value. Most of the difficult items do not discriminate well.

The department should consider modification or replacement of the more difficult and less discriminating items. These items could be changed to be fairly easier and cover language material that is more appropriate for intended levels of grammatical knowledge. In order to do that, it is important to be aware of all the components of a grammar construct and to specify which grammar components and functions are more prevalent in intended levels of language proficiency.

Suggestions;

- Items 4, 9, 10, 11, 13, 14, 15, 19, and could be revised to be easier.
- It should be kept in mind that assessors should use different kinds of methods to assess the grammar component, so that they can test different language functions and can also cover a variety of grammar functions. Other methods to assess the grammar component may include rewriting sentences by the help of a specific grammar function, constructing sentences using sentence parts given in their basic form, in response to a clue sentence given, students could be asked to write short sentences including a specific grammar function and could also be asked to substitute sentence parts with an alternate grammar function.

5.3.4. Vocabulary

Most of the items in the vocabulary component are extremely hard. A component containing many words that are not commonly used or are antiquated can contribute to low item facility. According to the item analysis of the vocabulary component, the items do not discriminate well. Except from the item 8, the mean value of all the other items is 1.2 or below. For this reason, these items should be revised.

Suggestions;

- All the items in this component could be revised and possibly substituted by more commonly used words.
- The type of questions that are asked to assess the vocabulary knowledge of students could be varied. Instead of asking only the synonyms of words, definitions, appropriate words and recognizing appropriate words for context, gap filling, and etc. could be asked to assess this skill.

5.3.5. Dialogue Completion

The item analysis of the vocabulary component also revealed that the items are very difficult in this section as it is in the vocabulary component. Only one item, which is item 7, seems to be easier than the other items with the value of 1.5. The rest of the items are far lower than the item 7.

The reason for those low values might be that the students' lack of sociolinguistic competence as it was mentioned earlier. Without the sociolinguistic competence, which is the ability to adjust one's speech to fit the situation, even the most perfectly grammatical sentences or utterances can convey a meaning entirely different from that which the speaker intended. It is usually difficult for second language learners to acquire this competence as there is a large amount of variance in cultural rules of speaking, which means what is appropriate to say in one culture may be completely inappropriate in another culture, even though the situation in which it is said is the same. For these reasons, students are usually unaware of these differences in cultures and speeches, and they use the rules of speaking of their own native culture when communicating in the foreign language.

Another issue that should be commented on is that, the dialogue completion component aims to assess the sociolinguistic competence of the students and the questions are designed as multiple choice items. The advantages of that is, it is time saving, easier to assess and hence more reliable. It might also be helpful for the students to do multiple choice questions as the questions clearly specify the audience, the social environment in written texts

and it gives time to think about the most appropriate answer among the given multiple items. However, sociolinguistic competence is assessed better through oral exams than of multiple choice written exams as the aim of this function of language ability is to determine whether or not the students could give appropriate responses to different audience and situations in their speech. Considering these issues, some suggestions to improve this component of the English Proficiency Exam are as follows.

Suggestions;

- All the items could be revised and it could be decided on exactly what aspect of culture it is aimed to assess.
- Context of the items could be chosen among the more familiar, common topics.
- Some open-ended questions might be included in this section instead of multiple choice items only and a detailed scoring key should be provided.
- Since it is a test of general language ability, the language used in questions should be everyday language, without any specialized technical terms or culture specific vocabulary.
- Even so, in a possible way, it would still be better to use a face-to-face interview test in order to elicit the students' real sociolinguistic competence ability.

5.4. Pedagogical Implications

When developing a language test, using a well-defined language ability construct, no matter it is adapted from an existing theory of language ability or tailored to the needs of the department, may have a good and useful effect on teaching. Valid and reliable test scores can give the teachers opportunity to be aware of the language skills and areas of teaching that need to be improved. They also can make use of this information to determine how the courses should be connected from the very beginning of the semester to the end. The language instructors can have more clarity and guidance for teaching the individual courses. Well developed, valid tests can provide students with meaningful feedback on their proficiency in the language skills. Therefore, this may have a positive wash back effect on the teaching and learning of language.

As a conclusion, well-designed tests can provide results that can be used in a variety of meaningful ways. In addition, valid and reliable test scores provide fair assessments that produce meaningful results and can also eliminate bias and prevent unfair advantages by testing the same or similar information under the same testing conditions.

5.5. Recommendations for Further Research

1. Further studies should be conducted in order to collect more evidence to get more and different information about the validity of the test as this study collected only a limited amount of validity related evidence using a few methods of data analysis among plenteous available evidence.
2. This study was limited only with the listening, reading, grammar, vocabulary and dialogue completion components of the English Proficiency Exam. Including Writing and speaking components may contribute to the validity of the exam more.
3. In further studies, a deeper analysis of the scoring procedures would be useful to improve the scoring process. Inter-rater reliability should be established in order to assess the components which need to be scored by multiple raters, such as the writing and speaking components.
4. Apart from quantitative evidence, multiple methods of data collection including qualitative ones, such as interviewing or filling out questionnaires about the experience of taking, administering or scoring the English Proficiency Exam could be used.
5. Students who failed the English Proficiency Exam and had to follow the Preparatory School program could be interviewed at the end of the program to gain more insight about how they used the results of the exam and what contributions it made to them to improve their English language skills.
6. The number of seminars or workshops could be hold in order to train all the scorers so that they could be familiar with the scoring procedures, including proficiency levels and scoring key.

6. REFERENCES

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

Anastasi, A. (1988). Psychological testing. In Bond, L. A. (Ed.) *Norm and criterion-referenced testing.* Practical Assesment, Research & Evaluation. Retrived from: http://ericae.net/pare/getvn.asp?v=5&n=2.

Anastasi, A. (1988). Psychological testing. In Weir, C.J. (Ed.). *Language Testing and Validation: An Evidence-Based Approach.* New York: Palgrave Macmillan

Bachman, L. F. (1990). Fundemental considerations in language testing. In Starr, T. (Ed.) *The German proficiency exam at Brigham Young University: a validation study.* (Unpublished master's thesis) Brigham Young University, USA.

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

Bond, L. (1996). Norm- and criterion-referenced testing. *Practical Assessment, Research & Evaluation, 5*(2). Retrieved from http://ericae.net/pare/getvn.asp?v=5&n=2.

Brown, H. D. (2004). *Language assessment: principles and classroom practices.* White Plains, NY: Pearson Education.

Brown, J.D. (1989). Language Program Evaluation: A synthesis of existing possibilities. In K. Johnson (Ed.) *The Second Language Curriculum*.(pp. 222-241). London

Brown, J.D. (1995) The Elements of Language Curriculum. In Muşlu, M. (Ed.) *Formative Evaluation of a Process-Genre Writing Curriculum at Anadolu University School of Foreign Languages.* Anadolu University, Eskişehir, Turkey.

Brualdi, A. (1999). *Traditional and Modern Concepts of Validity.* Retrived from: http://www.ericdigests.org/2000-3/validity.htm

Çağ University. (2010). Mission Vision and Values. Retrieved 14 January, 2012. http://www.cag.edu.tr/en/cag.php?Kod=Genel&Alan=Genel&sayfaID=116&akategori=&menuID=UNIVERSITY&fakulte1

Chapelle, C.A. (1999). Validity in Language Assessment. *Annual Review of Applied Linguistics, 19,* 254-272. doi: 0267-1905/99.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. Routledge.

Giri, R. A. (2002). Approach to Language Testing. *Journal of NELTA, 7*(1), 15

Hughes, A. (2003). *Testing for language teachers.* Cambridge, UK: Cambridge University Press

Jones, N. (2001). Reliability in UCLES' Examinations. In Weir, C.J. (Ed.) *Language Testing and Validation*: *An Evidence-Based Approach*. New York: Palgrave Macmillan

Kelly, T.L. (1927). Interpretation of Educational Measurements. In Weir, C.J. (Ed.) *Language Testing and Validation*: *An Evidence-Based Approach*. New York: Palgrave Macmillan.

Lado, R. (1961). Language Testing. In Hughes, A. (Ed.) *Testing for Language Teachers.* Cambridge, UK: Cambridge University Press.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York, NY: Macmillan

Muşlu, M. (2007). *Formative Evaluation of a Process-Genre Writing Curriculum at Anadolu University School of Foreign Languages.*(Unpublished master's thesis). Anadolu University, Eskişehir, Turkey.

Shepard, L.A. (1993). Evaluating Test Validity. In L. Darling-Hammon (Ed.), *Review of* Research in Education, 19. Washington, DC: AERA.

Shih C. (2008). The General English Proficiency Test. *Language Assessment Quarterly, 5*(1), 63- 76.

Shuttleworth, M. (2009). *Definition of Reliability*. Retrived from: http://www.experiment-resources.com/definition-of-reliability.html

Sireci, S.G. (2007). On Validity and TestValidation. *Educational Researcher, 36*(8), 477-481.

Starr, T.G. (2008). *A Validation Study*. (Unpublished master's thesis). Brigham Young University, Brigham, America.

Suvedi, M. (2002). *Introduction to Program Evaluation.* Retrieved from http://hostedweb. cfaes.ohio-state.edu/brick/suved2.htm#Introduction

Swannell, J. (ed.) (1986). *The Little Oxford Dictionary of Current English*, 6th edition. Oxford: Clarendon Press.

The National Center for Fair and Open Testing (2007). *Criterion-and Standards-Referenced Tests.* Retrived from: http://www.fairtest.org/facts/csrtests.html

Walt, J.L., & Steyn, F. (2008). The Validation of Language Tests. *Stellenbosch Papers in Education, 38,* 191-204.

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-Based Approach.* New York: Palgrave Macmillan

Wells, C.S., & Wollack, J.A. (2003). An Instructor's Guide to Understanding Test Reliability. *Testing & Evaluation Services.* Retrived from: http://testing.wisc.edu/ Reliability.pdf

Vahra, R. (2007). *Understanding Reliability of Test Scores*. Retrived from: http://ezinearticles.com/?Understanding-Reliability-of-Test-Scores&id=755306