A MODEL FOR A PROFICIENCY/FINAL ACHIEVEMENT TEST FOR USE AT ERCIYES UNIVERSITY PREPARATORY SCHOOL

A THESIS
SUBMITTED TO THE INSTITUTE OF HUMANITIES AND LETTERS
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF ARTS
IN THE TEACHING OF ENGLISH AS A FOREIGN LANGUAGE

BY

FARUK BALKAYA

AUGUST 1994

A MODEL FOR A PROFICIENCY/FINAL ACHIEVEMENT TEST FOR USE AT ERCIYES UNIVERSITY PREPARATORY SCHOOL

A THESIS
SUBMITTED TO THE INSTITUTE OF HUMANITIES AND LETTERS
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF ARTS
IN THE TEACHING OF ENGLISH AS A FOREIGN LANGUAGE

BY

FARUK BALKAYA

AUGUST 1994

ABSTRACT

Title: A model for a proficiency/final achievement
test for use at Erciyes University Preparatory
School
Author: Faruk Balkaya
Thesis Chairperson: Dr. Dr. Arlene Clachar Bilkent
University, MA TEFL Program
Thesis Committee Members: Dr. Phyllis L. Lim, Ms.
Patricia J. Brenner, Bilkent
University MA TEFL Program

The goal of this study was to develop and pilot a model of a test based on course objectives that could be used for both a proficiency test and an achievement test for Erciyes University Preparatory School (EUPS) and that could be demonstrated to have reasonable reliability and validity. Only general English and reading skills were included in this pilot study.

This newly developed model test, the Erciyes University Proficiency/Final Achievement Test (EUPFAT, or PAT, for short), consisted of 64 open-ended items such as short-answer, sentence completion, interrogatives, and rational cloze as recommended by a number of researchers (e.g., Heaton, 1988; Hughes, 1989; Hill & Parry, 1992). No multiple-choice items were included as it has been suggested that they can produce negative backwash (Hughes, 1989). Twenty-two items testing general English skills and 42 items testing reading comprehension were included.

There were 35 intermediate-level English as a Foreign Language students attending the prep school who volunteered to pilot the PAT. Of these 35 subjects, 30 also took the English as a Second Language Achievement

Test (ESLAT) (1984), which was one of two criteria for estimating validity of the PAT. Teachers' evaluations of the 35 subjects who took the PAT were also used, as the second criterion.

Following piloting, the PAT was scored independently by two scorers using an answer key prepared by the researcher. Inter-rater reliability was .99.

The PAT was then evaluated for reliability and validity. Item analysis was also performed to identify items that should be replaced or rewritten for future administration of the tests.

For internal consistency, the split-half reliability estimate of Pearson Product-Moment Correlation adjusted for length by Spearman-Brown Prophecy Formula, the Guttman split-half reliability estimate, the K-R 20, and the K-R 21 reliability formulas were used. The reliability coefficients estimated for internal consistency using these different split-half methods ranged from .87 to .96. The descriptive statistics of the PAT are as follows: $N = 35$, Mean = 29.86, Variance = 110.89, Standard Deviation = 10.53, Sum of Item Variance = 11.93.

To determine the correlation between the PAT and the ESLAT, and between the PAT and the teacher evaluations, Pearson Product Moment Correlation (PPMC) was used. PPMC between the PAT and the ESLAT is .61, $df = 28$, $p < .0004$, and the correlation between the PAT and the teacher evaluation of subjects is .74, $df = 33$, $p = .0000$.

Item analysis of the PAT has demonstrated that if the 19 of 64 items which are lying outside the acceptable range for item difficulty and discriminability, are eliminated from the test, the rest of the test items can be used as part of a proficiency/final achievement test.

Because the total number of subjects in this study was not very high (*N* = 35), generalizing the results to other EFL situations should be avoided. However, the results of this study should be taken into consideration while developing a new test, or evaluating existing tests by those who are interested or involved in language testing.

BILKENT UNIVERSITY

INSTITUTE OF HUMANITIES AND LETTERS

MA THESIS EXAMINATION RESULT FORM

AUGUST 31, 1994

The examining committee appointed by the Institute of Humanities and Letters for the thesis examination of the MA TEFL student

Faruk Balkaya

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

Thesis title	: A model for a proficiency/final achievement test for use at Erciyes University Preparatory School
Thesis Advisor	: Dr. Phyllis L. Lim Bilkent University, MA TEFL Program
Committee Members	: Dr. Arlene Clachar Bilkent University, MA TEFL Program Ms. Patricia J. Brenner Bilkent University, MA TEFL Program

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

Phyllis L. Lim

Phyllis L. Lim
(Advisor)

Arlene Clachar

Arlene Clachar
(Committee Member)

Patricia J. Brenner

Patricia J. Brenner
(Committee Member)

Approved for the
Institute of Humanities and Letters

A. L. Karao

Ali Karaosmanoglu
Director
Institute of Humanities and Letters

ACKNOWLEDGMENTS

I am greatly indebted to my study advisor, Dr. ***Phyllis L. Lim,*** for her patience, invaluable guidance, and support throughout this study. Without these, I could have never finished this study.

I am deeply grateful to Dr. ***Arlene Clachar*** and ***Patricia J. Brenner*** for their advice and suggestions on various aspects of the study.

I am also appreciative of the administrators of Erciyes University, especially of the Rector, Prof. Dr. Mehmet Sahin, the Secretary General, Associate Professor Mümin Ertürk; to the Director of the Foreign Languages Department, Prof. Dr. Cemal Özgüven, and to the Coordinator, Dr. Mustafa Dagli, who encouraged me and gave me permission to attend the Bilkent MA TEFL Program.

My greatest debt is to my admirable parents ***Ibrahim and Gülkiz Balkaya***, my invaluable wife ***Necla Balkaya***, my newborn son Ahmet Ibrahim Balkaya, and my brothers Hüseyin and Ilhan Oguz Balkaya, who have supported me with their patience, encouragement, enthusiasm, and understanding.

My grateful thanks also go to Ahmet and Rahime Erciyes (my parents-in-law), Seyrani and Emel Sukut, Ibrahim and Halime Ciklacoskun, Recep and Hatice Azgin, and Yasar Erciyes for their kind help and support.

My most special thanks go to all of my classmates at MA TEFL, especially to Abdul Kasim Varli and Gencer Elkiliç, who helped me with my computer problems all the time, all my colleagues and staff at Erciyes University Preparatory School who helped me with administration of the tests at EUPS, Filiz Ermihan and Demet Çelebi (BUSEL Testing Office), Cem Erçin (BUSEL Resource Room), Elif Uzel and Suzanne Olcay (instructors at Bilkent Freshman), MA TEFL secretary, Seref Ortaç, and Erciyes University Preparatory School students for their participation in my pilot study.

Finally, I must express my thanks to Andrzej Sokolovsky and Gürhan Arslan for their kind assistance with computer and statistical problems.

To my

Admirable parents

Mr. ***Ibrahim Balkaya***
Mrs. ***Gülkiz Balkaya***

invaluable wife

Necla Balkaya

and
newborn son

Ahmet Ibrahim Balkaya

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

Background of the Study

In recent years only, have language tests been given the attention they require. Alderson and Buck (1993) argue for the importance of language testing and the responsibility of testers in education. The fact that language tests are often used in education to make important decisions such as determining the academic achievement level of students, the selection of students for further studies, or the suitability of students for a profession, places an awesome responsibility on test makers. Davies (1990) points out the importance of language testing in education saying, "Language testing is central to teaching. It provides goals for language teaching, and it monitors, for both teachers and learners, success in reaching those goals" (p. 1). Unless we make sure that the goals in education have been achieved, we cannot assume that the process is complete.

It is clear that tests not only affect the students but also have an impact on teaching and learning, which is called backwash by Hughes (1989). Hughes briefly explains backwash as the effect of testing on learning and teaching. Wall and Alderson (1993) suggest that a good test which reflects the aims of the syllabus will produce positive backwash and that a bad test which does not will produce negative backwash. Davies (1990) also notes that testing can have an impact on the syllabus as well as measuring progress among learners.

Two issues have been much argued concerning language testing: the method of the test and the content of the test. The term method was used by Shohamy (1984) to refer to the specific technique or procedure used in language testing to assess a trait--the knowledge being measured. Methods may be classified under two main headings: closed-items and open-ended items. Closed-items are generally multiple-choice items, whereas open ended-items include such items as sentence completion, cloze, short answers, interrogative, and essays. Types of test items (method), their strengths and weaknesses have been an issue of concern for many researchers (e.g., Henning, 1987; Shohamy, 1984; Weir, 1990).

Henning (1987) discusses two different types of testing: objective (e.g., multiple-choice tests) and subjective (e.g., writing compositions or essay types of exams). Henning agrees that for objective tests, for example, a multiple choice test, the scoring is more objective than that for a subjective test, for example, for a free, written composition; however, he claims that subjective tests may test the examinee's performance on language better than objective tests.

According to Heaton (1988), today the most commonly used testing method in objective tests is the multiple-choice type test because of its being practical--easy to score and administer, economical, and objective in scoring. However, the validity of this kind of test has been argued. Hughes (1989) says that "the advantages of the multiple-choice type of test were once so highly regarded that it was thought that the multiple-choice technique was the only way to test" (p. 60). However, he does not support this idea and claims that this kind of test may have a harmful effect on learning and teaching (i.e., negative backwash). According to Aslanian (cited in Çelebi, 1991), multiple-choice tests should not be used in assessing the comprehension of readers in general and of ESL/EFL students in particular because they (i.e., multiple-choice tests) are an inadequate means of assessment. Moreover, Heaton (1988) argues that to prepare multiple-choice items is very time consuming and that there is no relation between what multiple-choice items measure and real-life situations. He says that objective tests of the multiple-choice type encourage guessing, and adds,

> They can never test the ability to communicate in the target language, nor can they evaluate actual performance.... Appropriate responses to various stimuli in everyday situations are produced rather than chosen from several options.... The length of time required to construct good multiple-choice items could often have been better spent by teachers on other useful tasks connected with teaching and testing. (p. 27)

Hill and Parry (1992) support the open-ended type of items and discuss the difference between the multiple-choice type of test which has been developed in the United States and the open-ended style which has been developed in Britain. Hill and Parry suggest that because the British-style open-ended test requires the test taker to draw on productive skills, the open-ended style should be used in language tests.

Weir (1990) compares the advantages and disadvantages of open-ended, short-answer types of items. According to him, short-answer questions are a question technique which refers to questions requiring the subjects to write the specific answers in the specific places determined by the test makers on the question or answer paper. He says that the short-answer test type is an extremely useful technique for testing both reading comprehension and listening comprehension. The only disadvantages he mentions are that these kinds of test items require students to write, which may interfere with the measurement of the intended construct, and that variability of answers in scoring might lead to marker unreliability.

Weir (1990) also argues in favor of the "selective deletion gap filling" (p. 48) or rational cloze. To Weir, because of the negative findings in recent studies on mechanical deletion cloze, there has been a tendency to support the use of a rational cloze, or selective gap filling, in which the test makers select the items to be deleted based on what he or she wants to measure.

Although some researchers (e.g., Heaton, 1988) argue in favor of the sentence completion type in which the testees are required to supply a word or a short phrase, some researchers such as Kirshner, Wexler, and Spector-Cohen (1992) suggest that the easiest format for students is the interrogative form.

The other issue which has been argued a lot is the content of the tests. Some researchers (e.g., Finocchiaro & Sako, 1983) suggest that in achievement tests, the test items should be based on the course content, whereas some researchers, for example, Hughes (cited in Hughes, 1988) and Wiseman (cited in Hughes, 1988), suggest that test items be based on the objectives of the course. Hughes argues for the objective-based achievement tests, saying that "one function of testing is to provide the kind of information that will help keep its partner on the right track. It can best do this when achievement test content is based not on the syllabus or textbooks but on course objectives" (p. 42).

Statement of the Purpose

In Turkey, at most of the English-medium universities and semi-English-medium universities, there are one or two-year preparatory programs to teach new entrants English so that they can do their undergraduate studies. Erciyes University is one of the Turkish universities which has a one-year preparatory school--Erciyes University Preparatory School (EUPS). Every year there are approximately 600 new students studying at EUPS. Most of these students take the proficiency test at the beginning of the term. Those who are not successful in the proficiency test have to attend the preparatory school. The objectives of the preparatory school at Erciyes University are "to prepare students so that they can study in their undergraduate classes, which are given in English, read and understand the issues published in English related to their subjects, and write answers in English to essay type questions" (M. Dagli, personal communication, April, 1994). However, in the current teaching system, the teaching is not carried out based on the course objectives; grammar accuracy is given primary importance by the instructors and the students. When the researcher had informal conversations with instructors at EUPS, the common point shared by almost all of the instructors was that almost all the instructors give primary importance in the classes to grammar accuracy rather than course objectives.

At Erciyes University Preparatory School there is a unit called the Testing Office, which is in charge of developing, administering, and scoring all the tests. The researcher knows the situation of the testing office very well because he worked there for two years. This emphasis on grammar mentioned above leads the Testing Office to focus on preparing grammar-based tests. In other words, the testing office prepares all the tests, including the final achievement test, based on the *content* of the courses rather than based on objectives. Because grammar accuracy is given much importance in this program and, subsequently, in the tests, both teachers and students focus mainly on grammar accuracy and little on reading comprehension; the other skills such as writing and listening are largely ignored by the instructors and students most of the time. Most of the test items are multiple-choice types of questions; a few of them are true/false, and occasionally there are cloze passages in which the options are given in multiple choice format. All these multiple-choice or true/false type of test items encourage students to *recognize* only rather than to *produce*.

In addition to problems mentioned above, at Erciyes University Preparatory School, neither final achievement tests nor proficiency tests have ever been evaluated for *reliability* or *validity*, nor has *item analysis* been done. Reliability is defined by Harris (1969) briefly as "How well does the test measure?" (p. 19). Anastasi (1988) defines reliability as "the consistency of scores (that would be) obtained by the same persons when reexamined with the same test on different occasions, with different sets of equivalent items, or under other variable examining conditions" (p. 109). According to Harris, validity means "What precisely does the test measure" (p. 19). Henning (1987) argues the importance of reliability for tests used for admission to universities, saying that "examinations that serve as admissions to university, for example, must be highly reliable" (p. 10). Henning also emphasizes the importance of validity of tests: "We should ascertain whether or not the test is valid for its intended use" (p. 10).

As well as reliability and validity, item analysis of tests is also important. According to Henning (1987), in most of language testing we are concerned with three issues: the writing, administration, and analysis of appropriate items--item difficulty, and item discriminability. For Henning, item difficulty (item facility, Oller, 1979) is, perhaps, the most important characteristic of an item which has to be accurately determined. Tests which consist of too difficult or too easy items for a given group of testees often show low reliability coefficients. Another important characteristic of an item to Henning is its discriminability index, that is, how well an item discriminates a strong student from a weak student. Henning claims that the item discriminability index is as important as item difficulty. When determining whether to include or exclude an item, both item difficulty and item discriminability should be taken into consideration. For Henning, the item difficulty index does not give sufficient information to make the ultimate decision to choose or reject an item. If the item difficulty is reasonable and it discriminates good and bad testees, that item is regarded as an ideal item. In fact, according to Henning, all these psychometric features (reliability, validity, item difficulty and item discriminability) interact with each other in some way.

The purpose of this study was to develop a model examination based on the objectives of the program using open-ended test methods which can be used for both final achievement and proficiency measurement functions, and after piloting, can be demonstrated to have reasonable reliability and validity. The test was also evaluated for item difficulty and item discriminability and recommendations were made for improving the test based on these results. The reliability of both test scoring and the test itself were evaluated, that is, inter-rater reliability which means "the correlation of the ratings of one judge with those of another" (Henning, 1987, p. 82), and internal consistency. Interrater reliability was determined by Pearson Product-Moment Correlation. Internal consistency was determined by the ordinary split-half method adjusted for length by the Spearman-Brown Prophecy Formula, the Guttman split-half estimate, which according to Bachman (1990), "provides a direct estimate of the reliability of the whole test" (p. 175), the K-R 20 and the K-R 21 reliability formulas. For test validity, two external criteria--teacher evaluations of students and the English as a Second Language Achievement

Test (ESLAT)--were used to evaluate concurrent (criterion) validity. According to Henning (1987), validating a test against concurrent cretirion involves the following:

> One administers a recognized, reputable test of the same ability to the same persons concurrently or within a few days of administration of the test to be validated. Scores of the two different tests are then correlated using some formula for the correlation coefficient and the resultant correlation coefficient is reported as a concurrent validity coefficient". (p. 96)

Correlations between the new developed Erciyes University Proficiency/Final Achievement Test (EUPFAT), or (PAT) for short, and the ESLAT and between the PAT and teacher evaluations were determined using Pearson Product-Moment Correlation (PPMC). Item analysis of the PAT, including both item difficulty and item discriminability, was done. Recommendations for inclusion or exclusion of items were made.

In this study, because of time constraints, only general language proficiency and reading were included. Listening and writing sections were not included.

CHAPTER 2 REVIEW OF THE LITERATURE

Introduction

In Chapter 1, the importance of language tests, their impact on learning, and teaching and various issues related to the method of language tests were briefly discussed.

This chapter gives a detailed review of related literature. The issues discussed in this chapter are (a) approaches to language testing, (b) types and purposes of language tests, (c) direct versus indirect tests, (d) subjective versus objective tests, (e) importance of backwash or washback effect of tests, and (f) basic qualities of a good test--reliability and validity.

Approaches to Language Testing

According to Baker (1991), before the Second World War, language testing was not usually a distinct activity, that is, was not looked at as a separate activity in language teaching. If teachers had to assess language proficiency, they would do it by using the same methods used when they were teaching the language, such as having students write a composition, do a translation, or take a dictation.

Madsen (1983) and Spolsky (cited in Grotjahn, 1988) discuss the history of language testing in three phases, whereas Heaton (1988) discusses it in four stages. Although the number of and names given for each stage, or phase, are somewhat different, actually, their meanings are very close. The common point concluded by these three researchers is that language testing in the first stage, which is called "intuitive or subjective stage" by Madsen (1983, p. 5), "pre-scientific or traditional" by Spolsky (cited in Gtortjahn, 1988, p. 159), and the "essay-translation approach" by Heaton (1988, p. 15) was largely based on subjective evaluation. Grotjahn (1988) argues that the examinations of this stage have been criticized for not having objectivity and reliability. The second stage was called the "scientific era" (Madsen, p. 6), the "psychometric-structuralist or modern phase" (Spolsky, p. 159) and the "structuralist approach" (Heaton, p. 15). As Madsen said, the objective tests were devised so that tests could measure students' performance on recognition of sounds, specific grammatical points, or vocabulary items. Also, in this stage, subjective tests began to be replaced by objective tests due to the fact that, according to Madsen, objective tests were easily and objectively scored by even an untrained person. In the third stage, a new vogue has appeared in which, according to Madsen, the emphasis is given to evaluation of language use rather than to language form. This era is called the "communicative stage" (Madsen, p. 6), the "psycholinguistic-sociolinguistic or post-modern phase" (Spolsky, p. 159), the "integrative approach" (Heaton, p. 16), and the "communicative approach" (Heaton, p. 19).

Heaton points out that in the communicative approach, which is sometimes linked to the integrative approach, importance is given to the meaning of utterances rather than form and structure.

According to Savignon (1991), communicative language testing has come to the scene along with the development of communicative language teaching which, Savignon says "puts the focus on the learner [and in which] learner communicative needs provide a framework or elaborating program goals in term of functional competence" (p. 266).

Although the names for each stage, or phase, are different, a common point shared by Spolsky (cited in Grotjahn, 1988), Heaton (1988), and Madsen (1983) is that a good test will frequently combine features of all there approaches.

Types and Purposes of Language Tests

Language testing has a major role in language teaching. According to Alderson, Krankhe, and Stansfield (1987), language tests are used for different purposes such as determining proficiency, achievement, placement, and diagnosis. Proficiency tests are administered to determine if candidates are proficient enough to do something such as to perform a certain job. (For example, at English-medium, or semi-English-medium universities in Turkey, proficiency tests are given at the beginning of the term to determine if students have the required proficiency to do their studies in their faculties.) Achievement tests, on the other hand, are defined by Alderson et al. as tests which are given after instruction and measure a student's success in learning the given instructional content of a course. According to Hughes (1989), achievement tests are divided into (a) progress achievement tests which are administered during the instruction to determine if the students are making desired progress through the instruction and (b) final achievement tests which are administered after the instruction to find out if the course has served its purpose, that is, if the candidate learned what he was expected to. Diagnostic tests, according to Hughes, are administered to find out students' weaknesses and strength and to ascertain what further teaching is necessary. Placement tests are administered to classify students according to their proficiency levels before instruction begins (for example, at preparatory schools to form groups in which the students have almost same proficiency level).

The relationship between proficiency and achievement tests (especially final achievement tests) has been a topic of discussion in the literature. There have been different ideas about the relationship between achievement and proficiency tests. Some researchers (e.g., Alderson et al. 1987; Finocchiaro & Sako, 1983; Henning, 1987) put a clear cut distinction between achievement tests and proficiency tests and say that achievement tests have to be based on the content of the syllabus. Finocchiaro & Sako point this out as follows:

> They [achievement tests] may be said to represent the incremental progress made by a student in a language course between two points in time. These tests are designed *solely to cover material that has been presented in the classroom* [italics added] (or in a language laboratory) during that period, but not the material which represents the total corpus of the foreign language and which therefore may not have been taught. (p. 15)

On the other hand, some researchers suggest that achievement tests be based not on the course content but on objectives. As Wiseman (cited in Hughes, 1988) states:

> The syllabus content approach tends to perpetuate ineffective educational practices; it is a reactionary instrument helping to encapsulate method within the shell of tradition and accepted practice...[but] the goal-oriented test is exactly the opposite; it evaluates learning--and teaching--in terms of the aims of the curriculum, and so fosters critical awareness, good method, and functional content. (p. 40)

Hughes (1989) also shares the same idea as Wiseman. For Hughes, if a test is based on the course objectives, and these objectives are equivalent to real language needs (just as would be expected in a proficiency test), then there is no reason to put a distinction between the form and content of these two types of tests, that is, proficiency and final achievement tests. Additionally, Hughes argues that if a test is not based on the course objectives but on the content of a poor or inappropriate course, the students who take the test will be "misled as to the extent of their achievement and quality of the course" (p. 12); however, if a test is based on the course objectives, it will give more useful information about the achievement of students and how well the objectives have been reached.

Direct versus Indirect Tests

Language tests can be administered either directly which is termed "authentic tests" by Shohamy and Reves (1985), or indirectly. According to Henning (1987), it is said that if a test is testing language use in real life situations, that test is testing language performance directly; otherwise, it is testing indirectly, in such tests as multiple-choice recognition tests. To Henning, "an interview may be thought of as more direct than a cloze test for measuring overall language proficiency. A contextualized vocabulary test may be thought more natural and direct than a synonym-matching test" (p. 5). Many language tests can be viewed as lying somewhere between these two points. In a study done in Egypt to devise an examination that could be used for both achievement and proficiency measurement of students, Henning, Ghawaby, Saadalla, El-Rifai, Hannallah and Maffar (1981) used the multiple-choice method for its being objective and less time consuming to score. They also assume that the multiple-choice items in reading comprehension would be more valid saying that "the response themselves are more valid in that they entail reading and recognition rather than writing" (p. 466). However, in A Guide to Language Testing (1987), Henning argues that because multiple-choice recognition tests are indirectly tapping true language performance, they are less valid for measuring language proficiency than direct tests.

Subjective versus Objective Tests

The terms subjective and objective seem to cause some confusion and attempts to clarify them can be seen in the literature. Heaton (1988) clarifies the terms subjective and objective as terms which should be used only to refer to the scoring of tests. In objective tests, scoring is very easy. Madsen (1983) says that they can be scored even by a person who is not trained in testing. According to Henning (1987), an objective test is "one that may be scored by comparing examinee responses with an established set of acceptable response or scoring key...a common example would be a multiple-choice recognition test" (p. 4). Henning shares the same idea as Madsen about the scoring of objective tests saying that no particular knowledge or training is required on the part of the scorer of this kind of tests. Subjective tests, on the other hand, "require scoring by opinionated judgment, hopefully based on insight and expertise, on the part of scorer. An example might be the scoring of free, written compositions" (p. 4). There is a common belief among instructors and students that only objective tests such as multiple-choice tests, are reliable and dependable, and that the other type of tests, subjective tests, are unreliable and undependable. Henning tries to correct this wrong assumption, saying that "the possibility of misunderstanding due to ambiguity suggests that objective-subjective labels for tests are of very limited utility" (p. 4).

According to Weir (1990), the multiple-choice test method, an objective type of test, has advantages and disadvantages. He concludes that multiple-choice tests have high rater reliability but that to prepare multiple-choice items in objective tests is very time consuming, whereas open-ended test method have low rater reliability and scoring the open-ended items is very time consuming. However, according to Heaton, all kinds of tests have some degree of subjectivity. Heaton argues that all test items require candidates to exercise subjective judgment, and adds that all kind of tests are constructed subjectively by the tester, who determines which areas of language are to be tested, how those particular areas will be tested, and what kind of items will be used for this purpose. Thus, according to Heaton, only the scoring of a test can be described as objective because the score will not change when the test is re-rated by another rater.

The disadvantages of subjective tests (the problem of subjectivity in scoring and their being time consuming to score) are generally accepted by researchers such as Heaton (1988), Henning (1987), and Weir (1990).

However, there is no agreement that objective tests are "good", and subjective tests are "bad." Heaton suggests that because it is difficult to achieve reliability, due to many different degrees of acceptable answers, "careful guidelines must be drawn up to achieve consistency" (p. 26). He claims that an objective test can also be a very poor and unreliable test. Objective tests are also criticized by researchers (Heaton, 1988) on the grounds that (a) they require far more careful preparation than subjective tests, (b) they are simpler to answer than subjective tests, (c) item difficulty can be made as easy or as difficult as the test constructor wishes (by analyzing student performance on each item and rewriting the items where necessary), and (d) multiple-choice objective tests encourage guessing. Hughes (1989) states his concerns about the use of multiple-choice objective tests as follows:

> If there is a lack of fit between at least some candidates' productive and receptive skills, then performance on a multiple choice test may give a quite inaccurate picture of these candidates' abilities.... The chance of guessing the correct answer in a three-option multiple choice item is one

> in three, or roughly thirty-three per cent. On average we would expect someone to score 33 on a 100-item test purely by guesswork. (p. 60)

Importance of Backwash (or Washback) Effect of Tests

There is no doubt that every language test has backwash effect on teaching and learning. Davies (1990) says that "testing always has a washback influence and it is foolish to pretend that it does not happen" (p. 25). In discussing the relationship between testing and teaching, he says that the kind of testing and the aspects of language which are tested find their ways into teaching situations: If grammar is tested, then grammar will be taught; if speaking is tested, then spoken [English] will be taught; if speaking is not tested, then it will not taught at all. He suggests that because the backwash effect of tests in teaching is inevitable, we should take advantage of this backwash effect in such situations where change has been very slow. Whereas Wall and Alderson (1993) say that "language tests are frequently criticized for having negative impact on teaching--so-called negative washback" (p. 41), according to Davies, what is important in backwash effect of testing is whether it is _beneficial_ or _harmful_.

Davies (1990) and Hughes (1989) argue that tests should have a beneficial backwash on learning and teaching. Davies states that "washback is so widely prevalent that it makes sense to accept it, to stop regarding is as negative, and then make it as good as it can be in order to improve its influence to the maximum" (p. 1). Hughes suggests that what we should demand of testing is that testing should be supportive of good teaching and where necessary have a corrective role on bad teaching. If the tests are based on course objectives and the testing technique is measuring the students' knowledge objectively, tests will have beneficial backwash effect. In a study of the development of a new test and its backwash effect at Bogaziçi University, one of the most important and reputable universities in Turkey, Hughes (1989) reports that the new test that was prepared based on the objectives of the courses, determined through a questionnaire, resulted in students' reaching a much higher standard in English than had been reached in the history of the university.

Pierce (1992) suggests that the TOEFL-2000 (Test of English as a Foreign Language) test development team at Educational Testing Service, in the process of reviewing the test, "needs to address the washback effect of the test in consultation with both ESOL teachers and TOEFL candidates internationally" (p. 665).

Wall and Alderson (1993), in a study to investigate the impact of a new examination on English language teaching in secondary schools in Sri Lanka, suggest, based on the results they got, that "testers need to pay much more attention to the washback of their tests" (p. 68).

However, although the importance of positive washback should not overlooked, there are basic criteria which must be met for any test to be considered a "good test" in any sense. These basic criteria are reliability and validity. According to Harris (1969), in test evaluation or test selection, generally reliability and validity should be considered.

Basic Qualities of a Good Test: Reliability and Validity

Unless tests have reliability and validity, it is hard to be sure if the test really serves the purpose it is intended to. As Harris (1969) said, while selecting a test for any purpose, two important questions must always be kept in mind: "(1) What precisely does the test measure? and (2) How well does this test measure?" (p. 19). These questions address validity and reliability, respectively.

Anastasi (1988) defines reliability as "consistency of scores [that would be] obtained by the same person when reexamined with the same test on different occasions or with different sets of equivalent items, or under other variable examining conditions" (p. 109), and according to Henning (1987), reliability is "a measure of accuracy, consistency, dependability, or fairness of scores resulting from administration of a particular examination" (p. 74).

Reliability in tests has always been considered very important. There are different kinds of reliability estimates such as test-retest, parallel forms, and internal consistency (split half reliability). Inter-rater reliability refers to consistency of scoring when the test scores are independent estimates by two or more judges or raters. In this kind of reliability estimate, correlation of the ratings of one judge with the ratings of others is computed. The test-retest method is the most direct way of calculating test reliability. The same test is readministered to the same group within a certain period of time, no more than two weeks, and Pearson correlation of two sets of scores is calculated. For the parallel forms method, there must be two independent but equal tests administered to the same sample of subjects whose scores are correlated using Pearson correlation. When test-retest or the parallel forms method are not possible, the best way to estimate reliability is to use one of the internal consistency, or split-half reliability, estimates. There are different methods for obtaining split-half reliability estimates such as the ordinary, or traditional, split-half Pearson Product-Moment Correlation adjusted for length by the Spearman-Brown Prophecy Formula, the Guttman split-half reliability estimate, and two the Kuder-Richardson formulas (K-R 20 and K-R 21). In splitting the halves for the traditional split half, sometimes special care must be given in order to select items that are independent of each other for each half (Anastasi, 1988). Bachman (1990) says that splitting the test into possible halves should be done in every possible way. However, Bachman points out its difficulty in real life situation saying that, "with a relatively short test of only 30 items we would have to compute the coefficient for 77,558,760 combinations of different halves" (p. 176). This need led to the development of the K-R 20 and K-R 21 formulas which yield reliability coefficient which are averages of all possible split halves.

In addition to reliability, validity is also an important characteristic of language tests to be taken into consideration. Although researchers such as Bachman (1990), Celce-Mercia (1990) and Henning (1987), and claim that any test must be reliable before it can be valid, this does not mean that reliability is important and validity is not important. For example, Raatz (1985) says that the most important characteristics of a test is its validity. There are different kinds of validity estimates used in testing, among which are content validity, criterion-related validity, response validity, and face validity.

According to Davies (1990) content validity is a professional judgment of the teacher or test maker (that the content is appropriate for testing whatever test is supposed to measure), and to Hughes (1989), in discussing language tests, it means that the test is a representative sample of the language skills, structures and so forth which the test is meant to measure. Hughes claims that "the greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure" (p. 22). As for criterion validity, Hughes says that there are two essential kinds of criterion-related validity: concurrent and predictive. Davies says that concurrent validity is "based on a measure that is already at hand, usually another test [known or assumed to be valid]" (p. 24). Concurrent validity is ordinarily determined by administering the test and criterion at about the same time, which can be in a few days time (Henning, 1987). Hughes argues that a test may be validated not only against another test, but also against the teachers' assessments of their students provided that the teachers' assessment of the students can be relied on. The other criterion-related validity is predictive validity which concerns the degree to which a test can make predictions about the candidates' future performance (Hughes). Henning defined the term response validity as the extent to which examinees have responded to the items in the way the test developers expected. The last type of validity is face validity. According to Magnusson (cited in Henning, 1987), content validity and face validity are synonyms. However, some researchers differentiate between content and face validity. For them, face validity is determined by asking the examinees whether the exam they took was appropriate to their expectations (Henning).

There have been a lot of discussions about the relative importance of reliability and validity. However, Bachman (1990) suggests that we should recognize them as "complementary aspects of a common concern in measurement" (p. 160). To him, reliability is concerned with answering the question, "How much of an individual's performance is *due to measurement error* [italics added], or to factors other than the language ability we want to measure?", and validity is concerned with the question, "How much of an individual's test performance is *due to the language abilities* [italics added] we want to measure?" (p. 161).

The issues in the literature discussed in this chapter have demonstrated that tests' formats, their contents, their backwash effect on learning and teaching, and their psychometric features such as reliability, validity, item difficulty, and item discriminability are very important and should be given the necessary importance. Tests are administered in education to make important decisions. Before making these decisions, the decision makers must be sure of the test itself. At EUPS, nearly all of the items in the current proficiency and final achievement tests are multiple-choice items which measure recognition rather than production, the content of the progress and final achievement tests are based on the course content rather than course objectives, the backwash effect of tests have not been taken into consideration, reliability and validity evaluation of final achievement tests and proficiency tests have not been assessed, nor have the item difficulty and item discriminability of the final achievement and proficiency tests been determined. In this study, the researcher developed the PAT in open-ended format because it would test production rather than recognize only. The content of the PAT was determined taking the course objectives into consideration hoping that future tests modeled on this test will have beneficial backwash effect on education. The reliability and validity of the PAT were estimated, assuming that in making important decisions about individual's life in future, we must be sure of the instrument we use. And consequently, item analysis was also done in order to recommend how the test can be improved.

CHAPTER 3 METHODOLOGY

Introduction

As mentioned in the previous chapters, language testing has long been an integral part of education. Through the history of language testing, different testing methods have been used, each having its advantages and disadvantages. Because tests often have a great impact on people's lives, the development of reliable and valid tests is very important. In this chapter, the researcher will give detailed information about subjects, materials and procedures used in this study.

Background

New students are accepted to Erciyes University through a university entrance examination. Medical faculty students must be good at mathematics and science, whereas the students of the Faculty of Economics must be good at Turkish and mathematics. New entrants to these two faculties must then take an English proficiency test, the Erciyes University Proficiency Test, for exemption. Those students who are successful in this exam begin their first year studies without attending the preparatory school; the students who fail in the exam take a placement test to be classified into three groups: Group-A (beginning), Group-B (intermediate), and Group-C (upper-intermediate) according to their proficiency levels. There are 13 groups (from Group-A through Group-

M). Every three weeks students take an achievement test. If a student is successful, he or she moves up a level. Each student has to finish at least Group-J to be eligible to take the final achievement test.

Subjects

Out of about 100 intermediate EFL students attending EUPS who were asked, with the permission of their class teachers and administration, if they would voluntarily participate in a study, 37 students ranging from 17-20 in age volunteered. These subjects were to take their final achievement test two weeks after they participated in this study. All the subjects in this study were J-level intermediate students; therefore, all the subjects were assumed to be eligible to take the final achievement test.

Of the 37 subjects in this study, 2 refused to continue shortly after the study began; thus, 35 subjects completed the PAT. The 2 subjects who refused to continue were dropped from the study.

Materials

The materials in this study included two exams, the model PAT and the ESLAT, and teacher evaluations of subjects.

Erciyes University Proficiency/Final Achievement Test (PAT)

This testing material consisted of consent page, general instructions, the body of the test, and the answer sheets. (see Appendixes A, B, & C).

The first page consisted of information which explained the aim of the study and informed participants that the test would not have any effect on their scores at preparatory school and that their identity would be confidental and a consent form. Students were asked to write their name, student number, and date and to sign to demonstrate that they understood that they were volunteers and could withdraw at any time.

The body of the PAT test consisted of seven reading passages selected from different textbooks along with instructions for different part of the test. According to Hughes (1989), types of texts to be asked in reading comprehension might include academic journals, textbooks, magazines, newspapers, and newspaper advertisements. The texts used in the model test included passages from academic textbooks and newspaper articles from second language textbooks. The texts were determined in consultation with the study advisor; Filiz Ermihan, who is in charge of the testing office at Bilkent University School of English (BUSEL); Cem Erçin (BUSEL Resource Room); some of the MA TEFL students; and instructors teaching at Erciyes University. After the texts were determined, the researcher began to write items. While he was writing the items, he took into consideration the recommendations [about test construction] made by Hughes (1989), the study advisor, Filiz Ermihan, Elif Uzel (instructor in English at Bilkent Freshman), and Suzanne Olcay (a native speaker of English and instructor in English at Bilkent Freshman). The body of the test consisted of two parts: Part One--Use of English Section and Part Two--Reading Comprehension Section. Almost all items in both Part One and Part Two were on the open end of the closed-to-open-end continuum of item types. No completely closed types such as multiple choice were asked. Only the items for Passage 1 in Part I offered the subjects choices from which to select the answers.

Part One, the Use of English Section, consisted of two different passages. The passage length for the first text was 256 words. There were 12 blanks in the text and there were 14 choices given in a box. Subjects were asked to fill in the blanks with an appropriate word or phrase given in the box. The length for the second passage was 257 words. Subjects were asked to fill in the blanks using the text only (rational cloze); the researcher deleted a content <u>or</u> a function word in almost every sentence. For this passage, the subjects were not supplied options. The total number of items in Part One was 22: 12 related to the first passage and 10 related to the second passage. In each passage, the missing words were both content and function words to measure students' knowledge in real life situations. Subjects were required to understand the whole sentence and also to have a general idea about the whole text to find the right words to fill in the blanks.

Part Two, the Reading Comprehension Section, consisted of five different reading passages. The passage lengths were, in order, 353, 464, 334, 673, and 786 words. Hughes (1989) suggests that "in order to increase reliability, include as many passages as possible in a test" (p. 119).

Because the objective of the PAT was to measure students' reading comprehension, "macro-skills" were given primary importance and "micro-skills" (subskills) were given secondary importance (Hughes, 1989; Lumley, 1993). Macro skills were tested with (a) sentence completion (e.g., for main idea: "Many of the worlds greatest writers have been concerned with whether _______________ can be avoided.", (b) short-answers (e.g., for scanning: "What date did the fire break out?"), and (c) interrogatives (e.g., for scanning: "Why is it difficult to know very accurately how much the number of people in the world changes? _______________"). The micro-skill items included (a) referents (e.g., "The word "it" in line 16 refers to _______________ ") and (b) guessing meaning of unfamiliar words (e.g., "What one word in paragraph (5) means nearly the same as ***unchanging*?").**

In Part Two, the Reading Comprehension Section, there were 42 items in total: 12 of the items were related to micro-skills, and 30 of the items were related to macro-skills.

A separate answer sheet which consisted of two pages was prepared by the researcher (see Appendix C). The subjects were cautioned to write all the answers only on the answer sheet; the answers on the test itself would not be scored at all.

English as a Second Language Achievement Test (ESLAT)

The second test, used to validate the PAT for concurrent (criterion) validity, was the English as a Second Language Achievement Test (ESLAT), which consisted of 65 multiple-choice items, each having 5 options, or distractors. The time allotted for this test was 45 minutes. This test was presumed to be reliable and valid.

Teacher Evaluation of Subjects

All the teachers who were currently teaching the subjects who participated in the study were asked to evaluate the examinees on a 1-to-5 scale, in which 1 meant the lowest level of proficiency and 5 meant the highest level of proficiency. The same scale was used by Demet Çelebi (1991). (see Appendix D)

Procedure

Development of the PAT

While developing the test, the researcher first piloted it with 6 MA TEFL students. Taking the recommendations made by the testees, he revised the test, dropping some of the passages and items and added new ones. The new test was piloted 3 other MA TEFL students and 1 instructor who is in charge of the BUSEL Testing Office and 1 instructor of English at Bilkent freshman. The researcher took the recommendations made by these testees and once more revised the test, omitting or revising some items and texts. This newly revised test was piloted with 2 instructors at Bilkent Freshman, 1 of whom is native speaker of English, and 1 of whom is non-native speaker of English. The researcher revised the test according to the suggestions made by these testees. Subsequently, the test was checked by the study advisor. Final revisions of the test were made according to her suggestions. The test was piloted with 4 intermediate students at EUPS (two weeks before the administration of the test) without informing them that some other students in the same program would take the same test. Henning (1982) suggests that piloting on the target populations is useful for (among other things) checking to see if time limits are appropriate.

The duration of the test was determined by this pilot test--150 minutes.

Administration of the Test

This was a two-part study on testing. Students were allowed 150 minutes to complete the new model test, the PAT. Two days later, the second test, the ESLAT, was administered, taking 45 minutes. Instructions, distribution, and collecting of papers are not included in these times. Each test was administered in four different classes. In each class, there were approximately 9 subjects and 1 proctor (the class teacher). Each test was administered in one session during the students' usual class hours with regular class teachers.

Hughes (1989) claims that even the best test may result in an unreliable and invalid outcome if it is not well administered. The researcher attempted to follow Hughes's recommendations for test administration. Just before starting the exam, all the students were informed about the test and requirements following the Hughes' suggestions. The followings procedures were followed by each class teacher (the proctor):

a) Tests were delivered to the subjects and they were asked not to start until they were told to do so;

b) Answer sheets were delivered to the subjects;

c) Students were asked to check the test booklet to find out if there were any missing pages;

d) When the exam started, the proctors wrote the time the exam started and when it would finish on the blackboard;

e) The subjects were also informed about procedures mentioned above in Turkish to make sure that all the subjects understood.

As the subjects finished the test, they were allowed to leave the exam room. At the end of the time allotted, the proctors collected the test and answer sheets separately and handed them to the researcher. All the subjects were able to finish the test within the allotted time.

Scoring and Data Analysis Procedure

All the exam papers were scored by two teachers from an answer key prepared by the researcher. Each scoring was carried out independently. The raters were told that correct items should be given 1 point and wrong items should be given 0 points. The raters were informed about grammatical mistakes [whether they (grammatical mistakes) would be counted wrong or correct, or would be given partial credit]. The researcher recommended to the raters that grammatical mistakes be ignored, and that partial credit not be given. However, one rater counted grammatical mistakes as errors and counted the item as wrong. After the raters finished scoring, the researcher recorded the scores and then compared the scores of these two raters. In 23 of the papers, there were discrepancies between final scores (1 to 4 points). The researcher witnessed that the discrepancies were either due to wrong computing of the scores (i.e., the scores were added up incorrectly) or due to failure to evaluate some of the items. The researcher rescored those papers and gave the final scores. These final scores were the scores used to determine internal consistency and validity of the test. These were also the scores on which item analysis was performed. There was no partial credit in any of the items.

After each rater finished scoring, the researcher evaluated the PAT for inter-rater reliability (on the original set of raters' scores), internal consistency, and validity as well as determining item difficulty and item discriminability.

CHAPTER 4 DATA ANALYSIS

Introduction

As mentioned above, language testing has been an integral part of education and has influenced important decisions. The format of tests has been debated. Although today the multiple-choice test technique is very often used, its validity has been argued by researchers (e.g., Heaton 1988; Hughes, 1989; Oller, 1979). Some researchers (e.g., Oller, 1979; Hughes, 1989) recommend that we not use the multiple-choice test technique. Heaton (1988) accepts the advantages of the multiple-choice test technique; however, he claims that the usefulness of this type of item is limited. In this study, assuming that the open-ended test technique would test students' true language performance better and more directly (i.e., would encourage students to produce rather than recognize), an open-ended test technique was used.

Reliability and validity are regarded as two complementary characteristics of tests (Bachman, 1990). Along with measuring reliability and validity, item analysis (item difficulty and item discriminability) also has been considered important. Henning (1987) suggests that both item difficulty and item discriminability should be taken into consideration when determining whether to accept or reject an item.

Data Analysis

Reliability

Bachman (1990) and Henning (1987) argue the importance of reliability and claim that without first being reliable, a test cannot be valid at all. The first issue the researcher addressed was evaluating the test for reliability. However, before evaluating the reliability of the test itself, the reliability of the scoring was estimated.

Inter-rater reliability. In order to see if there were any discrepancies between the two raters, the researcher first obtained the inter-rater reliability estimate by correlating the two raters' original scores using Pearson Product Moment Correlation (see Appendix E). The inter-rater reliability is very high, $\underline{r}_{AB}$ = .99.

Internal consistency: split-half reliability estimates. Since test-retest reliability was not feasible on this occasion, the researcher used internal consistency methods (split-half reliability estimates). The researcher estimated the reliability using the "usual" Pearson Product Moment Correlation (adjusted for length with the Spearman-Brown Prophecy Formula and the Guttman split-half reliability estimate in two ways, that is, for two different types of splits. In addition, Kuder-Richardson (K-R) 20 and K-R 21 formulas were also employed to estimate reliability.

First the researcher split the test items in two ways. The first split was the traditional odd-even number split (ODE); however, in the second split, care was taken to split items in such a way as to preserve independence of items across the two halves as much as possible (IND). Anastasi (1988) suggests that care be taken to keep items in one reading passage (presumably dependent on each other) together on one side of the split. This was done as much as possible in this split. The reliability coefficient estimated with Pearson Product-Moment correlation, adjusted for length with the Spearman-Brown Prophecy Formula for ODE, is .96. The reliability coefficient estimated using the Guttman split-half reliability estimate for ODE is .96 (for more information, see Appendix F).

The reliability coefficient estimated with Pearson Product-Moment Correlation, adjusted for length with the Spearman-Brown Prophecy Formula for the second split-half method (IND) is $r_{IND} = .88$, and reliability coefficient estimated using the Guttman split-half reliability estimate for IND is $r_{IND} = .87$ (for details, see Appendix G).

The other two reliability coefficient estimates used for internal consistency were K-R 20 and K-R 21.
The reliability coefficient of the PAT estimated using K-R 20 is .91, and the reliability coefficient of the PAT using K-R 21 is .88. The data used to estimate K-R 20

reliability estimate are $\Sigma_i^2 = 11.93$ and $\Sigma_t^2 = 110.89$; the data used to estimate K-R 21 are $\underline{M} = 29.86$, and $\Sigma_t^2 = 110.89$.

Validity

After the reliability of the PAT was evaluated, the researcher evaluated the test for concurrent (criterion) validity using two external criteria, the ESLAT and the teacher evaluations of the subjects. Pearson Product Moment Correlation (PPMC) was used to determine the correlation between the PAT and the ESLAT and the correlation between the PAT and the teachers' evaluation (for calculations for correlation between the PAT and the ESLAT, and between the PAT and teacher evaluations, see Appendix H and Appendix I).

The Pearson correlation between the PAT and teacher evaluations is $\underline{r} = .74$, $\underline{df} = 33$, $\underline{p} = .0000$, and the correlation between the PAT and the ESLAT is $\underline{r} = .61$, $\underline{df} = 28$, $\underline{p} < .0004$. Oller (1979) says that the higher the correlation, the more informative it is. Because the test methods in the PAT and in the ESLAT were different, and particularly because in the ESLAT, some of the students ($\underline{n} = 10$) gave double answers to some of the items (ranging from 2 to 35), which were counted wrong, this may have resulted in a lower correlation between the PAT and the ESLAT than between the PAT and the teacher evaluations (for more information, see Appendix H).

Item Analysis of the PAT

After the PAT was evaluated for reliability and validity, the PAT was subjected to item analysis. Item analysis was carried out in two parts: item difficulty and item discriminability. All the subjects (*N* = 35) were included to determine the item difficulty. Oller (1979) says that "items falling somewhere between about .15 and .85 are usually preferred" (p. 247), which means that items with a proportion of correct answer (p) between .15 and .85 should be accepted. The items which lie outside of these two points should be revised or omitted. In the PAT, there are 11 items which should be revised or omitted because they are either too easy (e.g., Item 38, proportion correct (p) = .97) or too difficult (e.g., Item 11, p = 0.00). (see Appendix J for a detailed picture of item difficulty statistics).

To determine item discriminability, the item discriminability with sample separation method (Henning, 1987) was used. According to Henning, using the upper 28% and the lower 28% of the scores of the total sample is suitable to evaluate how well an item discriminates a good student from a weak student. Henning suggests that a discriminability index of .67 be regarded as the lowest acceptable discriminability. Fifty-three of the items in the PAT had a higher discriminability index than .67 (the lowest acceptable discriminability index) (see Appendix K for item discriminability statistics).

Conclusions

In this chapter, the researcher has discussed the data analysis of the PAT which was developed using open-ended type of items such as interrogatives, sentence completion, and short-answers. After the administration of the PAT, the exam papers were rated by two independent raters. To estimate inter-rater reliability, the researcher calculated the correlation between the two raters using Pearson Product-Moment Correlation which resulted in a very high correlation coefficient $\underline{r}_{AB} = .99$. Internal consistency of the test was calculated in several ways. First, the researcher estimated internal consistency using two different splits: one used the traditional odd-even numbered scores of subjects, and the other used two halves in which care had been taken to select items independent from each other. For these split-half internal consistency measures, the researcher first calculated the correlation coefficient between the halves and then adjusted it with the Spearman-Brown Prophecy Formula. Reliability was also estimated by the Guttman split half reliability estimate. The reliability coefficients for these measurements were as follows: .96, .96 (for odd-even numbered scores) .88, and .87 (for independent halves). Two other reliability estimates were carried out by the researcher using K-R 20 and K-R 21 formulas. The reliability coefficient for K-R 20 was .91, whereas for K-R 21 the reliability coefficient was

.88. The validity of the PAT was evaluated for concurrent (criterion) validity using two external criteria, the ESLAT, and the teacher evaluations of subjects. There were some discrepancies in the scores of subjects which turned out to be errors in calculating total scores and a failure to evaluate a few items, apparently all careless errors. After correcting these careless errors the researcher calculated the correlation between the PAT and the ESLAT. To estimate the correlation between the PAT and the ESLAT, the researcher used Pearson Product-Moment Correlation. The PPMC between the PAT and the ESLAT is .61 (*df* = 28, *p* = .0004), and between the PAT and the teacher evaluations of subjects .74 (*df* =33, *p* =.0000) resulted in reasonably high correlation.

Item analysis of the PAT has demonstrated that the majority of items had acceptable item difficulty and item discriminability. Of 64 items, only 11 are lying out of the preferred item difficulty range, and only 11 are lying out of the lowest accepted discriminability index.

CHAPTER 5 CONCLUSIONS

Summary of the Study

In this study, the researcher investigated the possibility of developing model for a proficiency/final achievement test based on course objectives. The test format included open-ended items (short-answers, sentence completion, interrogatives, and rational cloze) as recommended by a number of researchers (e.g., Heaton, 1988; Hill and Parry, 1992; Hughes, 1989). Assuming that they might have harmful backwash effect (Hughes, 1989), no multiple-choice items were included in the test.

It was hoped that the new test would be a reliable and valid test. Because it was not feasible to check reliability with the test-retest method, the researcher, after determining inter-rater reliability, estimated internal consistency using the split-half reliability estimate with Pearson Product-Moment Correlation, adjusted for length with the Spearman-Brown prophecy formula, the Guttman split-half reliability estimate, the K-R 20 formula, and K-R 21 formula. For validity, two external criteria were used, the ESLAT and the teacher evaluations of subjects. Item analysis of the PAT (item difficulty and item discriminability) was carried out by the researcher.

Results and Discussion

Reliability

Inter-rater reliability. Because the test did not contain so-called objective items like multiple-choice items, it was important for the test to have inter-rater reliability to certify that the test was scored objectively. Two independent raters scored all the answer sheets of the subjects from an answer sheet prepared by the researcher. To estimate the inter-rater reliability, the researcher correlated the scores of two raters using Pearson Product Moment Correlation. The inter-rater reliability of the test scoring was r_{AB}= .99. It is clear from the coefficient that the inter-rater reliability is very high.

As mentioned in the previous chapters, the reliability of scoring the open-ended test method has been debated by researchers. However, because of the ease of administration and objectivity in scoring, today most of the tests include or almost exclusively consist of multiple-choice items. The PAT, which consisted of generally open-ended items has demonstrated that if the items are carefully written taking the recommendations of researchers (e.g., Hughes, 1989) about how to write good items into consideration, and if the test is pre-piloted on target population (if feasible more than twice as many the researcher did in this study), even the open-ended test items may be scored almost as objectively as, and be as _reliable_ as multiple-choice test items. However, in spite of obtaining a high degree of inter-rater reliability on the two raters' scores, discrepancies in their scores suggest that tests should be evaluated at least by two raters plus one moderator in case one of the raters may fail to score an item or make mistakes in adding up the scores of subjects. In scoring test with the open-ended items, a separate, well prepared answer sheet should be delivered to the raters, and the raters should be clearly informed about grammatical mistakes [whether they (grammatical mistakes) will be counted wrong or correct, or will be given partial credit]. In this study, the researcher "recommended" to the raters that grammatical mistakes be ignored but this proved to be not explicit enough.

Internal consistency--split-half reliability. In this study, because test-retest was not feasible, the researcher evaluated the PAT for internal consistency using different split-half reliability coefficient estimates. The reliability coefficient for internal consistency ranged from .87 to .96.

Anastasi (1988) points out that the correlation coefficient for desirable reliability coefficient usually fall in the .80s or .90s. All the correlations range between .87 and .96, all well with Anastasi's criterion of a desirable reliability coefficient. It is clear from the findings that the reliability coefficient estimated using different reliability estimate formulas have all resulted in reasonably high reliability coefficients.

Validity

The concurrent (criterion) validity of the PAT against the ESLAT ($\underline{r}$ = .61; $\underline{df}$ = 28; $\underline{p}$ <.0004) and the teacher evaluations of the subjects ($\underline{r}$ = .74; $\underline{df}$ = 33; $\underline{p}$ =.0000) were statistically significant.

The validity of the multiple-choice test method has been much debated also. It is clear from the results mentioned above that the reasonably high correlation of the PAT against the ESLAT and the teacher evaluations can be interpreted as meaning that a reliable test which includes the open-ended test items may be a valid test method. In the study, the significant but somewhat lower correlation between the PAT and the ESLAT might be due to the fact that some of the subjects gave some items more than one answer, which may mean that on the ESLAT the so-called objectivity in scoring of the multiple-choice test method cannot be assumed unless scorers guard against this problem.

The major difference between the multiple-choice test method and the open-ended test method is that in the former, testees are required to recognize correct answers rather than produce, and in the latter testees are required to both comprehend and produce. In the latter, because testees are required to integrate reading skills with writing skills, the test method can be regarded to be measuring the real performance with the language that the course objectives suggest they will need in their university studies. Because testees are not supplied any options in the open-ended method, testees have to produce what required, which should encourage testees to learn the necessary skills because, otherwise, they cannot be successful in the examinations.

Item Analysis

Item difficulty. Item difficulty of the PAT was done (see Chapter 4). There are 11 items which are outside of the accepted lowest and highest range for an ideal item--proportion correct (p)>.15 and <.85--(Oller 1979), (see Appendix J for item difficulty statistics for each item).

Item Discriminability. According to Henning (1987), item discriminability of a test is as important as item difficulty. Only 11 of items had lower discriminability than the accepted lowest discriminability index, .67 (Henning, 1987). Detailed statistics of item discriminability can be seen in Appendix K.

Item analysis, that is, item difficulty and item discriminability, of the open-ended test items can be done easily. These are two important characteristics of a language test along with reliability and validity. According to the criteria given by Oller (1979), of the 64 items in the PAT, only 11 were out of the accepted extremes, and according to the criteria given by Henning

(1987), of 64 items only 11 were less than the lowest acceptable discriminability index. However, because there was overlap between these two groups, that is, of the 11 items outside the accepted difficulty extremes, 3 were also in the group of items that fell outside the lowest acceptable discriminability index. These statistics suggest that only 19 items should be revised or omitted before administering the test again.

Implications for Further Study

In this study, because of time constraints, writing and listening sections were not included. The number of subjects was very limited and all of them were volunteers. It is recommended that another similar study be carried out which will include writing and listening sections with a larger sample of subjects selected randomly. This study can be done comparing the multiple-choice test method and the open-ended test method including reading comprehension, writing, and listening comprehension in terms of reliability and validity of each test method.

References

Alderson, J. C., & Buck, G. (1993). Standards in testing: A study of practice of UK examination boards in EFL/ESL testing. Language Testing, 10, 1-26.

Alderson, J. C., Krahnke, K. J. & Stansfield, C. W. (Eds.). (1987). Reviews of English language proficiency tests. Washington, DC: TESOL.

Anastasi, A. (1988). Psychological testing. New York: Macmillan Publishing Company.

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

Baker, D. (1989). Language testing: A critical survey and practical guide. London: Edward Arnold.

Çelebi, D. (1991). An experimental study of the multiple-choice test technique for measuring reading comprehension of EFL students. Unpublished master's thesis, Bilkent University, Ankara.

Davies, A. (1990). Principles of language testing. Cambridge, MA: Basil Blackwell.

English as a Second Language Achievement Test. (1984). Collage Entrance Examination Boad. NY: Hato Rey.

Finocchiaro, M. & Sako, S. (1983). Foreign language testing: A practical approach. New York, NY: Regents.

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. Language Testing, 3, 159-185.

Harris, D. P. (1969). Testing English as a second language. New York: McGraw-Hill.

Heaton, J. B. (1988). Writing English language tests: A practical guide for teachers of English as a second or foreign language. Hong Kong: Longman.

Henning, G. (1987). A guide to language testing: Development, evaluation and research. Cambridge, MA: Newbury House.

Henning, G. H., Ghawaby, S. M., Saadalla, W. Z., El-Rifai, M. A., Hamallah, R. K., & Maffer, M. S. (1981). Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. Tesol Quarterly, 15, 457-466.

Henning, G. (1982). Twenty common testing mistakes for EFL teachers to avoid. FORUM, 20, 33-37.

Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy in reading assessment. Tesol Quarterly, 26, 433-461.

Hughes, A. (1989). Testing for language teachers. Cambridge: Cambridge University Press.

Hughes, A. (Ed.). (1988). Testing English for university study. ELT Documents 127. London: Modern English Publications.

Kirschner, M., Wexler, C., & Spector-Cohen, E. (1992). Avoiding obstacles to student comprehension of test questions. Tesol Quarterly, 26, 537-556.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. Language Testing, 10, 211-234.

Madsen, H. S. (1983). Techniques in testing. Oxford: Oxford University Press.

Oller, J. W. (1979). Language tests at school: A pragmatic approach. London: Longman.

Pierce, B. N. (1992). Demystifying the TOEFL® reading test. Tesol Quarterly, 26, 665-691.

Raatz, U. (1985). Better theory for better tests? Language Testing, 2, 60-75.

Savignon, S. J. (1991). Communicative language teaching: State of the art. Tesol Quarterly, 25, 261-277.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. Language Testing, 1, 147-170.

Shohamy, E., & Reves, T. (1985). Authentic language tests: Where from and where to? Language Testing, 2, 48-59.

Theunissen, T. J. J. M. (1987). Text banking and test design. Language Testing, 4, 1-8.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. Language Testing, 10, 41-70.

Weir, C. (1990). Communicative language testing. New York: Printice Hall.

Appendix A

Written Consent Form

I am conducting a two-part study on testing. I hope my results will be useful in helping Erciyes University to develop more valid and reliable tests in future. Your scores on the two tests will ***NOT*** affect your grade or your admission to the university. I will post your scores by your school number for your interest only. Your scores will ***NOT*** be given to anyone. Your identity will remain confidental. Your voluntary cooperation is greatly appreciated.

Faruk BALKAYA

BILKENT UNIVERSITY

MA TEFL

I understand that I am participating in a two-part study on testing and that I am free to withdraw at any time and that my identity will remain confidental.

Name : ____________________

Student No : ____________________

Signature : ____________________

Date : / /1994

Appendix B

Model for a Proficiency/Final Achievement Test (PAT)

ERCIYES UNIVERSITY ENGLISH LANGUAGE PROFICIENCY TEST

This test has two-parts: Part I is the Use of English, and Part II is Reading Comprehension. You will have 150 minutes in total to finish the test. Do not spend too much time on any one question on Part I, and Part II. If you are not sure of an answer, go on the next question. You can come back to the question later. If you do not know an answer, make a guess. There is no penalty for guessing.

Write all your answers for Part I and II on the answer sheet. Be careful to write legibly. Use a dark pencil. If you change your mind about and answer, erase completely and rewrite your answer. When the time allotted for Parts I and II (150 minutes) is finished, your answer sheet will be collected.

PART I: USE OF ENGLISH SECTION

A) Read the following words and phrases in the box below and select one for each blank in the passage that best completes the meaning. Do not use any word or phrase more than one time.

better than, inconclusive, conducting, measure, lack, measurement,
studies, effective, observers, observation, successfully, carried out,
comparisons, hypothesis, comparative

Some Problems in the Study of Language-teaching Methods

Researchers and teachers (as well as some students) have long been interested in whether any one method of teaching a second language is more (1)____________ than another. Several comparative studies of language teaching methods have been (2)____________, notably in Britain, Sweden and the United States. Results have generally been (3)____________, yet it is hard to believe that teaching methods make no difference at all. Therefore, attention has been focused on the research methods used by the researchers in (4)____________ the studies themselves.

Several possible reasons for the (5)____________ of clear findings have emerged. First, very few (6)____________ have taken individual differences among students into account; they have looked instead for methods which could be used (7)____________ with students of all types. Thus, method A may indeed be (8)____________ method B for more intelligent adults or for those with certain kinds of learning styles, but the studies have rarely been designed in such a way that this (9)____________ could be tested. Second, the tests of language proficiency used to (10)____________ students' achievement have often been inadequate. They have sometimes simply been unreliable and, hence, invalid; on other occasions, they have tended to reflect the aims of one method rather than another, making true (11)____________ difficult. Lastly, control over what actually took place inside the classrooms studied has frequently been insufficient. Observations by trained (12)____________ have been infrequent or even nonexistent. This means that one cannot be sure that one really is looking at data from method A and B and not at the results of several teaching and learning activities which had little or nothing to do with either.

Note. From Reading English for Academic Study (p. 15) by M. H. Long, W. Allen, A. Cyr, C. Pomeroy, E. Ricard, N. Spada & P. Vogel, 1980, Massachusetts: Newbury House.

B) Read the following passage and fill in each blank with one suitable word. A contraction (e.g., isn't) or a hyphenated word (e.g., twenty-two) is one word.

Refugees

Refugees are people (13)____________ have had to leave their homes in order to survive. War, lack of food, and lack of freedom are three reasons that refugees seek new homes. Life is difficult for refugees. They seldom want to leave their homes, but it is necessary. Very often, they move into a culture that is very (14)____________ (15)____________ their own. Therefore, they must adapt to the new culture by getting used to a new language, new foods, customs, and other social practices. They learn new way of life. They are often successful at (16)____________ new lives for themselves, but they almost always dream (17)____________ returning home.

Refugees are most successful when they move to developed nations. In these countries the governments can help refugees (18)____________ providing free social services. For example, the governments can pay for a refugee to learn a skill (19)____________ (20)____________ he or she can find a good job. The governments can also provide food and a dwelling. When the governments can afford to help them, refugees usually become independent after six month or a year. They (21)____________ need government help anymore.

Many refugees move to developing countries. In these nations, survival is often difficult. The governments must pay attention to the needs of their own people first. Sometimes they cannot afford to help the refugees. They must ask assistance of international charity organizations like the International Red Cross or peace-keeping ones like the United Nations. These organizations try to provide the refugees (22)____________ basic necessities. They also work to establish good conditions so that the refugees can become independent.

PART II: READING COMPREHENSION

Read the following passages and give enough information referring to the text to answer the items following each text.

War!

The question of whether war is inevitable is one which has concerned many of the world's great writers. Before considering this question, it will be useful to introduce some related concepts. Conflict, defined as opposition among social entities directed against one another, is distinguished from competition, defined as opposition among social entities independently striving for something which is in inadequate supply. Competitors may not be aware of one another, while the parties to a conflict are. Conflict and competition are both categories of opposition, which has been defined as a process by which social entities function in the disservice of one another. Opposition is thus contrasted with cooperation, the process by which social entities function in the service of one another. These definitions are necessary because it is inevitable in a world of limited resources, but conflict is not. Conflict, nevertheless, is very likely to occur, and is probably an essential and desirable element of human societies.

Many authors have argued for the inevitability of war from the premise that in the struggle for existence among animal species, only the fittest survive. In general, however, this struggle in nature is competition, not conflict. Social animals, such as monkeys and cattle, fight to win or maintain leadership of the group. The struggle for existence occurs not in such fights, but in the competition for limited feeding areas and for the occupancy of areas free from meat-eating animals. Those who fail in this competition starve to death or become victims to other species. This struggle for existence does not resemble human war, but rather the competition o f individuals for jobs, markets, necessities of life that are insufficient to satisfy all.

Among nations there is a competition in developing resources, trades, skill, and a satisfactory way of life. The successful nations grow and prosper; the unsuccessful decline. While it is true that this competition may induce efforts to expand territory at the expense of others, and thus lead conflict, it cannot be said that war-like conflict among nations is inevitable, although competition is.

23) Many of the world's greatest writers have been concerned with whether ____________ can be avoided.
24) According to the passage, the meaning of "opposition" has been stated as ____________ .
25) The common point between "conflict" and "competition" is that ____________ .
26) How is opposition different from cooperation? ___________ .
27) Many authors have based their arguments related to the inevitability of war on the idea that ___________ .
28) The most important quality of the struggle for survival in nature is ____________ .
29) According the conclusion the author makes, ____________ cannot be avoided.

Note. From Reading English Discourse (p. 220) by L. F. Bachman, 1986, New Jersey: Printice-Hall.

Weather Forecast

Although the weathermen's forecasts for a month ahead are only a little better than guesswork, they are now making long-term forecasts into the next century with growing confidence. For the dominant trend in the world's climate in the coming decades will, scientists say, be a predictable result of man's activities.

At the start of the industrial revolution nearly two centuries ago, man innocently set off a gigantic experiment in planetary engineering. Unaware of what he was doing, he spared no thought for the consequences. Today the possible outcome is alarmingly clear, but the experiment is unstoppable. Within the lifetime of many of us, the Earth may become warmer than it has been for a thousand years. By the middle of the next century it may be warmer than it has been since before the last Ice Age. And the century after that may be hotter than any in the past 70 million years.

Superficially, a warmer climate may seem welcome. But it could bring many hazards--disruption of crops in the world's main food-producing regions, famine, economic instability, civil unrest and even war.

In the much longer term, melting of the great ice-caps of Greenland and Antarctica could raise sea-levels throughout the world. The average sea-level has already risen a foot since the turn of the century. and if the ice-caps disappear entirely, it would rise by nearly 200 feet. Complete melting might take many centuries, but even a small increase in the sea-level would threaten low-lying parts of the world such as the Netherlands.

The man-made agent of climatic change is the carbon dioxide that has been pouring out of the world's chimneys in eve-increasing quantities since the industrial revolution began. And in the past few years scientists have begun to suspect that there is a second man-made source of carbon dioxide which may be as important as the burning of fossil fuels, namely the steady destruction of the world's great forests. Trees and other vegetation represent a huge stock of carbon removed from circulation, like money in a bank. As the vast tropical forests are cut down, most of the carbon they contain finds its way back into the atmosphere as carbon dioxide.

This amount of CO_2 (Carbon dioxide) in the atmosphere is still tiny. But it has climatic effect out of all proportion to its concentration. It acts rather like the glass in a greenhouse, letting though short-wave radiation from the sun, but trapping the longer-wave radiation by which the Earth loses heat to outer space.

Computer studies have suggested that if the concentration of carbon dioxide in the atmosphere were to be twice that of today's, there would be a rise of between 2^0C and 3^0C in average temperature.

30) The world's climate appears to be getting warmer as a
result of the ____________ which began about 200 years ago.
31) Scientists are fairly sure that, by the year 2050 ____________ .
32) In the 22nd century it may ____________ .
33) In the 19th century the average sea level was
____________ than in the 20th century
34) What specific evidence is there that the earth is
warmer today than it was around the year 1900? ____________ .
35) Man has changed the world's climate by destroying
forests and by ____________ .
36) ____________ are important because they keep a large
amount of carbon removed from circulation.

Note. From Longman First Certificate Coursebook (p. 159) by R. Kingsbury, 1988, United Kingdom: Longman.

Another Firebomb in Large Store
Newscastle, 16 June

(1) Fire broke out in the early hours of yesterday morning in the large A& B store in Newscastle. Fortunately the only casualty was in watchman, who was taken to hospital but was released this morning. There was extensive damage to the third floor of the building.
(2) "From what we can gather at the moment," the Fire Officer said, "We don't think there was an electrical fault. In fact, we suspect the fire was started by an incendiary device which someone had set to go off at about 2 a.m., but are not absolutely certain yet".
(3) The only person in the store was Jim London, the 57-year-old night watchman. He was overcome by fumes and was taken to the General Hospital unconscious. When he came to, he told reporters, "I had already done my third inspection of the store-I go round four or five times during the night-and was setting down to write my report when I noticed an odd smell and thought I heard something. I broke off and went to look into it. It wasn't until I'd made absolutely sure there was a fire and I couldn't do anything about ***it*** myself that I rang the fire brigade. And by that time, smoke was billowing everywhere so ***I*** didn't know how big it was".
(4) The manager told our reporter this morning, "We have had a number of threats during the past few weeks, but the police have not been able to find out where ***they*** have come from. There was a minor fire in the store the same time last year and we had received a number of warnings before that one, too".
(5) He went on, "When the Fire Prevention people inspected the store after that fire, they were lightly critical of our fire precautions, but since then we have installed a complete new fire prevention system".
(6) "But for Mr. London," ***he*** added, "it could have been much worse. We shall be showing our appreciation to him with a gift".

37) What date did the fire break out? ____________ .
38) How many people were there in the store when the fire broke out? ____________ .
39) When the watchman was taken to the General Hospital overcome by fumes, he was ____________ .
40) Jim called the fire brigade after discovering the fire only when he was sure that ____________ .
41) What did the store people do between the two fires in order to decrease the fire danger? ____________ .
42) The word "it" in line 16 refers to ____________ .
43) The word "I" in line 27 refers to ____________ .
44) The word "they" in line 21 refers to ____________ .
45) The word "he" in line 27 refers to ____________ .

Note. From Longman First Certificate Coursebook (p. 65) by R. Kingsbury, 1988, United Kingdom: Longman.

Population Growth and Industry

(1) We have looked at some of the ways in which biological factors affect human population growth. However, although biological laws underlie all the phenomena of population, once societies reach an advanced level of technology and culture it is more meaningful to explain what is happening in terms of sociological, economic and political influences.

(2) The study of population statistics in themselves is called "demography". All advanced countries now collect detailed statistics on births, marriages and deaths, and very few years a census of the population is taken. In England these figures are published by the General Register Office in London. World figures for population changes are much more difficult to compile because many underdeveloped countries do not keep complete records. However, a very detailed list of the available statistics is published every year in the United Nations Demographic Yearbook.

(3) From a careful study of these figures, demographers have worked out a description of what they think happened in the history of the population of a modern industrial nation. Throughout most of human history, ***they*** believe, man has had a very high death-rate and a high birth-rate. The death-rate may have been due to infanticide, epidemic disease or starvation, but it was typical of traditional tribal and peasant societies. Since it was balanced by large numbers of births the size of the population remained stable. modern populations in Africa, and much of South America and Asia, are examples of what may have been universal in the past. In these countries, a very large proportion of the population belongs to an age-group capable of becoming parents. This means that, compared with modern industrial countries, the birth-rate will be very high, not only because women have bigger families, but because the proportion o women capable of having children is also much higher.

(4) This is a stage of high potential growth because, if the death-rate could be reduced, the population would increase very rapidly. In abut one-fifth of the world, modern medicine has reduced the death-rate and here the population explosion is greatest. South-eastern Europe, some South America countries and India are all more or less at this stage. It seems almost certain that many more countries will arrive at this situation by the end of the century. The available statistics suggest that the modern industrial nations o the West passed through a phase like this in the nineteenth century.

(5) After this transitional growth stage, a third change took place in the Western nations. The birth-rate began to drop, and by the 1930s several north European countries had reached a new stable level with low birth-rates combined with low death-rates. In some countries the population declined, and governments actively encouraged people to have more children.

(6) The three stages in this transition can be summarized in a graph. Each has a distinctive economic arrangement. In the earliest phase there is a very low level of productivity, energy sources are primitive, and the standard of living is very low. At the middle stage, agriculture becomes more productive but does not always keep up with population growth, and industrial growth begins. The third stage has a very high standard of living, great efficiency and universal, sophisticated technology.

(7) This "transition" theory of population growth is based on what happened in modern industrial nations. If the theory is applicable to the underdeveloped countries, we would expect that if they industrialize and modernize there will be a decline in fertility until the population is stabilized. If industrialization is not achieved in the next one hundred years there are two other possibilities for slowing the growth of the population. The death-rate could begin to rise again because medicine and hygiene cannot keep up with the continued rise in population. Alternatively, there could be a decline in fertility before industrialization. ***This*** has never happened before, but it is just possible that a peasant population might be influenced by a widespread birth-control campaign if they had enough help and encouragement from the government.

46) Why is it difficult to know very accurately how much the number of the people in the World changes? _______________ .
47) For almost all the history of man, the _______________ has been about the same as the _______________ and the population has remained relatively constant.
48) What three forces tended to keep the death rate high in traditional peasant and tribal societies? _______________ .
49) What generally happens to the birth rate in the third stage of the population transition theory? _______________ .
50) What generally happens to the population during the agricultural stage? _______________ .
51) The word "they" in line 19 refers to _______________ .
52) The word "this" in line 64 refers to _______________ .
53) What one word in paragraph (5) means nearly the same as ***unchanging***? _______________ .
54) What one word in paragraph (6) means nearly the same as ***stage***? _______________ .

Note. From Beyond Words (p. 58) by A. Malay and A. Duff, 1987, Cambridge: Cambridge University Press.

The Secrets of Sleep

(1) The secrets of sleep were a mystery for centuries simply because there was neither the means to explore them, nor the need. Only when candles gave way to gaslight, and gas to electricity, when man became able to convert night into day, and double his output by working shifts round the clock, did people seriously start wondering if sleep could possibly be a waste of time. Our ability to switch night into day is very recent, and it is questionable if we will ever either want, or be able, to give up our habit of enjoying a good night's sleep. However, a remarkable research project in London has already discovered a few people who actually enjoy insomnia. Even chronic insomniacs often get hours more sleep than they think. But, by placing electric contacts beside the eyes and on the head, it is possible to check their complaint by studying the tiny currents we generate which reveal the different brainwaves of sleep and wakefulness. This has shown that for some people seven or eight hours of sleep a night are quite unnecessary.

(2) A lot of recent work has shown that too much sleep is bad for you, so that if you are fortunate enough to be born with a body which needs only a small amount of sleep, you may well be healthier and happier than someone who sleeps longer.

(3) Every attempt to unravel the secrets of sleep, and be precise about its function, raises many problems. The sleeper himself cannot tell what is going on and, even when he wakes, has only a very hazy idea of how good or bad a night he has had. The research is expensive and often unpopular, as it inevitably involves working at night. Only in the last few years have experts come up with theories about the function of sleep and the laws which may govern **it**.

(4) The real advance in sleep research came in 1937 with the use of the electroencephalogram. ***This machine*** showed small--50 micro volt --changes in the brain, so, for the first time, we could observe sleep from moment to moment. Before that time one could put the person to bed, watch him mumble, toss, turn, bring back a few rough memories of dreams, and that was about all. In 1937 it was possible to read out these changes, second by second. Then in 1959 two other things happened. Kleitman and Aserinski, as they were looking at eye movements, trying to understand the brainwaves, noticed that after about ninety minutes there would be a burst of the EEG, as if the person was awake, and the eyes would move rapidly. It was not hard to guess that maybe that was a dream. And indeed it was. Waking people up during that period, they found they were dreaming; waking them up other periods, they found no dreams.

(5) The electroencephalograph shows that when we fall asleep we pass through a cycle of sleep sages. At the onset of sleep, the cycle lasts about ninety minutes during which you pass through stages one, two and three to stage four. This is the deepest form of sleep, and from it you retreat to stage two, and from there into REM, or rapid eye movement sleep. Here, for ten minutes on the first cycle and then gradually longer, it is thought that we do most of our dreaming.

(6) Studies of people who volunteered to be locked up for weeks in an observation chamber with no idea of whether it is night or day, give remarkable results. We are not, in fact, twenty-four-hour creatures. Put people in such circumstances and, even though the patterns of sleep continue, the day is extended to about twenty-five and a half hours. Without any clues to time, these people go to sleep the first night abut an hour later than usual, the next night an hour later, *and* the next night. So that, after about ten days, the person is going to sleep at three o'clock in the afternoon, thinking that he is still going to sleep at midnight.

(7) Today, jet-lag is a familiar hazard for the seasoned traveler. Travel across time zones play havoc with the biological clock rhythms of the human body. For the active pilot, who is rarely in one place long enough to know if it is time for breakfast or dinner, impact of jet-lag on his sleep is critical. Several air disasters have been partly caused by overtired pilots ignoring the natural laws of sleep. Much research is directed to finding out what there laws are and to what extent pilots and astronauts dare disobey them. But they are laws which affect all of us, not just pilots.

(From an article in the *Listener*)

55) Paragraph one suggests that if people didn't need to sleep, they could ______________ more.
56) Because it is necessary to do sleep research at night, this kind of research ______________ .
57) The electroencephalograph made it possible to know exactly when a person was having a ______________ .
58) When a person dreams, it is possible to see his eyes ______________ .
59) Paragraph (5) suggests that our early morning dreams take ______________ time than the first dream we have at night
60) If you normally go to bed at 10:00 p m , what time would you probably go to bed if you had no way of knowing what time it was? ______________ .
61) What one word in paragraph (1) means the same as ***change***? ______________ .
62) What one word in paragraph (1) means the same as ***people who have difficulty in sleeping***? ______________ .
63) The word "it" in line 27 refers to ______________ .
64) The words "this machine" in line 29 refer to ______________ .

Note. From Beyond Words (p. 69) by A. Malay and A. Duff, 1987, Cambridge: Cambridge University Press.

Appendix C

PAT Answer Sheet

E.U. FOREIGN LANGUAGES DEPARTMENT

GROUP : ____________ **DATE:** / /1994

NUMBER: ____________

PROFICIENCY/FINAL ACHIEVEMENT TEST

ANSWER SHEET

Directions: Write all your answers on the answer sheet. Do *NOT* write on the tests.

PART I: USE OF ENGLISH

01) ______________

02) ______________

03) ______________

04 ______________

05) ______________

06) ______________

07) ______________

08) ______________

09) ______________

10) ______________

11) ______________

12) ______________

13) ______________

14) ______________

15) ______________

16) ______________

17) ______________

18) ______________

19) ______________

20) ______________

21) ______________

22) ______________

PART II: READING COMPREHENSION

23) ____

24) ____

25) ____

26) ____

27) ____

28) ____

29) ____

30) ____

continues

PAT Answer Sheet

31) ____________________

32) ____________________

33) ____________________

34) ____________________

35) ____________________

36) ____________________

37) ____________________

38) ____________________

39) ____________________

40) ____________________

41) ____________________

42) ____________________

43) ____________________

44) ____________________

45) ____________________

46) ____________________

47) ____________________

48) ____________________

49) ____________________

50) ____________________

51) ____________________

52) ____________________

53) ____________________

54) ____________________

55) ____________________

56) ____________________

57) ____________________

58) ____________________

59) ____________________

60) ____________________

61) ____________________

62) ____________________

63) ____________________

64) ____________________

Appendix D

Subjects' Scores on the PAT, the ESLAT, and the Teacher Evaluations of Subjects (TEVAL)

Subjects'	PAT	ESLAT	TEVAL
no	*N* = 35	*N* = 30	*N* = 35
01	50	44	5.0
02	48	31	4.5
03	45	43	4.0
04	43	35	4.5
05	44	48	4.0
06	37	39	4.5
07	38	46	4.0
08	37	42	4.5
09	37	51	3.0
10	36	43	3.5
11	33	39	4.5
12	32	40	4.0
13	33	40	3.5
14	32	38	4.0
15	31	26	3.5
16	31	37	3.5
17	29	29	4.0
18	30	39	3.5
19	27	37	4.0
20	27	35	3.5
21	27	39	4.0
22	27	39	3.0
23	26	39	3.5
24	24	37	3.0
25	19	31	3.5
26	23	36	3.0
27	22	35	3.0
28	12	25	2.0
29	40	33	4.5
30	10	15	3.0
31	33		4.5
32	17		3.0
33	10		3.5
34	26		4.0
35	9		3.0

table continues

Subjects' Scores on the PAT, the ESLAT, and the Teacher Evaluations of Subjects (TEVAL)

Subjects' no	PAT *N* = 35	ESLAT *N* = 30	TEVAL *N* = 35
Mean	29.86	37.03	3.71
Variance	110.89	51.76	.42
Standard Deviation	10.53	7.19	.65

Note. PAT = Erciyes University Proficiency/Final Achievement Test; ESLAT = English as a Second Language Achievement Test; TEVAL = Teachers' evaluation of subjects.

Appendix E

Subjects' Scores Given by the First Rater (R1), the Second Rater (R2), and the Researcher (RSC)

Subject			
No	R1	R2	RSC
01	49	50	50
02	44	48	48
03	44	45	45
04	40	43	43
05	44	41	44
06	37	38	37
07	37	38	38
08	36	38	37
09	36	37	37
10	36	36	36
11	30	32	33
12	31	32	32
13	33	33	33
14	30	32	32
15	29	31	31
16	28	31	31
17	28	29	29
18	28	30	30
19	27	28	27
20	26	27	27
21	26	27	27
22	27	28	27
23	26	26	26
24	24	24	24
25	19	19	19
26	23	23	23
27	21	21	22
28	12	12	12
29	38	40	40
30	10	10	10
31	33	33	33
32	17	17	17
33	10	10	10
34	26	26	26
35	09	09	09

Note. Pearson Product Moment Correlation between R1 and R2 is .99.

Appendix F

Subjects' Odd (OD) and Even Numbered (EV) Scores on the PAT

Subject		
No	OD	EV
01	26	24
02	22	26
03	23	22
04	22	21
05	22	22
06	18	19
07	20	18
08	19	18
09	18	19
10	19	17
11	16	17
12	16	16
13	18	15
14	16	16
15	15	16
16	14	17
17	13	16
18	14	16
19	14	13
20	13	14
21	11	16
22	14	13
23	13	13
24	11	13
25	09	10
26	12	11
27	11	11
28	03	09
29	21	19
30	04	06
31	16	17
32	09	08
33	06	04
34	15	11
35	05	04

	Sum	Mean	Variance	SD
OD	518	14.80	30.75	5.55
EV	527	15.06	27.11	5.21

p = .0000 *df* = 33 r_{tt} = .96

Note. PAT = Erciyes University Proficiency/Final Achievement Test, r_{tt} = Reliability estimated by the split half method.

Appendix G

Subjects' Scores on Two Different Independent Halves (H1) and (H2)

Subject No	H1	H2
01	21	29
02	21	27
03	21	24
04	23	20
05	21	23
06	19	18
07	20	18
08	16	21
09	14	23
10	15	21
11	15	18
12	19	13
13	14	19
14	19	13
15	13	18
16	15	16
17	13	16
18	15	15
19	13	14
20	12	15
21	14	13
22	11	16
23	10	16
24	11	13
25	10	9
26	13	10
27	11	11
28	5	7
29	21	19
30	6	4
31	19	14
32	9	8
33	4	6
34	11	15
35	5	4

table continues

Subjects' Scores on Two Different Independent Halves (H1) and (H2)

	Sum	Mean	Variance	S. Dev.
H1	499	14.26	26.31	5.13
H2	546	15.60	36.37	6.03

p = .0000; *df* = 33; $\underline{r}_{tt}$ = .88

Note. H1 = Subjects' scores for the first half (Items 1-12; 23-29; 37-45; 55-58); H2 = Subjects' scores for the second half (Items 13-22; 30-36; 46-54; 59-64); r_{tt} = Reliability estimated by the split half method.

Appendix H

Pearson Product Moment Correlation between the PAT (X) and the ESLAT (Y)

Subject No	X	Y	X2	Y2	XY
01	50	44	2500	1936	2200
02	48	31	2304	961	1488
03	45	43	2025	1849	1935
04	43	35	1849	1225	1505
05	44	48	1936	2304	2112
06	37	39	1369	1521	1443
07	38	46	1444	2116	1748
08	37	42	1369	1764	1554
09	37	51	1369	2601	1887
10	36	43	1296	1849	1548
11	33	39	1089	1521	1287
12	32	40	1024	1600	1280
13	33	40	1089	1600	1320
14	32	38	1024	1444	1216
15	31	26	961	676	806
16	31	37	961	1369	1147
17	29	29	841	841	841
18	30	39	900	1521	1170
19	27	37	729	1369	999
20	27	35	729	1225	945
21	27	39	729	1521	1053
22	27	39	729	1521	1053
23	26	39	676	1521	1014
24	24	37	576	1369	888
25	19	31	361	961	589
26	23	36	529	1296	828
27	22	35	484	1225	770
28	12	25	144	625	300
29	40	33	1600	1089	1320
30	10	15	100	225	150

	Sum	Mean	Variance	S. Dev.
X	950	31.67	91.47	9.56
Y	1111	37.03	51.76	7.19

$p<.0004$; *df* = 28; r_{xv} = .61

Note. PAT = Erciyes University Proficiency/Final Achievement Test; ESLAT = English as a Second Language Achievement Test; r_{xy} = Pearson Product Moment Correlation between the PAT and the ESLAT.

Appendix I

Pearson Product Moment Correlation between the PAT (X) and the Teacher Evaluations of Subjects (Y)

Subject No	X	Y	X2	Y2	XY
01	50	5.0	2500	25.00	250.0
02	48	4.5	2304	20.25	216.0
03	45	4.0	2025	16.00	180.0
04	43	4.5	1849	20.25	193.5
05	44	4.0	1936	16.00	176.0
06	37	4.5	1369	20.25	166.5
07	38	4.0	1444	16.00	152.0
08	37	4.5	1369	20.25	166.5
09	37	3.0	1369	9.000	111.0
10	36	3.5	1296	12.25	126.0
11	33	4.5	1089	20.25	148.5
12	32	4.0	1024	16.00	128.0
13	33	3.5	1089	12.25	115.5
14	32	4.0	1024	16.00	128.0
15	31	3.5	961	12.25	108.5
16	31	3.5	961	12.25	108.5
17	29	4.0	841	16.00	116.0
18	30	3.5	900	12.25	105.0
19	27	4.0	729	16.00	108.0
20	27	3.5	729	12.25	94.5
21	27	4.0	729	16.00	108.0
22	27	3.0	729	9.000	81.0
23	26	3.5	676	12.25	91.0
24	24	3.0	576	9.000	72.0
25	19	3.5	361	12.25	66.5
26	23	3.0	529	9.000	69.0
27	22	3.0	484	9.000	66.0
28	12	2.0	144	4.000	24.0
29	40	4.5	1600	20.25	180.0
30	10	3.0	100	9.000	30.0
31	33	4.5	1089	20.25	148.5
32	17	3.0	289	6.250	42.5
33	10	3.5	100	12.25	35.0
34	26	4.0	676	16.00	104.0
35	9	3.0	81	9.000	27.0

	Sum	Mean	Variance	S. Dev.
X	1045	29.86	110.89	10.53
Y	130	3.71	.42	.65

$\underline{p} = .0000$; $\underline{df} = 33$; $\underline{r}_{xy} = .74$

Note. $\underline{r}_{xy}$ = Pearson Product Moment Correlation between the PAT and the Teacher evaluations of subjects.

Appendix J

Item Difficulty of the PAT as Proportion Correct (p) and Proportion Incorrect (q), and Item Variance (pq)

Item					
No	N = 35	Correct	p	q	pq
01		19	.54	.46	.25
02		15	.43	.57	.25
03		06	.17	.83	.14
04		07	.20	.80	.16
05		11	.31	.69	.22
06		06	.17	.83	.14
07		10	.29	.71	.20
08		28	.80	.20	.16
09		02	.06	.94	.05
10		16	.46	.54	.25
11		00	0.00 *	1.00	.00
12		08	.23	.77	.18
13		30	.86 *	.14	.12
14		20	.57	.43	.25
15		18	.51	.49	.25
16		21	.60	.40	.25
17		28	.80	.20	.16
18		14	.40	.60	.23
19		07	.20	.80	.14
20		07	.20	.80	.16
21		29	.83	.17	.14
22		01	.03 *	.97	.03
23		24	.69	.31	.22
24		05	.14 *	.86	.12
25		16	.46	.54	.23
26		11	.31	.69	.22
27		17	.49	.51	.25
28		20	.57	.43	.25
29		07	.20	.80	.16
30		24	.69	.31	.22
31		26	.74	.26	.19
32		20	.57	.43	.25
33		18	.51	.49	.25
34		03	.09 *	.91	.08
35		10	.29	.71	.20
				Σpq=	11.93

* = p<.15 or >.85

table continues

Item Difficulty of the PAT as Proportion Correct (p) and Proportion Incorrect (q), and Item Variance (pq)

Item No	*N* = 35	Correct	p		q	pq
36		27	.77		.23	.18
37		26	.74		.26	.19
38		34	.97	*	.03	.03
39		20	.57		.43	.25
40		16	.46		.54	.25
41		10	.29		.71	.20
42		33	.94	*	.06	.05
43		31	.89	*	.11	.10
44		28	.80		.20	.16
45		31	.89	*	.11	.10
46		24	.69		.31	.22
47		17	.49		.51	.25
48		18	.51		.49	.25
49		10	.29		.71	.20
50		11	.31		.69	.22
51		19	.54		.46	.25
52		14	.40		.60	.24
53		17	.49		.51	.25
54		22	.63		.37	.23
55		05	.14	*	.86	.12
56		03	.09	*	.91	.08
57		16	.46		.54	.25
58		18	.51		.49	.25
59		08	.23		.77	.18
60		08	.23		.77	.18
61		15	.43		.57	.25
62		17	.49		.51	.25
63		16	.46		.54	.25
64		28	.80		.20	.16
					Σpq =	11.93

* = p<.15 or >.85

Appendix K

Item Discriminability (IDISC) of the PAT with Sample Separation

	Groups			
	Upper	Lower		
Item No	n=10 (28%)	n=10 (28%)	IDISC. Index	
01	9	3	.75	
02	8	2	.80	
03	3	2	.60	*
04	3	1	.75	
05	6	1	.86	
06	3	3	.50	*
07	7	0	1.00	
08	9	9	.50	*
09	1	0	1.00	
10	6	3	.67	
11	0	0	.00	*
12	3	2	.60	*
13	10	5	.67	
14	7	3	.70	
15	6	4	.60	*
16	9	1	.90	
17	9	4	.69	
18	7	3	.70	
19	2	1	.67	
20	2	0	1.00	
21	9	6	.60	*
22	0	0	.00	*
23	9	2	.82	
24	3	0	1.00	
25	4	1	.80	
26	5	0	1.00	
27	8	1	.89	
28	9	2	.82	
29	4	0	1.00	
30	7	6	.54	*
31	9	2	.82	
32	8	0	1.00	
33	6	1	.86	
34	2	0	1.00	
35	4	1	.80	
36	9	2	.82	
37	10	7	.59	*
38	10	7	.59	

Note. * = Discriminability level < .67.

table continues

Item Discriminability (IDISC) of the PAT with Sample Separation

	Groups			
	Upper	Lower		
Item No	_n_=10 (28%)	_n_=10 (28%)	IDISC. Index	
39	7	3	.70	
40	5	2	.71	
41	5	0	1.00	
42	10	6	.63	*
43	10	5	.67	
44	9	3	.75	
45	10	5	.67	
46	10	3	.77	
47	7	2	.78	
48	7	2	.78	
49	5	0	1.00	
50	7	0	1.00	
51	9	1	.90	
52	6	0	1.00	
53	10	1	.91	
54	10	2	.83	
55	3	0	1.00	
56	2	0	1.00	
57	8	1	.89	
58	7	2	.78	
59	5	0	1.00	
60	6	0	1.00	
61	8	2	.80	
62	8	3	.73	
63	8	0	1.00	
64	10	4	.71	

Note. * = Discriminability level < .67.