A STUDY ON DEVELOPING A WRITING ASSESSMENT PROFILE FOR ENGLISH PREPARATORY PROGRAM OF ANADOLU UNIVERSITY SCHOOL OF FOREIGN LANGUAGES

ANADOLU ÜNİVERSİTESİ YABANCI DİLLER YÜKSEKOKULU İNGİLİZCE HAZIRLIK PROGRAMINDA YAZILI ANLATIM BECERİLERİNİ ÖLÇMEDE KULLANILABİLECEK BİR ÖLÇEĞİN GELİŞTİRİLMESİ ÜZERİNE BİR ÇALIŞMA

By

Murat POLAT

Thesis Submitted for the Degree of Master of Arts

English Language Teaching Department

Advisor: Asst. Prof. Aynur YÜREKLİ

Eskişehir

Anadolu University

The Institute of Educational Sciences

July, 2003

ANADOLU ÜNİVERSİTESİ YABANCI DİLLER YÜKSEKOKULU
İNGİLİZCE HAZIRLIK PROGRAMINDA YAZILI ANLATIM BECERİLERİNİ
ÖLÇMEDE KULLANILABİLECEK BİR ÖLÇEĞİN
GELİŞTİRİLMESİ ÜZERİNE BİR ÇALIŞMA

A STUDY ON DEVELOPING A WRITING ASSESSMENT PROFILE FOR ENGLISH PREPARATORY PROGRAM OF ANADOLU UNIVERSITY SCHOOL OF FOREIGN LANGUAGES

Murat POLAT

(Yüksek Lisans Tezi)

Eskişehir, 2003

YÜKSEK LİSANS TEZ ÖZÜ

ANADOLU ÜNİVERSİTESİ YABANCI DİLLER YÜKSEKOKULU İNGİLİZCE HAZIRLIK PROGRAMINDA YAZILI ANLATIM BECERİLERİNİ ÖLÇMEDE KULLANILABİLECEK BİR ÖLÇEĞİN GELİŞTİRİLMESİ ÜZERİNE BİR ÇALIŞMA

Murat POLAT
İngiliz Dili Eğitimi Ana Bilim Dalı
Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü, Temmuz, 2003
Danışman: Yrd. Doç. Dr. Aynur Yürekli

Yabancı dil eğitiminde yazılı anlatım becerilerinin güvenilir ve tutarlı ölçümü eğitim kurumları için çoğu zaman problemli olmuştur. Bu tip ölçümlerde öznel değerlendirmenin, yani insan faktörünün değerlendirme sürecine daha fazla girmesiyle notlayıcı tarafından verilen notların güvenilirliği ve hangi kriterlere uygun verildiği daha çok sorgulanır hale gelmiştir. Yazılı anlatım notlayıcılarının kendi aralarında, hatta kendi notlamaları arasında doğabilecek olası tutarsızlıklar, öğrencilerin eğitim kurumuna olan güvenini azaltırken, her eğitim sürecinde yapılması gereken öğrenci (ve dolayısıyla program) başarısını belirleyici değerlendirmeleri de tehlikeye atar. Bu durumu dikkate alarak, yabancı dil eğitimi veren kurumlarda yazılı anlatım becerilerini ölçmede yaşanan bu problemleri en aza indirgemek amacıyla uzun bir süredir yazılı anlatım değerlendirmelerinde güvenirliği arttırıcı çalışmalar yapılmakta ve yeni ölçüm araçları geliştirilmektedir. Ancak, hazırlanan bu araçların geliştirme aşamasında kullanılacakları eğitim kurumlarının isteklerine cevap verecek tarzda desenlenmeleri ve güvenirlik derecelerinin buna göre belirlenmesi, yabancı dil eğitimi veren başka kurumların bu araçları kullanmalarına engel olmaktadır, çünkü çoğu zaman başka eğitim programlarında aranılan kriterler ölçektekileri tutmamakta, ölçekte revizyona gidilmekte, bu durum da orijinal ölçeğin güvenirliğini önemli ölçüde azaltmaktadır. Dolayısıyla bu çalışmada, her eğitim kurumunun halihazırda kullandıkları ölçeklerin güvenirlik düzeylerini belli aralıklarla ölçtürmeleri ve yetersiz sonuçlar aldıklar takdirde kendi program ve öğrenci profiline uygun güvenirliği kanıtlanmış ölçekler geliştirmesi ve bunları kullanmasının daha iyi olabileceği fikriyle, Anadolu Üniversitesi Yabancı Diller Yüksekokulu İngilizce

Hazırlık Programı yazılı anlatım sınavlarını değerlendirmede kullanılan ölçeğin güvenirliğinin ölçülmesi amaçlanmıştır.

Bu maksatla, bu okulda daha önce yapılmış bir yazılı anlatım sınavından 3 ayrı başarı seviyesinde (iyi-orta-kötü) toplam 50 kağıt seçilmiş ve bunlar o okulda çalışan ve yazılı anlatım değerlendirmesinde en az 3 yıl deneyimli 10 hocaya yine bu okulda kullanılan bütünsel performansı ölçücü tipte bir kriterle birer ay arayla iki kere okutulmuştur. Elde edilen sonuçların okuyucuların kendi aralarında ve kendi okumaları arasındaki tutarlılıkları göz önüne alınarak güvenirlik hesaplamaları yapılmış ve sonuçların düşük çıkması nedeniyle yeni bir ölçek geliştirilmiştir. Analitik tipte hazırlanan bu yeni ölçekle aynı kağıtlar aynı hocalara yine birer ay arayla okutulmuş ve elde edilen sonuçların okuyucuların kendi aralarında ve kendi okumaları arasındaki tutarlılıkları göz önüne alınarak güvenirlik hesaplamaları yapılmıştır. Bu hesaplamalardan 6 ay sonra aynı kağıtlar her iki ölçekle aynı hocalara bir ay arayla birer kez daha okutulmuş ve güvenirlik düzeyleri karşılaştırılmıştır.

Uygulama sonuçlarının istatiksel analizleri, notlayıcıların her iki kriterle yaptıkları notlamalarında farklı davrandıklarını ve yeni geliştirilen analitik ölçekle notlama yaptıklarında belirlenen güvenirlik derecesinde hem kendi aralarında hem de kendi notlamaları arasında diğer ölçeğe göre çok daha tutarlı olduklarını ortaya koymuştur. Üçüncü defa yapılan notlamaların sonuçları da, bundan önce yapılan iki notlamada elde edilen sonuçların uzun dönemde de aynı kaldığını gösterir niteliktedir. Bu sonuçlara göre yeni geliştirilen analitik ölçeğin Anadolu Üniversitesi Yabancı Diller Yüksekokulu İngilizce Hazırlık Programı Yazma Becerileri sınavlarını değerlendirmede bütünsel performansı ölçücü kritere oranla hem notlayıcılar hem de notlayıcıların kendi okumaları göz önüne alındığında daha güvenilir sonuçlar verebileceği söylenebilir.

MASTER OF ARTS THESIS

ABSTRACT

A STUDY ON DEVELOPING A WRITING ASSESSMENT PROFILE FOR ENGLISH PREPARATORY PROGRAM OF ANADOLU UNIVERSITY SCHOOL OF FOREIGN LANGUAGES

Thesis Submitted for the Master of Arts

English Language Teaching Department

Advisor: Asst. Prof. Aynur YÜREKLİ

In foreign language education, reliable and consistent measurement of writing abilities has quite often been challenging. The potential for subjective evaluation or the interference of the "human factor" in such measurements has led to questioning of the reliability of the graders' decisions and justifications. Possible inconsistencies among graders and gradings of their own may lessen the respect of the students for the institution as well as endangering a number of assessments, which should never be ignored in education programs, since it affects the learners', and the school's success. Taking such inconsistencies into consideration, a significant amount of studies have been done with an aim to enhancing reliability. Thus, new measuring instruments have been developed in an attempt to minimize the current problems of language schools in the assessment of writing. However, most of those criteria are not found particularly suitable for the needs of each language school, and a number of revisions are made on those grading profiles according to their needs. Eventually, these revisions on the criterion decrease the reliability of the original version to a great extent. For this reason, it is claimed that language schools would better have their current scoring standards tested regularly. In case low reliability levels are taken, to design and use their own grading criteria according to their goals and learner profiles which have proven to be reliable would be better for such schools. In this study, it has been aimed to find out the reliability levels of the holistic-analytic instrument that is

being used at Anadolu University School of Foreign Languages English Preparatory Program.

With this aim in mind, a total of 50 papers of different achievement levels (unsuccessful, moderate, successful) were selected from a previous final exam held in this school and given to ten graders who have a minimum of 3-year-experience in grading writing papers in this school. These graders were asked to grade these papers using the holistic-analytic criterion twice with 1-month interval. Both the inter/intra-rater reliability levels were computed from the data and a need to develop a new criterion emerged after it was found that the reliability levels of the holistic-analytic criterion were considerably low. With the new criterion designed to measure analytically, the same papers were again graded by the same graders twice with a month interval. To confirm those calculations, the same pile of papers was graded with each criterion by the same graders for the 3rd time after six months, and the results of the reliability tests were compared.

Statistical analysis of the above application revealed that the graders acted differently in their gradings using both instruments; what is more, the inter/intra-rater consistency degrees were found to be significantly higher than the ones gathered from holistic-analytic instrument when the new analytic criterion was used as the medium of evaluation at a certain significance level. Findings of the 3rd grading with each criterion confirmed the results of the previous gradings in the long-term. All these results seem to suggest that in the evaluation of writing exams in Anadolu University, School of Foreign Languages English Preparatory Program, the new analytic criterion would provide better inter/intra-rater reliability degrees than the holistic-analytic criterion.

JÜRİ VE ENSTİTÜ ONAYI

Murat POLAT'ın "İngilizce Hazırlık Programında Yazılı Anlatım Becerilerini Ölçmede Kullanılabilecek Bir Ölçeğin Geliştirilmesi Üzerine Bir Çalışma" başlıklı tezi 25.06.2003 tarihinde, aşağıda belirtilen jüri üyeleri tarafından Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca Yabancı Diller Eğitimi Anabilim Dalı İngilizce Öğretmenliği yüksek lisans tezi olarak değerlendirilerek kabul edilmiştir.

		Adı-Soyadı	İmza
Üye (Tez Danışmanı)	:	Yard.Doç.Dr.Aynur YÜREKLİ	
Üye	:	Doç.Dr.Hülya ÖZCAN	
Üye	:	Doç.Dr.Handan YAVUZ	
Üye	:	Y.Doç.Dr.Aysel BAHÇE	
Üye	:	Y.Doç.Dr.İlknur MAVİŞ	

Prof.Dr. İlknur KEÇİK
Anadolu Üniversitesi
Eğitim Bilimleri Enstitüsü Müdürü

ACKNOWLEDGEMENTS

I would like to thank many people for their encouraging assistance in writing this thesis.

First, I wish to express my deepest gratitude to my thesis advisor, Asst. Prof. Aynur YÜREKLİ for her never-ending guidance and support throughout this study. I am also very much indebted to Nesrin ORUÇ for the constant inspiration and help she has provided.

Next, I would like to express my gratefulness to Prof. Dr. Gül DURMUŞOĞLU-KÖSE for her helpful suggestions and direction.

I owe special thanks to my friends, Neslihan DOĞAN, Sema GÜN, Eylem KORAL Bülent ALAN, Nazan ARMAĞAN, Şener EŞ, Nazmi TASLACI, Hüseyin KAFES, İlkay GÖKÇE, Meral Melek ÜNVER and Emel ŞENTUNA who voluntarily accepted to be participants in this study.

I am also indebted to Jonathan ROSS for his proofreading, and Erdal KARA for his help and patience during the very complex statistical calculations.

Finally, my special thanks go to my wife, Emine POLAT, for her everlasting love and encouragement, and to my son, Eren, for his existence in my life.

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

1.1. Background to the Problem

The first time a language teacher grades a student's writing paper is generally a confusing experience since the task of grading students' written performance presents most EFL teachers with an uncomfortable dilemma. Although they are mostly provided with some kind of grading criteria, they have difficulties deciding which aspects of the paper to grade or how to assign a grade to those aspects of the composition that even they themselves are not sure about.

Consciously or not, some teachers put emphasis on content. They may grade a paper according to what was said, how well it was said, or how much information was transferred. Some may consider the effectiveness of the writing and grade a paper according to such rhetorical concerns as unity, coherence or organization. Others may stress the mechanics such as spelling, vocabulary, punctuation etc., and apart from those, a great number of teachers believe that a good piece of writing should be connected, contextualized and should have appropriate pieces of communication (Johns, 1991 in Ruetten, 1994, p: 85). It should thus be realized that some aspects of students' writing must, due to the variations among teachers' grading philosophies and the nature of the task, be evaluated from a subjective, global perspective. Having no valid proofs, researchers (Pollitt-Murray, 1996; Ruetten, 1994) claim that these perspectives should be ensured by a carefully designed training programme and with the help of teachers' experiences from language teaching classes, as Scott (1995) states, language teachers are experts in their field, and experts can make valid subjective judgments. Needless to say, language teachers should accept the necessity of assigning grades based on subjective evaluations when communicative writing skills are involved.

Leaving the issue of subjective evaluations beside, the search for objectivity has led to various measures for evaluating students' writing. EFL students' writing is frequently tested by means of essays, which are pieces of writing, planned and written by the students on pre-determined purposes and topics. The priority of essay writing in the context of testing writing in EFL stems from the fact that essays reflect the students' mastery of important organizational, structural and mechanical skills in writing better than the other devices which generally affect their writing preferences in one way or another. In the phase of objective evaluation, the primary goal is to quantify selected aspects of the writing. Teachers may choose from among five objective measures that can be used to evaluate essays. Kameen, P.T. (1983) lists these five measures as: (1) length, (2) subordination and relativization, (3) sentence connectors, (4) number and types of errors, and (5) T-unit which is the number of errorless T-units. With the help of these approaches to the selected features, teachers can arrive at a reliable rating of assigned compositions (Homburg, 1984).

Contrary to the use of objective measures to evaluate a student's writing, Perkins, (1983 cited in Bacha, 2001) claims that objective measures are not valid measures of writing quality associated with the communication of meaning, and they are not adequate to quantify factors such as cohesion, coherence, organization, tone and focus which contribute to effective writing. He concludes that objective measures are "impractical, tedious, and time consuming for classroom use" (p. 662). Instead, he suggests three types of subjective ratings for essays: (1) holistic (2) analytic and (3) primary trait.

A language teacher who is following a holistic approach uses one rating scale to assign a grade; one following an analytic approach rates selected aspects of the writing such as organization, wording, or ideas independently. Finally, a teacher following a primary trait approach evaluates characteristics unique to the particular audience and purpose of the writing (Bachman, 1991).

Of these three, two recent trends, the "holistic" and "analytic" approaches, are mainly discussed in the literature related to the grading of compositions. Both of these trends have been widely used in grading writing papers and both of them have supporters claiming that holistic grading is superior to analytic grading or vice versa. As a result,

writing programs need to choose either holistic grading with its focus on the total product rather than on separate aspects of the student's work, or analytic grading with its focus on each pre-determined quality of the paper not on the total product. Certainly, the choice of which type of grading to choose rests with the school and/or the teacher and depends upon the writing course objectives. Not only the school, but also the evaluators should make these objectives and type of grading style the focus of attention while testing and evaluating; otherwise, serious problems could be expected to emerge in the crucial merits like validity and/or reliability of the given test.

1.1.1. Problems in Grading Writing Papers

In the field of language testing, the classification of productive and receptive skills has been carried out with great sensitivity so as to clarify their methods of evaluation. As commonly known, receptive skills, which are reading and listening, cannot be directly observed in test performance, since the processes of comprehending take place within the mind. In contrast, the productive skills, which can be named as speaking and writing, are directly observable in the test performance in which the student reflects his/her competence. In the light of this classification, a simple definition of writing which defines this ability as a productive skill in the written mode should be made. A more detailed definition for writing was made by Boughey (1997), who referred to it as a form of communication in the written code exclusive to humans. Heaton (1988) lists the following skills that a good piece of writing should entail:

- language use: the ability to write correct and appropriate sentences;
- mechanical skills: the ability to use correctly those conventions peculiar to the written language- e.g. punctuation, spelling;
- treatment of content: the ability to think creatively and develop thoughts, excluding all irrelevant information;
- stylistic skills: the ability to manipulate sentences and paragraphs, and use language effectively;
- judgment skills: the ability to write in an appropriate manner for a particular purpose with a particular audience in mind, together with an ability to select, organize and order relevant information (p: 135).

A quick glance at these skills reveals that successful writing is more complicated than it seems, and often seems to be the hardest of all skills, even for native speakers of a language, since it involves, as Schoonen, Vergeer and Eiting (1997) state, not only a

graphic representation of speech, but also the development and presentation of thoughts in a structured way. If the major concern of a language is to communicate, an activity which is very special to humanity, expressing thoughts is of course an important quality to be searched in students' writing exams testing their mastery in a foreign language

Furthermore, the process of evaluating writing is even more complicated, since a quantitative degree, which is troublesome in productive skills, is necessary to identify the degree of success in the language learning/teaching process. The final results of language tests are most often presented in quantitative degrees or scores, which the test users benefit from. Such scores or grades are beneficial in language testing since they are commonly used in making decisions about students, methods and the intended goals of the language program (Bachman-Palmer, 1996). However, scoring or grading writing has been a difficult issue in language teaching programs for quite a long time. Grading writing involves subjective scoring in which the "human" factor appears to be dominant in the grading process since it is impossible to include all possible answers in an answer key. Thus, an acceptable level of reliability is expected from the grader and/or the criterion to evaluate the success of the written task.

In contrast to such expectations, a considerable amount of research, which addresses the problem of unreliability in the evaluation of EFL writing, exists. A number of studies (Bacha, 2001; Homburg, 1984; Hamp-Lyons, 1995; Connor-Linton, 1995; Sweedler-Brown, 1993; Ruetten, 1994; Turner-Upshur, 2002) conclude how extremely unreliable graders are – not only in their own inconsistency but also in their failure to agree with the other graders on the scores of EFL students' written works. Studies on graders' reactions to error in EFL writing reveal that students are graded by different standards developed by the graders themselves. Sweedler-Brown (1993) reports that most of the EFL writing graders he investigated have influence (which is impossible to be controlled completely) on the grade of a student by their perception of overall writing quality rather than the determined qualities in the grading criterion. Research reveals that graders may grade the papers on what a student has written, on what they believe the student meant when he or she wrote, or on some previous knowledge which was not the primary concern (Bachman, 1991). Thus, two graders who are grading the same paper may differ

enormously, whereas, they were supposed to be consistent as a consequence of using the same grading criterion (inter-rater reliability). The problem of reliability also appears when the same grader reads the same paper after some time. There can be no guarantee that the same paper read by the same grader with the same criterion will be awarded the same score (intra-rater reliability).

Since reliability is one of the most important characteristics of a good test, Brown (1996) goes so far as to call it "vital", as the qualities of inter/intra-rater reliability appear to carry primary importance in the process of grading. Brown (1994) asserts that a consistent and dependable scoring may be achieved by high inter/intra-rater reliability. Moreover, he proposes that in tests of writing skills, scorer reliability is not easy to achieve since writing evaluation involves numerous other aspects (p: 254).

In addition to the problems which focus on the amount of reliability, researchers have also investigated the influence of various types of scoring and their degrees of reliability in competency exams (Janopoulos, 1995; Upshur-Turner, 1995; Bachman, 1991; Nunn, 2000; Schoonen, Vergeer and Eiting, 1997; Ruetten, 1994; Song and Caruso, 1996; Sweedler-Brown, 1993; Homburg, 1984; Bacha, 2001). For the evaluation of proficiency exams of writing, two types of scoring can be stated as widely used; analytic and holistic grading. Of these two, the latter has been harshly criticized for more than a decade in terms of reliability. Elbow (1993, cited in Song-Caruso, 1996) criticizes holistic grading for being "unreliable, uncommunicative and harmful to the atmosphere for teaching and learning" (p:164) and advises the use of analytic grading in place of holistic scoring. Brown (1994) also states that a careful design of an analytic grading criterion can increase the scorer reliability better than a holistic instrument.

In contrast to these studies, Miller (1997, cited in Oruç, 1999) claims that holistic scoring presents a reliable and timesaving way of grading students' papers. Homburg (1984) reports the same advantage of holistic scoring and considers this type of grading as adequately reliable and valid (p: 103). Literature reveals that considering the process of grading students' writing, holistic scoring can be a reliable way that enables the group of readers to discuss different issues and consequently form a consensus on the problematic

ones. Moreover, Hughes (1989, cited in Oruç, 1999) recommends holistic scoring as it enables a rapid grading time that lasts a couple of minutes or less if the graders are experts. Bacha (2001) also focuses on the importance of issues like expertise and training in grading writing and supports the idea that higher inter-and intra- reliability values can be attained from holistic scoring through rater training and experience.

To check the validity of the above thesis and to have the reliability level of such a scoring system, the holistic-analytic instrument that is currently being used at Anadolu University School of Foreign Languages was tested. By ten graders, a number of 50 papers were graded twice with the holistic-analytic instrument which has three distinctive components including task achievement, essay organization and accuracy of the written skills. The scores after two grading sessions were compared in terms of inter/intra-rater reliability. The level of intra-rater reliability was found to be satisfactory (around 0.74) whereas the inter-rater reliability level, which is in fact the main concern of the institute, was found rather low (0.38).

Moreover, a considerable variation was observed on the scores (differences ranging from 10-60 points) given to a certain paper with the holistic instrument by ten graders. Having such dissimilar scores on papers, it was inevitable to question which component or components caused this variation among the graders. Thus, the factor analysis of all the scores assigned by 10 graders was done for the components of the holistic-analytic criterion.

Literature permits maximum 10% difference among the scores assigned to a single component. In addition, if a standard deviation testing procedure is implemented, the value of 1.00 is the maximum tolerable degree (Traub, 1994). For the holistic-analytic criterion, the values of standard deviation calculated for each component were significantly higher (task achievement 5.28, essay organization 6.065, accuracy 3.05) than the tolerated level (See Appendix O). The proportion of difference for each component was also calculated higher (task achievement 40%, essay organization 47%, accuracy 46%) than the permitted amount (See Appendix O). In addition to these calculations, the ANOVA test was

implemented on the sores of graders to see if there was a significant difference among their scores for each component at the significance level of .05.

The findings of ANOVA revealed that there was a significant difference among the scores assigned to each component; in other words, a consistent scoring could not be achieved among the components of the holistic instrument; therefore, great score differences were found at the end of that scoring procedure. Given that the use of the holistic instrument may cause huge differences among the scores, a new way of grading was sought, and analytic grading was determined as an alternative way of evaluation. Literature reveals that analytic scoring guides have superiority against holistic type even when inexperienced graders are the raters. Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981 cited in Sasaki-Hirose, 1999) suggest the use of on analytic instrument for measuring the success of writing where inexperienced teachers are the graders. In their study, Jacobs et al. developed an analytic criterion to test a number of compositions of an English L2 proficiency test battery with inexperienced teachers. The results were considerably successful and the new instrument was found to be fairly reliable. In addition, Ruetten's (1994) recent study on the success rate of ESL students as compared with native English speakers on a writing proficiency exam graded by both experienced and inexperienced teachers confirms the fact that despite being time saving, holistically scored competency exams are difficult for ESL students to pass (p.94).

Moreover, the use of a holistic-analytic criterion is not suitable for the assessment needs of Anadolu University School of Foreign Languages English Preparatory Program, since such (holistic based) scales were empirically proved to involve rater believes much more than analytical scales (Sweedler-Brown, 1993). When it is remembered that most of the raters who grade the writing papers in this school are inexperienced and not trained for language testing, the scale which seems to have a great influence on raters and the control of their ratings requires more importance. Thus, the use of an analytic scale may provide better control over raters' judgments, since it provides more explicit and detailed instructions.

In spite of a vast amount of support in the literature for the superiority of analytic grading, Brown (1991) suggests further study on an analytic rating scale, which is designed to produce separate scores for different language features. He claims that a better and more reliable way of grading could only be found by means of further empirical studies, as there is no "perfect" instrument to measure the success in productive skills. Also, a great number of researchers (Homburg, 1984; Hamp-Lyons, 1995; Connor-Linton, 1995; Sweedler-Brown, 1993; Ruetten, 1994; Turner-Upshur, 2002) agree with Brown and support the theory that it is impossible to call an instrument "the best" since all the instruments of measuring writing ability will be insufficient considering the fact that it is impossible to include all the possible answers in the criterion. Although, analytic scales seem to be more advantageous than holistic ones on account of their detailed content and superiority in considering the written work from different aspects, it is still often asked whether they can be proven empirically to be reliable.

1.2. Aim and Scope of the Study

This study aims to develop an alternative-grading criterion to the one that is currently being used at Anadolu University School of Foreign Languages English Preparatory Program. With this aim in mind, both intra and inter-rater reliability values of the new criterion will be measured and if the results of the new criterion appear to reflect higher reliability values, this alternative criterion will be suggested for the use of Anadolu University School of Foreign Languages English Preparatory Program.

Anadolu University School of Foreign Languages English Preparatory Program was established in 1998 in an attempt to provide foreign language education. This education is thought to be helpful for its students' further academic careers or future professions. The school has more than 1700 students and 100 lecturers at present. A skill-based program is being run and students take 4-hour-writing courses at 6 different levels ranging from beginner to advanced. A bottom–up process is implemented in the writing syllabus which enables students develop from producing the elements of a simple

paragraph to complex essay structures. At the end of each year, students take final exams, which involve writing an essay, evaluated with Oruç's criterion (See Appendix C) adapted from her previous holistic grading instrument (Oruç, 1999) (See Appendix D).

The graders who accepted to participate in the study were volunteer instructors working at the same school. Since the task they were supposed to achieve was not an easy and simple one, the number of the available teachers was no more then 10. This number can be stated as a limitation, because better statistical analysis could have been made if there had been more graders to score the papers. The more the number of graders, the better a comparison for a measurement instrument could be made.

What is more, at least 3 years of experience in grading writing papers was another qualification seen as a condition, because it is commonly believed that a considerable amount of experience is crucial to measure the inter-rater reliability levels. This experience-limit (at least three years) narrowed the possibility of more graders, which led to the grader sample being limited to 10. Finally, all the graders were quite busy with their own duties at work; therefore, they might not have been able to focus on the papers at the required level because of the lack of time.

The students' papers used in this study were chosen from the June 2000 Final Writing Exam Papers. The number of these papers was 50 and included 14 very low graded papers, 20 average graded papers and 16 high graded papers. Their distribution as 14, 20 and 16 may be considered a limitation, but having the biggest amount in the average group was on purpose, because it has been proved to be harder to have more reliable papers in an average group than the others. Moreover, it would have been more extensive if the total number of papers had been increased. Having the comparison results of two different tests applied on 100 papers would surely be better than applying them on 50 papers. In addition, having the risk of graders being too familiar with these papers and reading the same papers three times with each instrument may be seen as a limitation. However, if the concern was to create the same grading atmosphere or environment, reading the same papers was crucial not only for the grader but also for the researcher.

To conclude, not only research on grading EFL writing but also statistical findings indicate that holistic grading is unreliable; therefore, a more reliable way of grading students' writing papers will be sought. This leads on to the necessary step of developing an alternative grading instrument, and finally, the newly developed instrument will be tested in terms of reliability. If the results prove that there is a significant difference of reliability of the new instrument, in other words, if the new criterion appears to be more reliable than the previous holistic-analytic criterion, it will be recommended for the evaluation of the writing papers of the students of Anadolu University School of Foreign Languages English Preparatory Program.

1.3. Variables

The variables in this study are :

Dependent Variable : Final scores of the writing papers.

Independent Variable : The criteria used in the grading process.

1.4. Research Questions:

The study aims at answering the following Research Questions:

1- What are the intra / inter- rater reliability values of the new analytic criterion suggested for the writing assessment of Anadolu University School of Foreign Languages English Preparatory Program?

2- Is there a significant difference in terms of reliability between the holistic-analytic criterion (the one presently used at Anadolu University School of Foreign Languages English Preparatory Program) and new analytic criterion?

a) Which criterion has the wider range between the maximum and minimum mean scores among the ten raters' overall grades?
b) Which criterion has higher correlation coefficients after all gradings?
c) What do the long-term reliability calculations indicate? Does a significant difference appear when the long-term scores are compared with the previous ones?

1.5. Definition of terms

- Objective scoring: Scoring procedures for test items which do not require markers to make subjective decisions. All acceptable answers are clearly specified in a scoring key, thus inter- rater reliability should be perfect.

- Subjective scoring: Any scoring procedure which involves the exercise of judgment by the scorer where all possible answers are not specified.

- Holistic Scoring: A method of subjective scoring often used in the assessment of speaking and writing skills where a single score is assigned to the final product.

- Holistic-analytic scoring: A method of subjective scoring used in the assessment of writing skills through adaptation of some components on a holistic scale.

- Analytic Scoring: A method of subjective scoring often used in the assessment of speaking and writing skills where a separate score is awarded independently for each of a number of features such as organization, content, language etc.

- Criterion: A list of determined quality on which test performance is judged.

- Descriptor: A single word, phrase or sentence used to identify specific information in any kind of instruction or guidance.

(Quoted from The Cambridge Dictionary of Language Testing, 1999)

CHAPTER II

REVIEW OF LITERATURE

2.1. Writing in EFL

A visible increase in the use of English as a foreign language (EFL) in the world has been quite noticeable for a long time. Many reasons could be mentioned here for this incredible spread of English; however, 'the need to communicate' is probably the most favorable and reasonable one among the others. Being the international language of the United Nations, the official language of command, aviation or unofficially the language of sport and science, English as a language presents an inevitable tool for communication for one-fourth of the world's population's day-to-day needs (Harris&McCann, 1994). Considering this vast amount of usage, teaching and learning EFL is inevitably the center of focus of many institutes and schools. Thus, a great amount of research has been conducted in order to clarify needs while setting goals in EFL programs.

When people learn a foreign language like English, they learn to communicate with other people: to talk to them, understand what they say, read what they have written and write to them. For language learners, a crucial part of participating fully in a setting where a foreign language is the medium of communication is learning how to communicate when the other person is not right there in front of them, listening to their words and looking at their gestures and facial expressions. University students or academicians will often have to take language proficiency exams like TOEFL, fill out forms for some meetings or write abstracts in English. In other words, EFL learners have to write in English for different reasons, and writing (which no language programs dare to ignore), appears to be one of the four important skills (reading, writing, speaking and listening) to be gained in the language- learning process.

At this point, there is the need to define the term "writing" and its importance in language learning, because many researchers consider 'writing' to be more than a simple system of graphic representation of speech.

2.1.1. What Is Writing?

Literature presents a vast array of definitions of the term 'writing'. Hedge (1988, cited in Bilash, 1998) defines writing as:

> "a complex set of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer" (p: 163).

This definition stresses the very basic usage of writing as a graphic setting. In this sense, a broader definition was made by Porto (2001). She defines writing as:

> "a well-planned system of visible or tactile signs used to represent units of language in a systematic way, with the purpose of recording messages which can be retrieved by everyone who knows the language in question and the rules by virtue of which its units are encoded in the writing system "(p:40).

The term 'tactile signs' was used in the above definition, since all writing systems use visible signs with one exception: Braille, the system of raised dots used by visually impaired people. Of these two definitions, the latter presents a broader sense than the former, yet both of them still remain very basic since they ignore the purpose of writing.

Gannon (1985, cited in Heaton, 1988) argues that no one definition of writing can cover all the writing systems that exist and have ever existed. Instead, he states that a 'complete writing' system should fulfill all the following criteria:

a. complete writing must have as its purpose communication;
b. complete writing must consist of artificial graphic marks on a durable or electronic surface:

c. complete writing must use marks that relate conventionally to articulate speech or electronic programming in such a way that communication is achieved.

Gannon's definition and requirements of writing seem quite satisfactory at the first glance; however, the very important factor 'reader' is forgotten or ignored. To include this factor, another definition from Pollitt & Murray (1996) can be mentioned here. They define writing as:

> "an instrument for conveying ideas from one mind to another; therefore, the writer's job is to make the reader(s) understand the meaning quickly and precisely "(p:75).

Schallek (1999) however, claims that writing is a kind of storage exclusive to humans and comes up with the definition of writing which seems to be superior to the other definitions, since it is clear, concise and all encompassing. According to him, writing is a form of communication, which is special for human beings. Humans developed this form of communication because writing presents many advantages than speaking. First, writing enables humans to store their thoughts, thus the forgetful nature of minds becomes less of a problem. Second, it allows humans to communicate to their descendants so as not to 'reinvent the wheel'. Third, writing gives humans another way to express themselves; for instance, to read a poem is not the same as to hear it spoken aloud.

All in all, in the light of all these definitions, the basic and final definition of writing can be stated:

> "writing is a set of ideas, feelings, directions, instructions, descriptions and thoughts of humans legibly marked on some media that allows another or one's self to review and interpret the markings at a later time in order to accomplish one or many tasks and to persuade or inform the intended reader" (Schallek, 1999, p:2).

Thus, it is easily understood that writing is obviously a tool for communication and one of its most important functions today is to allow one person to reach out to another and persuade their thoughts. Yet, the fact that people frequently have to communicate with

each other in writing is not the only reason to include writing as a part of foreign language teaching program, there are other reasons that make writing essential in language teaching/learning.

2.1.2. The Importance of Writing in Language Programs

Like many researchers in ELT, proponents of writing across the curriculum are quick to clarify that writing to learn is not the same as learning to write but different sides of a single coin; the two support one another (Sorenson, 1992; Nunan, 1998). Walker (1988, cited in Sorenson, 1992), defines these two parts the 'virtuous circle', she asserts that when foreign language teachers incorporate writing in all areas of the curriculum, students benefit in three ways:

. they have a resource for better understanding content
. they practice a technique which aids retention
. they begin to use the language better

In addition to these points, Raimes (1983) believes that writing helps the students learn a foreign language. She lists her reasons as: (1) writing reinforces the grammatical structures, idioms, and vocabulary that they have been learning; (2) when the students write, they also have a chance to be adventurous with the language, to go beyond what they have just learned to say, to take risks; (3) when they write, they necessarily become very involved with the new language; the effort to express ideas and the constant use of eye, hand, and brain is a unique way to reinforce learning.

A final benefit of writing in language learning/teaching is that writing reinforces and enables the students to think, and this leads us to the conclusion that the close relationship between writing and thinking makes writing a valuable part of any language course (Raimes, 1983 p:3). Literature verifies the theory that writing and thinking are related. Turrisi (2000) claims that 'the better, the more astutely and precisely you think, the better you may write; however, the more you write to learn, the better you may think'

(p:2). She defines writing as a kind of logical diagram of thoughts, a visible map of ideas and their organization.

Writing, then, supplies a model of thinking, and models can be modified to better suit their purpose; therefore, writing can be a useful logical exercise to foster thinking in language learning. Recent theories suggest that thinking and reasoning promote involvement in the language learning process; thus, the more a student is involved the better practice he makes. This is a fine example of the fact that writing has positive effects on language learning since it reinforces students' involvement, and encourages them to practice and produce. Although, the initial goal of an EFL writing program is to help students become better and more active language learners, there are also many other goals attached to such programs, as their students will need to write for many other reasons apart from practice.

2.1.3. The Goals of an EFL Writing Program

The goals of a writing program, like the goals of the other three language skills, may vary from institute to institute. Some institutes focus almost entirely on the language itself, some on communication, and others on both the forms and the message. In most EFL courses, language is taught sentence pattern by sentence pattern, with vocabulary being fitted in according to the situations used to illustrate the sentence patterns being presented. Even in courses designed on modern lines, there is a tendency for language to be presented as a number of separate items, related to situation or communicative act, and when writing is used to reinforce the work which has been initially presented, it often reinforces either at the direct sentence level, or in relation to dialogues or situations which are not those usually expressed through writing. Hence, it is the responsibility of the writing program to train students to produce sequences of sentences which express the intended meaning most effectively.

For a student in a writing program, the goal on a linguistic level is to learn to acquire the facility to manipulate grammatical forms accurately (Raimes, 1983); on the

communication level, the goal is to adapt the goals of the writer to the needs of the reader (Bachman-Palmer, 1996). In general, writers should consider the reader; the effect they want to achieve (informing, persuading etc.); the relationship they want to establish with the reader, and the use of grammar. Writing programs; therefore, should design such syllabuses that encompass all these qualities.

The writing syllabus of Anadolu University School of Foreign Languages English Preparatory Program was prepared in order to provide a writing course to meet a number of needs. This school strives to achieve the following objectives:

- to increase the English language skills of its students to the level needed to be successful in the advanced courses of the further Academic Programs,
- to provide detailed instruction in each of the four language skill areas: reading, writing, listening / speaking and grammar,
- to apply sound language teaching methods,
- to provide students with strategies that help them learn more effectively, and efficiently,
- to provide students with supportive services that help them overcome their problems in language learning.

In relation to the school's general language teaching objectives, the aims of its writing program are :

. to enable students to write formal and informal pieces in English,

. to provide students with the ability to communicate in a foreign language via writing,

. to make the students ready to write their application forms, CV's, special notes etc.,

. to train them to write a complete essay from sentence to paragraph to the complete essay.

The micro-skills involved in this writing program should also be listed in accordance with what they praise while evaluating. Thus, a competent writer should:

. make a text's main ideas distinct from its supporting ideas or information,
. make a text coherent, so that others can follow the development of the ideas,
. make the main sentence's constituents, such as subject, verb, and object, clear to the reader,
. use a style appropriate to the given task and audience,
. use vocabulary correctly,
. use orthography correctly, including spelling, punctuation and capitalization.

The evaluation cycle for the above skills in writing works according to the language program's objectives, testing instruments and grading styles. More or less, the objectives of language programs with respect to the goals to be attained in teaching writing do not differ extremely; however, the testing policies, kinds, phases, purposes and even the distribution of the scores of these tests may differ according to the objectives and intended goals of language programs. Thus, it is better to focus now on the term "testing" and its usage in foreign language evaluation, which will lead us to the process of testing and evaluation of writing.

2.2. Language Testing

Whether they realize or not, a great deal of language teachers' time, attention and effort is devoted to understanding or measuring the degree of their learners' progress in language learning. From the time when the issue 'teaching and learning a foreign language' appeared, a need to measure the degree of these processes appeared as well. Not only in language skills, but also in every cognitive effort we make, we feel the necessity of testing (Brown, 1994). Brown (1996) states that "in most of our daily routines, whenever we complete a difficult task, try a new method, or even when we buy a new T-shirt, we feel the necessity of testing so as to approve our hypotheses and make judgments about

them." (p:185). Thus, testing seems inevitable in any circumstances where a kind of competency is the center of attention.

Moreover, the necessity of testing is crucial in today's modern language classes where many new teaching techniques and materials are tried for the sake of improving the quality of language teaching/learning processes. With regard to such innovations, since measuring the degree of success is the only way to understand how valid a theory is, testing again seems to be extremely important and unique. Therefore, each language teacher needs to know what testing is and how it is classified.

2.2.1. The Term 'Testing' and Testing Types

In a broad sense, testing is simply defined, as a method of measuring a person's ability or knowledge in a given area (Brown, 1994). However, Heaton (1988) considers testing an issue which is much more complicated. According to him, a test has three distinct meanings that can be listed as:

> "**a**. a carefully prepared measuring instrument, which has been tried out on a sample of people like those who will be assessed by it, which has been corrected and made as efficient and accurate as possible using the statistical techniques appropriate to educational measurements.
> **b**. a short, quick teacher-devised activity carried out in the classroom, and used by the teacher as the basis of an on-going assessment.
> **c**. an item within a larger test, part of a test battery, or even sometimes what is often called a question in an examination" (pp: 5-10).

Brown (1994) states two additional functions of testing. According to him, testing is first a method in which a set of techniques, procedures and test items constitute an instrument of some sort. Second, testing has the goal of quantifying knowledge in other words, competence. A test samples performance but involves certain competence, too. Some techniques of testing, which could be named "informal", are rather broad and inexact, whereas others, which could be called "formal", are quantified in mathematically precise terms (Brown, 1994 p: 253).

Informal testing is widely used by language teachers in their everyday individual judgments which are difficult to measure, while formal testing offers effective techniques of evaluation, quantification and the ability of comparison within an individual or across individuals. Such formal tests are administered and scored under conditions uniform to all students, and they are used for a variety of purposes. Bachman (1991) groups these purposes into two broad categories; first, the results of language tests may be used to make inferences about test takers' language abilities or to make predictions about their capacity for using language to perform future tasks in contexts outside the test itself. Next, decisions such as selection, placement or proficiency can be made about language learners on the basis of what is inferred from their test scores.

Alderson (1990, cited in Bachman, 1991) states that tests can provide valuable information about an individual's competence, knowledge, skills, or behavior. He claims that tests can be used to:

.conduct needs assessments which determine whether a special program is needed, and if so, what kind,
. plan a specific content of a program,
. select those students who need a special program,
. determine when a student may no longer need a special program,
. determine whether students in the program are progressing as intended,
. determine which parts of a program may need to be revised,
. demonstrate program effectiveness (p:672).

In order to understand the role and uses of language tests better, there is also the need to mention the different types of tests, which are being used for various aims of measurement; namely, achievement, proficiency, affective and placement or diagnostic. According to the needs of a language program, these four types of formal tests provide the test user with specific types of information.

Of these four, achievement tests sample a student's current level of learning across a range of general skill areas and estimate what a student knows and can do in a specific subject as a result of schooling (Brown, 1996). Placement or diagnostic tests identify

students' strengths and weaknesses in specific content areas and determine how best to help the student to overcome any particular limitations (Lien, 1971). Affective tests measure student attitudes or interests such as how a student thinks about types of people, specific situations, experiences or other areas. Finally, language proficiency tests, the type which gains the center of attention in this study, determine how well a student is functioning in regard to a specific spoken and/or written language, and they can measure a wide range of language acquisition skills, ranging from those necessary for conducting basic interpersonal communications to those necessary for handling more difficult activities such as school learning (Bachman-Palmer, 1996).

Except affective tests, the other three tests (achievement, placement/diagnostic and proficiency) are widely used for making program-level and classroom-level decisions. Brown (1996) classifies these types into two general categories as norm and criterion referenced tests. The category of tests most useful for program level decisions consists of tests specifically designed to compare the performances of students to each other. These are called norm-referenced tests because interpretation of the scores from this category of tests is linked closely to the notion of the normal curve which is also known as the 'bell curve'. On the other hand, criterion-referenced tests are specifically designed to evaluate how much of the material or set of skills taught in a course is being learned by the students. Using criterion-referenced tests, the major goal of the institutes is to look at the performance of each learner comparing the curriculum at hand, rather than comparing the performances of learners to each other. As Brown (1996) suggests, these tests are often used to diagnose the strengths and weaknesses of students and they may also be used to assess achievement, in the sense of how much each student has learned. Thus, criterion-referenced tests are valuable tools for language testing since they are useful for deciding whether to promote a language learner to upper levels or classify them as competent or not.

The literature presents one more classification of language testing, based on the difference of the evaluation process of the tests measuring different language skills. Thus, tests are grouped as "objective" and "subjective" not only according to the skills they are designed to measure, but also according to the techniques, grading references and graders involved in the scoring procedure. Harris and McCann (1994) label tests as 'objective' or

'subjective' because of the way they are marked. According to them, an objective test can be marked by any person capable of interpreting and applying a marking key which gives the correct answers that are totally accepted as valid. Moreover, such tests can be marked easily and rapidly since they have only one correct answer or at least a limited number of correct answers. The answers are usually given by a tick, a cross in a box, a circle round a number or letter, or through writing of a letter, a number or a word at most. Typically, these tests take the multiple choice format or blank-filling to ease the measuring but no real linguistic judgments can be made by them, since there is no visible product of the language learner that is solely his/her performance of communicative skills. In terms of language testing, objective tests are frequently criticized on the ground that they are simpler to answer than subjective tests (Heaton, 1988 p: 26). For the sake of making the items clear, an objective test constructor generally feels the necessity to consider some points that may confuse the students, ignoring the need to prevent students from answering a question not by using their competence but by chance. Moreover, as Heaton (1988) also states, objective tests enable the students to guess, rather than give exact and accurate answers, since they present a number of possible answers.

Subjective testing, on the other hand, is not based on such simple counting, but depends on somebody's opinion, a judgment, or a decision about candidate performance (Heaton, 1988). The possible responses that test-takers give may vary enormously depending on their language use, vocabulary or personal experiences. Therefore, the grader plays an important role in deciding whether the given respond should be penalized or rewarded. In this sense, in order to make accurate judgments in subjective testing, the one who is to make the judgment should be trained, qualified and experienced, since he/she is the only one to decide whether the given answer deserves a good grade or not.

Having said all this, it must be clearly recognized, however, that the creation and setting of both kinds of tests is ultimately subjective, since the choice of items, their relative prominence in the test and so on depend on the knowledge, skill and judgment of the test-setter (Bachman- Palmer, 1996). Furthermore, evaluating a piece of language like a free composition is entirely a subjective matter, a question of individual judgment, and

involves some analytic procedures like grading content or language use which involve complete subjectivity.

It should also be mentioned that there are no theories to be found in the literature which claim that objective tests are regarded as measures of the students' communicative abilities (Ruetten,1994; Heaton,1988; Brown,1996; Kunnan,1995). Receptive skills (reading and listening) should successfully be tested by objective methods where the possible responses may be listed; productive skills (writing and speaking); however, are better measured using subjective methods, since they evaluate the actual performance and communicativeness. Nevertheless, that does not mean that subjective testing guarantees the accurate measurement of actual performance and communicative ability of a test taker; for this, a number of qualities are required.

2.2.2. Qualities of Subjective Testing

The experts of language testing (Brown, 1996; Heaton, 1988; Bachman, 1991) agree on a number of qualities which make a test "good". Brown (1996) lists these qualities as usability, validity, and reliability considering the immediate necessities that a test should be practical to use, should measure exactly what it is supposed to measure, and should give the same or exactly the same results every time it measures. Such qualities gain much importance when different language skills are involved in the testing process, thereby demanding distinct approaches to the testing of language learners' competence and performance. It is generally accepted that for both groups of skills -productive and receptive-, ensuring usability, validity and reliability at the same time is something truly hard. Thus, language testing inevitably involves making some kind of compromises between what is ideal and what is practicable for a certain language skill.

Since the aim of this study is to take the ability of writing in a foreign language as the primary concern, the qualities of subjective tests will be discussed in detail rather than the objective ones, because a considerable amount of researchers claim that the ability of writing is better tested by means of subjective tests, as the "human graders" are superior in

measuring communicative qualities to computers or "mechanic graders" (Ruetten, 1994; Sasaki-Hirose, 1999; Turrisi, 2000; Brown, 1996; Heaton, 1988).

2.2.2.1. Usability in Testing

For many EFL programs, a "good" test ought to be usable in terms of its cost, administration, time constraints, and ease of scoring. A writing test which is prohibitively expensive is impractical. Considering the limited budgets of language courses, tests should be designed so as to measure maximum items at the lowest cost. Thus, all the processes – design, implementation, and evaluation- in writing exams should be considered in terms of time (and therefore cost) involved in paying teachers to do the ratings. The ease of test administration is also important in making a test usable. The degree to which a test is easy to administer will depend on the amount of time it takes, the number of subtests involved, the amount of equipment and materials required, and finally the amount of guidance that the students need during the test (Brown, 1996).

A three-hour-writing test, including many tasks for the learners, with unclear instructions or vague terms will inevitably be difficult to administer. In addition, ease of test scoring is a crucial issue, since a test which was designed in a way that provides easy scoring is cheaper and likely to present similar results.

2.2.2.2. Validity in Testing

Homburg (1984) defines validity as "the degree to which a test actually tests what it is intended to test" (p:27). If a test claims to measure the ability to write in English, then it is valid if it does nothing else but test writing ability in English. If what a test is measuring is actually knowledge of grammar, then it is not a valid test for testing ability to communicate. This definition thus leads us to two very important aspects. The first is that

validity is a matter of degree rather than a unique quality. Tests cannot be classified as valid or not as there are degrees of validity, and some tests are more valid than others.

The other important aspect is that tests are valid or invalid in terms of their intended use (Doshisha, 2000). If a test is designed to test writing ability, but it also tests the student's vocabulary, then it may not be valid for testing writing, but it may test writing and vocabulary at the same time. Therefore, validity of a test must be carefully examined to avoid testing something different from that which the test was intended to measure. Eventually, the validity degree of a test affects the value of not only the test itself but also the consistency of the language program as well. However, a high degree of validity can be provided not only by means of careful studies but also with some degree of reliability.

2.2.2.3. Reliability in Testing

Nunn (2000) defines reliability as "the actual level of agreement between the results of one test with itself or with another test" (p:3). Brown (1994) calls a test as reliable if it is consistent and dependable according to the results it presents. To design consistent and dependable tests, language testers should first understand the potential sources of consistent and inconsistent test score variance. In objective tests, the source variance that affect the degree of reliability (like validity, reliability is reported with numbers between 0.00 and 1.00; the higher the number, the more reliable the test) is known to appear from the selection of specific items and time of testing. In such cases, methods like test-retest or split-half are used to measure the reliability. In the split-half method, test items are randomly assigned to two groups and the results are compared; in the test-retest technique, the same test is being given to the same group of students twice within a considerable amount of time (at least 15 days), and the results are measured in terms of reliability. However, both of these methods are especially effective for measuring the reliability of objective tests; for measuring the reliability of subjective tests, the main concern of variance is the grader and the grading criterion. To measure the reliability of subjective tests such as tests of written essays, inter-rater and intra-rater reliability values are important. The actual level of consensus between two or more independent graders in their

judgments of language learners' performance is called inter-rater, the extent to which a particular grader is consistent in using a grading criterion is called intra-rater reliability (Bachman and Palmer, 1996).

In the process of measuring inter-rater reliability, the raters or graders should not communicate and exchange their ideas in terms of grades given to students' works. To ensure intra-rater reliability, the order of the papers/performances should be changed on each occasion and the grader should not have access to the original set of scores so as to guarantee that former grades were totally forgotten (Cumming, 1997 cited in Clapham and Corson, 1997). Furthermore, the sensitive measurement of those qualities is extremely important, since many researchers (Upshur-Turner, 1995; Pollitt-Murray, 1996; Brown, 1996) state that subjective techniques in testing foreign language learners' performances generally present higher validity but lower reliability values when compared with objective techniques, unless they are proved to be reliable in terms of inter/intra-rater reliability degrees. This is a natural consequence of the complicated relationship between validity and reliability in language testing.

It should be clarified that a test can be reliable without having high validity values; however, if a test is valid, it must also be reliable. A test that gives different results at different times cannot be valid while it is possible for a test to be reliable without being valid. That is, a test can give the same result time after time but not be measuring what it was intended to measure. At this point, it is better to give an example used by Brown, (1996) concerning the conflict of reliability and validity in language testing:

> "if the TOEFL were administered to a group of foreign students as a test of their abilities in maths, the reliability would be high because the test would spread the students out rather consistently along a continuum of scores. However, the TOEFL is clearly not valid for the purpose of testing mathematical ability. This is not to say that anyone ever claimed that TOEFL should be used to test mathematics or that TOEFL is not valid for measuring proficiency in EFL. The point is that, a test can be reliable without being valid" (p: 231).

Heaton (1988) argues that the fundamental problem lies in the conflict of test designers' attitudes towards reliability and validity in subjective testing. He claims that the greater the reliability of a test, the less validity it usually has; for instance, while

communicative tests in productive skills are more popular and considered "to the point" since they provide validity ignoring the necessity of reliability in the long term, considering the need that an ideal test should be both reliable and valid. However, the problem waiting for the test designer is whether to attempt to increase the validity of a test known to be reliable or else to increase the reliability of a test known to be valid (Heaton, 1988, p: 165).

Furthermore, Purpura (1999) claims that if a test designer tries to increase the validity of a reliable test, he/she will have to change the items that make the test reliable, but it seems impossible to modify a reliable test into a valid one. However, it is possible to increase the reliability of a valid test by means of a carefully structured grading criterion with clear and concise descriptions of various characteristics of students' performance at each level- without damaging the very features that make the test valid.

The roles of the graders in this attempt are also important, as they are the ones to apply these criteria. An inclusive grading criterion enables the grader to identify precisely what he/she expects for each band and then give the most appropriate grade to the students' written work. In addition, graders are guided and instructed for both to be consistent in their scoring and changing their testing judgments into numbers with the help of grading criteria. At this point, it seems inevitable to focus on the term "grading", and on its importance and processes in language testing, including the issues of instruments and graders. This will enable us to divide the term "subjective testing" into more specific aspects.

2.3. Grading in Language Testing

For a great number of EFL teachers, one of the most unpleasant aspects of language education is the process of grading. This uneasiness seems less annoying if the language test is planned to grade objectively. However, many language programs tend to measure the ability of writing, or productive skills in general, by means of subjective methods. The

results of these tests are reported as numbers or scores, letters and sometimes as words such as: "excellent", "good", "satisfactory", "poor", and "failure". The method, in which the grades are assigned on a 100- point scale, with acceptable marks between 70-100 and failure below 70, is widely used in most EFL programs in Turkey.

Such grades, as Lloyd-Jones (1989) notes, are "the records of a teacher's evaluation of work by each language learner" (p:155). There is no doubt that it would be better for the language teacher if she/he did not have to worry about such records and could concentrate solely on teaching. However, these records or grades, so to speak, open or close many doorways in a student's life; thus, they are crucial, and should never be seen as simple numbers or records.

Lien (1971) defines grading as "a real judgment of the learner by its teacher" (p: 200). Grading an essay, homework or a pairwork in the class are all the tasks of a teacher that would lead to make a judgment about the value of the results. While doing these, the teacher may use either objective or subjective tests, but eventually both of them are still judgments of the teacher. Thorndike and Hagen (1981, cited in Lien 1971) state that judgments are always relative and a judgment about a student's performance is closely related to the school system and/or language program.

Nevertheless, it is a fact that there are variations in grading and reporting learners' progress evident among systems and teachers, which can easily lead one to question the value of grading. Unfortunately, just because of inconsistent grades, there are many instances in which a person's future has been radically changed by a misgiven grade. However, because of the fact that grades are the only quantitative tools of student's progress, they are inevitable and necessary not only for language programs, but also for individuals as evidence of achievement, allowing for competency, in communicating via a foreign language.

2.3.1. The Functions of Grades

Lloyd-Jones (1989) claims that grading language skills is an essential task since it represents an effort to record quantitatively the quality of the product of learners so that administrators can make policies about educational programs. Tests are given and scores are assigned for different uses of language programs. Clapham and Corson (1997; p. 297) list these uses or functions as:

. to give information on learner's progress,
. for promotion or graduation,
. to motivate the learner for school work,
. to guide through learning,
. to guide for educational and vocational planning,
. to honor,
. to collect data for curriculum studies.

Nunan (1998) classifies the functions of grades under four headings: (1) administrative, (2) guidance, (3) information, and (4) motivation functions. He proposes that these four functions are not separate and exclusive; they overlap. Grades can also be used in transferring students to other schools and, at the same time, may be used to provide the necessary information to school personnel. Another use of grades can be seen in awarding scholarships to learners to help them overcome their financial problems.

Consequently, each language program should decide on the basic functions or purposes it wants its grades to serve, and then the program itself should provide the ways and means by which these functions can be accomplished most adequately. For sure, any language school will prefer a certain type of grading system for each skill that is reliable and provides data for the staff better to assist the learner to advance toward intended goals. Considering all these necessities and the importance of grading, the focus shifts to the people who grade and the methods they use. How best to grade, or who grades best are crucial factors, since irrespective of all theories and suggested strategies, the final word is theirs.

2.3.2. Graders

Grading students' written work has always been an important part of a writing program as it reveals the actual level of students' performance as well as the programs' success. Graders of writing thus have extremely important roles, as they are the ones to measure and decide for a number of grades which will have a radical effect on a learner's education. Having a second task as graders, teachers of writing should strive to be as consistent and fair as possible, overcoming the negative factors that would cause them differ.

It is certain that writing teachers are trained and educated in different ways and in different schools. Each writing teacher has a distinct personality, different experiences, interests and skills. Their writing classes are thus not the same, neither are their syllabuses. This verifies the theory that their grading styles and the qualities that they praise would most probably differ while reading their students' papers (Boughey, 1997). For this reason, to form a basis where the graders of writing in a program should meet is of crucial importance. To reach an agreement between graders, two well-known issues demand the attention. The training phase and graders' expertise in teaching and testing are important aspects of the reliability of the grader.

2.3.2.1. Training Graders

Given that the grading process of a written work is a discipline which should have certain rules to be obeyed by the grader, it is advisable that teachers be trained according to a certain style or standards of grading. A glance at the literature shows that for more reliable results from writing tests, there should be consistency among graders, and a great part of this consistency could easily be accomplished by means of well-organized training sessions. Sakyi (2000) concludes that grader training and the use of the right criterion application should provide the writing staff with higher consistency and reliability.

Moreover, Cumming (1997) reports the problem of human interpretation in composition grading, and refers to a test development project intended to overcome possible inconsistencies, using carefully refined criteria and by training and maintaining pools of graders. Weigle (1994, cited in Schoonen et. al., 1997) also focuses on the importance of training graders of ESL compositions. However, she does not make it clear whether the graders who are trained have significant differences between them compared to the others.

Finally, Jacobs et al. (1981, cited in Oruç, N. 1999) stresses the importance of training and claims that training sessions are vital since the more the number of graders increases, the more diverse their grading standards are likely to be. The fact that great diversity in the grading results decreases the level of reliability, sufficient amount of time and effort should be spent for training and much practice. This leads one to conclude that this process should come to an end only after certain agreements and consistencies are uniformed (p: 28).

2.3.2.2. Expertise of Graders

Considerable research in the past decade has addressed a common set of concerns that affect the reliability and practicality of grading students' written performances. In this set, besides the importance of training the graders, expertise or the quality of "being an expert reader or grader" has been discussed by many researchers (Schoonen et al., 1997; Kunnan, 1995; Ruetten, 1994; Clapham and Corson, 1997; Bachman and Palmer, 1996). To improve the quality of grading writing, graders should have a vivid picture of what adequate writing looks like. A way to achieve this is having enough experience or being an expert rater, a point sometimes ignored, which is defined as having all the skills or knowledge acquired by much practice, study and training.

Schoonen et al. (1997) claim that the grader of writing should at least have some domain knowledge which can be interpreted as the knowledge of language use, text organization, appropriate adequacy in the mechanics of writing, and certainly

communicative adequacy in the particular foreign language. The misconception that anyone who is a member of the language teaching family is a potential grader to judge the quality of a written text, threatens the reliability of the assigned grades. In order to prove this theory, a number of studies have been done on the role that grader's expertise plays in grading essays, the findings of which suggest that this expertise affects the quality of the grading. McDaniel (1985, cited in Ruetten, 1994) reports that "the judgments of graders who were not trained and inexperienced in evaluating ESL writing were dominated by error when they graded ESL essays" (p: 87).

Cumming (1990, cited in Schoonen et al., 1997) investigated the grading reliability of six expert readers (with at least 4 years of experience) and seven novice readers (with no prior teaching experience) in ESL writing proficiency exams. He found that the grades given for content and organization parts between experts and non-experts differed significantly from each other. There were also large differences between these two groups when the total grades were completely counted. Kunnan (1995) concludes that expert readers who taught writing and graded writing for a couple of years seem to be more reliable graders than the novice graders since the experts are found to be more stable and show more agreement among each other.

In addition to 'grader qualities', the literature presents the need for the standardization of the grading process, since there is no guarantee that raters will be consistent (even if they are trained or/and experienced), unless they are provided with some tools to use as the basis. Standard grading criteria have thus been used for quite a long time in many different versions. The major goal in the use of such criteria is to avoid probable discrepancies of graders at a maximum level and to guide or give verbal definitions of the suggested qualities of students' performance. These criteria should keep a pool of graders on the same track and enable them have stable qualities or levels of performance by means of visible measuring tools. However, these criteria differ extensively in terms of grading techniques, length of grading time and, unfortunately, levels of reliability. Therefore, we now need to make further analysis of such criteria describing them in details, noting the advantages or disadvantages they offer in grading writing.

2.3.3. Grading Criteria

In recent years, writing tests have increasingly raised the issue of 'grading criteria' in the evaluation process of productive skills. A grading criterion should be defined as a practical means of testing the actual level of a student's communicative performance by using a number of descriptive bands for a particular skill, on a scale of competence ranging from excellence to failure (Nunn, 2000). Alderson (1991, cited in Nunn, 2000) reports the major reasons of using grading standards or criteria in testing:

> "Firstly, rating scales (grading standards) provide an easily understandable report for candidates, administrators, course designers, and teachers on the level of performance of individuals or groups at the same time as providing descriptions of what candidates can do. They can report on typical or likely behaviors of candidates at any given level or on the proportions of candidates at each level. Secondly, rating scales can guide the rating process standardizing the criteria for an individual rater or act as a common standard for different raters. Finally, they also help to guide the construction of tasks which allow students to display the described behaviors at their own level (p:171).

From the above reasons, it can be concluded that the main general advantage of developing and using grading criteria for productive skills is the exact harmony that can be achieved between the potentially conflicting and inconsistent perspectives and language views of graders, course teachers and administrators.

2.3.3.1. Types of Grading Criteria

Three common types of criteria; analytic, holistic and primary trait are found in literature when the issue is the ways of grading students' written works by means of exams. Since the analytic method is probably the most popular and the main concern of this study, additional detail will be given about it in this section.

Holistic scoring, despite many criticisms, is one of the popular ways of grading in which the graders respond to the whole essay by a quick reading and giving an overall score according to an established criterion that is introduced by a group leader. As Kroll (1998, cited in Connor and Mbaye 2002) states, "holistic scoring assigns a single score to a piece of writing based on its quality" (p: 16). Holistic scoring then, is a method of grading, which focuses on the whole rather than on parts. While using a holistic criterion, the graders periodically undergo reliability checks during the process in order to maintain the stated standards. In this process, first, a group of readers are formed, a criterion is created, a consensus among the graders on how to use the criteria is achieved, and nonconformist readers are retrained or dismissed (Miller, 1997; cited in Oruç, 1999). Raimes (1983) classifies holistic scoring into two types as: focused holistic and analytic-holistic. Focused holistic scoring centers on the product as a whole and produces only one final grade that is assigned to represent the student's work as a whole. Holistic-analytic, on the other hand, has two or three bands including specific information for some certain qualities which are expected to be achieved by the language learner. Despite their slight differences in their designs, both type of holistic scorings follow the same procedures and principles in order to measure the written performance.

Another type, primary trait scoring, involves the scoring of a piece of work (usually writing) in relation to one principal trait or characteristic specific to a certain task (Pollitt&Murray, 1996). This kind of criteria is mainly based on the view that a piece of work must be judged in relation to its specific purpose and content. This method begins with a writing assignment or task that requires rhetorical specifications that direct the writing process. When grading such tasks, graders score the degree to which writers met the specifications. If a task requires writers to select a topic and explain why it is relevant to their personal lives, graders would score the ability to select, explain and connect.

The third and final type is 'analytic'. Though it has three different types (point-off, band or rubric based, analytic-holistic), the analytic method is commonly known as a method where a separate score is given for each of a number of features of a certain task, as opposed to one global score in the holistic method. The analytic method, which was

originated by Diederich, French, Carlton (1961, cited in Sasaki and Hirose) who factor analyzed 53 English LI readers' remarks about 300 college compositions, includes a number of major traits which graders tended to value while grading. Diederich et.al., devised the first analytic criterion using a variety of traits gathered from the remarks of graders concerning their preferences. Their criterion, despite the criticism of its mechanical or unquantifiable content, pioneered the methodology of developing analytic criteria for grading compositions (Harmer, 1991).

In the field of testing writing, the superiority of these different types of criteria has always been a matter of argument. Of these three, analytic and holistic criteria were mainly the center of discussion since they have considerable differences in theory and practice. For some researchers (Miller, 1997; Hughes, 1989; Oruç, 1999), holistic grading has many advantages that no other way of grading can offer. However, recent studies (Hamp-Lyons, 1999; Cumming 1990, Upshur and Turner 1995; Schoonen et.al., 1997; Bacha, 2001) suggest that the analytic method presents a more reliable and consistent way of grading than the holistic method does. A better comparison can be made between these two grading methods by listing their advantages and disadvantages considering the stated problems present in the literature.

2.3.3.2. Analytic Versus Holistic Method in Grading

In order to make a vivid comparison between analytic and holistic methods, it is better to list the major advantages that each method offers. Sasaki and Hirose (1999) present four reasons that make the analytic method more advantageous than the holistic style. According to them the analytic method is superior because:

1. It is comparable with the criteria used for students' English L2 writing ability in courses.
2. It has proved to be more reliable than other types of scales.
3. It reinforces the grader's focus on the task rather than on the ideas presented.
4. It provides useful diagnostic information not found in other methods (p: 458).

On the other hand, Oruç (1999) presents the advantages of holistic grading as opposed to analytic grading under three sub-headings:

1. Time; having less detailed and relatively shorter bands, holistic scoring offers the reader a quite short time to grade a writing paper.
2. Emphasis is on communication; considering the student's work as a final product, holistic scoring focuses mainly on the given message and the way it was given. Communicative qualities, therefore, are regarded as of prime importance.
3. Rater agreement; since each paper, when graded holistically, is read by more than 2 teachers, higher rater agreement and reliability is expected.

Bacha (2001), however, claims that holistic scoring focuses on what the writer does well rather than on the writer's specific areas of weakness, which is of more importance for decisions concerning promotion. Thus, readers seem to have agreement on the student's performance level, since what they favor is only the positive sides of the paper. Kunnan (1995) also stresses the problematic and unreliable nature of holistic grading since it involves 'human instruments' (more than that of analytic grading) whose behavior cannot be completely understood. He concludes that results of empirical studies collectively show that both discourse-level and sentence-level features in an essay influence holistic scores, and these effects change according to the grader's view and the context of the particular study.

Another criticism against the holistic method arises because of the major concern which makes the 'product' more important, and takes it as the center of evaluation in holistic grading. A recent trend in the process of evaluating writing is to focus on the "process" as much as on the "product". Sweedler-Brown (1993) defines holistic grading as insufficient since it judges a product rather than a process. Unfortunately, there are two common trends in the attitude towards writing, both of which seem to support holistic techniques. From one point of view, writing is seen as an innate ability, that is one either can or cannot write in the mother tongue or in a foreign language. From the other point of view, to be able to write generally means only to be able to use the language correctly with

surface features of writing in which the learner is often said to be able to write if his final "product" shows that he can apply grammar rules and spell the words accurately. In teaching EFL, it is common to see that students' final products may show some degree of correctness with regards to the surface features of writing but very often lack content quality, organization or appropriateness. Thus, while grading students' written works, holistic methods should be thought about twice since they favor the "product", ignoring the "process".

Another weak point of holistic grading is related to the relative influence of rhetorical and sentence-level skills in the evaluation of writing. Song and Caruso (1996) report that sentence level error was the only significant influence on holistic score and was the critical factor in causing EFL students to fail. Moreover, Homburg (1984) asserts that writing variables, which took error into account, were the best discriminators among essay scores. Finally, McGirt (1984; cited in Sweedler-Brown, 1993) presents an interesting study that the pass rate for ESL essays raised from 20 % to 60 % when sentence- level errors were corrected. He found that lexical correctness was the best predictor, whereas organization was the worst, supporting the theory that sentence-level error has a significant influence on holistic grading.

Analytic grading, however, allows minimal rhetorical and sentence level effects, since a separate grade is given to a number of different features in the student's paper. Grading different features also reduces the 'human' factor in grading. While the analytic method governs the scoring within each quality of writing, the holistic approach seems to be unsatisfactory since there is no separate grading when reading a paper. The grader thus consciously or unconsciously decides on a grade mainly just because of some certain features.

Having detailed bands and descriptors, an analytic criterion is easier to use as it gives clear-cut guidelines in measuring writing ability. In contrast, Harris and McCann (1994) report a disadvantage of holistic scoring concerning the problems with descriptors. Students' written performances can often cut across the descriptions (e.g. one quality may belong to level A whereas the other belongs to level B), and in these cases graders feel

confused and try to create their own solutions that damage the level of consistency among teachers.

All the factors that were pointed out about holistic and analytic methods up to this section are of great importance in maintaining reliability among test scores. There is no doubt that in order to be fair and effective in language teaching, all institutes and writing programs try to adapt or develop reliable ways of evaluating their students' performance. However, there are a great number of researchers who claim that holistically scored competency exams are unreliable and risky in terms of validity (Brown, 1994, Hamp-Lyons, 1995; Sweedler-Brown, 1993; Heaton, 1988; Song and Caruso, 1996). In their study, Sweedler-Brown (1993) compared analytic and holistic scoring, considering the influence of sentence-level and rhetorical features. It was found from the results that the holistic method has lower reliability values compared with the analytic grading when sentence-level errors were corrected. Elbow (1993, cited in Song and Caruso) also finds holistic grading unreliable and harmful for testing writing since it includes many other things but consistent features.

A very recent study dealing with the effect of writers' and graders' personalities on holistic evaluation (Carrel 2002), reveals that the personality types of writers affect the scoring their essays receive, and the personality types of raters affect the grades they give to essays. Hence, unreliable results are inevitable in such cases where holistic scoring standards are used, but in analytic grading, problems including writers or graders personalities are at a minimum level, because factors like personality are not focused or ignored through grading.

Another important quality that an analytic criterion presents is the ease of giving feedback or diagnostic information to both the administrators and the students. Bachman (1991) states that analytic grading may also be used for diagnosis. The problems in the syllabus, disagreement among graders and points that need much attention can be clearly seen through analytic instruments. Students may also benefit from such a grading system as it provides a visible chart of what qualities were awarded (Upshur-Turner, 1995).

Homburg (1984), on the other hand, reports holistic methods to be an adequately reliable and valid way of grading writing. In his study, he discusses the relationship between subjective evaluation and objective measures of ESL writing proficiency. The relationship between the statistical procedures in his study confirmed that with training and expertise, results of holistically graded papers could be consistent. Reid (1993, cited in Bacha) also supports the idea that holistic instruments are reliable and useful tools in testing writing with various practical qualities offering new dimensions.

Hamp-Lyons (1995); however, criticizes holistic scoring and defines it "a closed system offering no windows through which teachers can look in and no access points through which researchers can enter" (p: 760). In addition, he adds that the grades which are given with holistic criteria cannot be explained to other readers in the same grading group; diagnostic feed back is out of the question. Thus, for the researchers of writing, holistic techniques fail in acting as qualitative research tools while analytic methods offer diagnostic details.

Consequently, Bacha (2001) recommends analytic evaluation instruments on which one can base decisions concerning the extent to which the students are ready to begin more advanced language courses. Brown (1996) for his part claims that it is the grader and /or its grading style that decides whether a student fails or passes the course. Grading students' writing papers is a complicated issue and it is the responsibility of the administrator or the head of the writing committee to ensure that grading is organized in such a way that validity and reliability are the main concerns to avoid possible discrepancies and fallacies which may lead to irreversible damage in a student's life. Thus, Brown (1996) recommends all language programs check or test their measuring instruments for different language skills periodically in order to assure language learners that their testing services really work. Moreover, the ever changing needs and developments of language teaching and testing methodology force the test designers to upgrade or re-develop their criteria considering the present problems of their program, teachers and students.

2.3.3.3. Problems Encountered in the Present Criterion

From the day grading criteria started to be used, research was made to suggest alternative instruments that offered higher values of validity and reliability. Parallel with increasingly communicative language teaching methods, writing tasks and scoring procedures need to be designed to keep up with the novelties in testing writing in a foreign language. Thus, writing teachers and other professionals using tests in instructional situations need applicable and handy grading standards to evaluate students' mastery of language use.

Unfortunately, the need for better grading instruments does not automatically lead to effective and efficient grading tools. Upshur and Turner (1995) report that commonly employed rating scales present major problems of reliability and validity. They present the most frequently registered complaints under two sub-headings.

Problems of reliability:

1- Raters of the same students will not agree on the meaning of descriptors. Therefore they give different scores to the same student performance. One teacher giving generally higher average scores than another can reflect this.
2- Published grading criteria of language performance are often too broad; only a portion of the criterion is applicable for a particular situation.
3- A criterion which was developed for a particular setting may have unwanted ceiling or floor effects when used on other students. That is, the description of the highest-grade category may be too high or too low for the new setting.
4- Because grading descriptors are not precise enough, a rater's standards may change during a single grading session.

Problems of validity

1- Grade descriptors often do not conform to a teacher's own objectives. Typically, descriptors list a number of features a performance must incorporate in order to receive

a given score. Teachers might not, however, have all those features as objectives in their teaching.

2- Descriptors often include features that do not occur in actual student performances; these features may be too difficult or too easy for the students.
3- Some descriptors reflect questionable assumptions about the order of importance; for example, one mechanical descriptor may be assigned a grade which is twice as high as an organizational quality (pp: 5-6).

Considering a number of sample grading criteria presented in the literature for public use, the above problems are recognized to be shared more or less by each of the grading criteria. Oruç's criterion (Appendix C) also includes a number of problems, some of which are common with the above ones.

First of all, a grading standard should be designed in such a way that it should carry the distinctive qualities of one of those criteria types. It should be holistic, analytic or primary trait, but ultimately it should make the grader aware of its type and features. Next, as Brown (1996) states, a certain type of criterion should not be adapted into another type without careful empirical research in terms of validity and reliability. However Oruç's study (1999), the holistic criterion was tried to be adapted into an analytic scale without any empirical research to justify it, though keeping the original form was very important. The third problem concerns the descriptors used in this criterion; because they were designed holistically, they give general senses that need further clarification. Holistic instruments are designed in such a way that their descriptors are intended to be kept as short and concise as possible in order to increase the speed of perception and grading, but this causes great problems when such criteria are adapted into analytic styles.

In addition, the amount of the penalty assigned in Oruç's grading standards for essays which have different topics or which accomplished totally different tasks, seems insufficient since it enables a student to pass who writes an essay which was prepared beforehand. Moreover, mechanics such as punctuation, capitalization and spelling are all ignored, despite the common trend in literature (Sweedler-Brown, 1993; Jacobs et al., 1981, cited in Bacha, 2001; Homburg, 1984) that recommends the scoring of these

qualities, and despite the effort spent on teaching and evaluating them throughout the year. Finally, since it carries the flavor of holistic features, the total grade given by this criterion is under the danger of reflecting some of the features that its grader favors, in other words a particular amount of "human factor" is involved in the grading rather than the focus on the performance itself since there is no accurate way of controlling what was valued. For instance, a student can be penalized or awarded for writing an essay about a certain politician, football team or a singer who is hated or admired by the grader him/herself.

Because a considerable amount of research favors the reliability and validity degrees of language test scores, and empirical measurement is of crucial importance, a number of reliability tests were implemented on the scores given with the holistic-analytic criterion after two grading sessions by ten graders. Due to the fact that a great diversity emerged among the scores of the graders, and the factor analysis of the constituents stress no exact components as the ill (since all of them were found to be problematic), the need for developing a new criterion emerged. This study initially aims to develop and measure the reliability values of an alternative-grading criterion, which was planned to be analytic. In the next step, not only inter-rater but also intra-rater reliability levels of the new criterion will be compared with the holistic-analytic one. Provided that the results of the analytic instrument have higher degrees of inter/ intra reliabilities, this new criterion will be suggested as a reliable grading instrument for the future use of the writing department of Anadolu University School of Foreign Languages English Preparatory Program.

CHAPTER III

METHODOLOGY

3.1. Introduction

Grading students' written performances has always been difficult given that literature presents no certain way or grading standards that should be used for any kind of language testing. Alderson, Clapham and Wall (1995, cited in Connor and Mbaye, 2002) claim that every language test has a theory of language behind it. Thus, it may be inferred that any test used in language programs is designed in a way that reflects (directly or indirectly) the objectives and theories very particular to that program. The issue of designing and grading language tests is therefore not only related with its test cycle, but also reflects the overall goals of language teaching programs. Considering this fact, developing its own grading standards should be more useful for a language teaching program than using or adapting one from the literature which was originally designed for a specific group of learners or for specific objectives which are usually different from the institute's own goals.

This study, therefore, aims to develop a grading criterion for the use of Anadolu University School of Foreign Languages, in the writing exams of its English Preparatory Program. There will be two main phases in this process; first, the number of components, their content and ratio in the new criterion will be clarified. Next, the reliability degrees of this new criterion will be identified and compared in order to elucidate whether the use of an analytic grading criterion will result in an increase in the inter/intra-rater reliability degrees of grading students' writing papers. In this chapter, all the steps of developing a new criterion, including the necessary materials, data collection and the graders as the participants will be given in details.

3.2. Participants

The participants of this study were chosen from the writing instructors of Anadolu University, School of Foreign Languages English Preparatory Program. All the participants had at least three years of experience in grading students' writing exam papers. Of the 19 writing instructors of this language program, 10 were found to be suitable in terms of experience in grading writing papers. All the instructors are non-native speakers of English. The ages of these instructors range from 27 to 35, and three of them were male while seven of them were female. All these instructors contributed to this study willingly.

3.3. Materials

The materials used in this study were sample papers, question sheets, two different grading criteria including holistic and analytic standards, and two questionnaires in order to form the new criterion according to the graders' suggestions.

3.3.1. Sample Papers

To check the consistency among the graders of the final writing exam, 50 papers (See Appendix B) were used to form a standard of input for the participants. They were given files containing 50 sample papers that were chosen from the June 2000 final writing exams of Anadolu University School of Foreign Languages English Preparatory Program. These papers were chosen according to their original scores; three categories (low, average, high) were determined, and 15 papers whose marks were ranging from 25 to 50 were included as the "low category". 20 papers whose scores were ranging from 65 to 80 were included as the "average" papers, and finally, 15 papers whose scores were ranging from 90 to 100, were regarded as the "high category".

To have 20 papers (which was more than the other two groups) as the "average" category was on purpose. Literature reveals that rater-reliability appears mostly in papers which are very close to the pass-line, or in the ones that have hardly passed the deadline (Upshur & Turner, 1995; Bachman & Palmer, 1996). Thus, to see that difference better,

the number of average papers was increased. The original scores were never declared to the participants and none of the scores given to these papers were compared with the others during the times of grading. Keeping the original forms, each paper was typed, and no other changes were implemented on the papers. The order of the papers was changed in each grading so as to avoid the negative effect of recognizing or remembering a certain paper. In each grading, the papers were given numbers and the graders named the papers by these numbers. The names of the writers of these papers were kept secret.

The papers that were used in the training session for the new criterion were also chosen from the June 2000 final writing exams of Anadolu University School of Foreign Language English Preparatory Program (See Appendix G). These papers were also classified as "low, average, and high" and they were typed keeping the original forms constant.

3.3.2 Question Sheets

Before grading the sample papers, the participants were informed about the tasks that the students were supposed to achieve. Including the questions of the exam, each instructor was given a paper about the instructions and topics that the students used in the writing exam (See Appendix A). The content of this paper was also explained in detail in the training session in order to avoid misunderstandings about the topics.

3.3.3 Grading Criteria

50 sample papers were graded 6 times using two different criteria in order to obtain the necessary data for this study. First, in the initial 2 gradings a holistic-analytic criterion was used; next, an analytic criterion was used in the next two gradings to see if there occurred an increase in the degree of reliability of the final scores of the papers. Later, all the papers were graded for the third time in order to have sufficient data to measure the rater reliability of these grading standards in the long run. Since the major focus of this study is to develop a new analytic criterion, the steps of developing the analytic criterion will be discussed in detail later, more than that of holistic-analytic scale.

3.3.3.1 Holistic-analytic Criterion

The criterion which was originally holistic, but adapted into a holistic-analytic figure for the needs of the language program, had been developed by Oruç (1999) (See Appendix C and D). There was no specific information on how the criterion had been developed or what qualities were focused on most in the development process of the holistic criterion. This holistic-analytic criterion was adapted as the grading standards of Anadolu University School of Foreign Languages English Preparatory Program, and since 1999 all the final writing papers of the students of this school have been graded by means of this criterion. It has 3 separate parts, identifying the characteristics of papers from best to worst under the headings, 'task achievement', 'essay organization', and 'accuracy of written skills'.

Next, if the validity of this criterion is questioned, it will be found out that as a common trend holistic-analytic instrument in grading writing is a rarely preferred means of scoring since literature recommends the use of one of the three techniques; holistic, analytic and primary-trait. A mixture of those or both of them is generally avoided since the nature of evaluation in each of these techniques is quite apart. In primary trait, a single item or purpose of writing is identified and evaluated; in holistic scoring, the final quality and the overall performance is assigned a total score. Finally, in analytic scoring, each component of writing that should be listed as a quality is graded distinctively. In other words, in holistic scoring, the forest as a whole is evaluated, whereas the trees, one by one, are considered cautiously in analytic scoring. Thus, the attempt to combine the above techniques in the measurement of writing is generally avoided by many experts (Brown, 1996; Gannon, 1985; Bachman & Palmer, 1196).

The criterion developed by Oruç is identified as a holistic-analytic instrument since it has 3 separate components which is, in fact, a quality of analytic scoring; however, Oruç's criterion is originally a holistic instrument (See Appendix D). The adapted version; therefore, looks like analytic but measures as a holistic instrument since it was originally designed to evaluate the whole. In addition, because their grading points are quite different, an originally holistic or originally analytic scale would never be adapted into another type.

Moreover, the inter-rater reliability results of the scorings with the holistic-analytic criterion were found significantly low (0.38), and the factor analysis test proved that all the components of this criterion were problematic. For that reason, a thorough revision on the criterion was found to be necessary since each component revealed a great diversity (task achievement 40%, essay organization 47%, accuracy 46%) among the scores of a group of graders. The experts regard such an inconsistency as not surprising since the issue of reliability can be conducted only when an exact validity is achieved (Heaton, 1998; Brown, 1996; Purpura; 1999). Finally, not only the literature but also the statistical findings directed us to a result that Oruç's criterion that is currently being used at Anadolu University School of Foreign Languages English Preparatory Program is not a valid and acceptably reliable instrument in grading students written works. Taking this result into account, the development of an alternative criterion was inevitable.

3.3.3.2 Analytic Criterion

The other instrument that was used to measure the students' writing papers was a newly developed analytic criterion. It was designed with the aim of grading the separate qualities of students' writing abilities rather than giving a global score that is usually done by means of holistic scoring. Considering the common procedures employed for developing an analytic criterion mentioned in the literature, it was decided to proceed according to the following five steps to form the final criterion.

Step 1: The first thing to be made clear was to determine the qualities that would be graded in students' papers. Despite various conflicting suggestions, Sasaki and Hirose (1999) present twelve qualities which were acquired from their questionnaires responded by 102 language teachers in Japan. They suggest: organization, content, language, social awareness, vocabulary, appeal to the readers, spelling, punctuation and capitalization, neatness, hand writing, transitions and finally the title of the essay as the probable features of grading in EFL learners' written works.

Step 2: These qualities were included in a questionnaire and given to the graders of writing papers (See Appendix F). The graders were asked not only to determine the ones

they favor, but also state their possible weightings for each quality. At the end of the questionnaire, a list of the favored items was added so as to summarize the process for the contributor.

Step 3: To shape the first draft of the new criterion, the data collected from the graders by means of the questionnaire were closely scrutinized. First, the qualities on which the majority of the graders agreed were identified (the ones at least half of the raters agreed on), and then their probable score distributions were calculated by taking the mean score of the overall scores (See Table 3.1). Finally, a sample criterion was shaped in the light of these findings.

Table 3.1: Graders' suggested score distributions for the analytic criterion

	Grader 1	Grader 2	Grader 3	Grader 4	Grader 5	Grader 6	Grader 7	Grader 8	Grader 9	Grader 10	**Final Criterion**
Organization	20%	25%	25%	25%	20%	30%	25%	30%	25%	30%	**25%**
Content	30%	25%	20%	20%	20%	20%	25%	30%	25%	25%	**25%**
Language	10%	15%	20%	15%	20%	15%	10%	15%	15%	15%	**15%**
Social awareness	---	---	---	---	---	---	---	---	---	---	---
Vocabulary	10%	10%	10%	15%	15%	10%	15%	10%	10%	10%	**10%**
Appeal to reader	---	5%	---	---	---	---	---	---	---	---	---
Spelling	5%	5%	5%	5%	5%	5%	5%	5%	3%	5%	**5%**
Punctuation and capitalization	10%	5%	5%	5%	5%	---	5%	3%	10%	5%	**5%**
Neatness	5%	---	---	---	---	---	---	2%	---	---	---
Handwriting	---	---	---	---	---	---	---	---	---	---	---
Transitions	10%	5%	10%	10%	10%	10%	10%	5%	8%	5%	**10%**
Title	---	5%	5%	5%	5%	10%	5%	---	4%	5%	**5%**
Penalty (different topic)	20%	40%	40%	30%	25%	20%	25%	30%	30%	25%	**30%**

Step 4: In the fourth step of developing the new criterion; the draft, which had appeared just after the mean calculations of the suggested items in the questionnaire, was discussed with two other writing teachers who were not only the instructors of the same language school, but also the coordinators of that writing program. The bands of each part and their wordings were made clear and the first draft of the criterion was reshaped by adding the necessary details including the point-ranks. Purpura (1999) points out that the design of each band in a grading standard is as important as the decision of what to grade. Therefore, considering not only the program coordinators' but also the participants' suggestions, each band was designed in a way that was suitable both to the goals of the language program and students' language competence. For instance, in the language part of the new criterion, the highest point band was set in view of the fact that the students were non-native users of that language and should not therefore be harshly penalized for simple grammatical errors that do not interfere with meaning.

Step5: In the last step of the development process, the latest draft of the criterion was pilot-tested with another pair of writing teachers from the same school. An essay, which was also taken from the June 2000 final writing exams, was graded by these two teachers after a brief training on how to use the new criterion with the intended goals of the whole study. After grading the paper, it was seen that the grades of each teacher were quite consistent. However, according to their suggestions which mainly dealt with practicality, a number of further modifications were also accomplished (e.g. the number of bands in vocabulary and transition sections were reduced and all the numerical markers of each section were equalized) but, the final version of the criterion still had some problems which were also discussed and mostly solved in the training session through the data collection process.

3.4. Data Collection Procedure

The data of this study were collected in six phases which were carefully planned and implemented parallel with a time-table, which the literature reveals as an important factor while measuring intra-rater reliability values (Brown, 1996).

Phase 1: At the very beginning of the data collection process, necessary permissions were taken from the Head of the Basic Languages Department, and writing co-coordinator, in order to use the final writing papers of the institute. Next, the graders who were determined beforehand were asked to contribute to the study and the ones who agreed to participate were informed about what they were supposed to do.

Phase 2: 10 graders who were listed after the first phase were called to a meeting for the preparations of the first grading. In the meeting, they were given details about the scope of the study and an extensive workshop was done to practice for the first grading with the old holistic-analytic criterion. Being quite familiar to use this scale, such training was not really necessary for the raters, but to ensure the identical circumstances for both instruments, this training was conducted. Since all the teachers were accustomed to grading papers using this holistic-analytic criterion, only a few questions were asked and the overall meeting lasted for only one hour. At the end of the meeting, all the graders were given a file including 50 sample papers, a holistic-analytic criterion, a grading sheet (Appendix E), and finally the topics on which the students were supposed to write. They were told not to interact while grading the papers, and an amount of time (1 week) was given to them to grade and give the papers back.

Phase 3: After collecting the initial grades and the files, the order of the students' papers was changed so as to militate against graders recalling the papers they had already marked. 4 weeks after the first grading, the files including the papers were distributed to the graders again to be graded with the holistic-analytic instrument, just as in the previous grading, in order to have the necessary data for quantifying the intra-rater reliability. The graders were again given 1 week to grade.

Phase 4: After collecting the results of the second grading using the holistic-analytic instrument, necessary statistical reliability tests were implemented on the data collected. Calculations regarding inter/intra rater reliability degrees (Inter-rater: 0.38; intra-rater: 0.74) were done and significantly low degrees were found in terms of inter-rater reliability. To see whether a single component caused the diversity among the scores, a number of factor analysis were implemented on the components of each single score.

Finding all the components erroneous (the diversity in scoring for each component was found as: task achievement 40%, essay organization 47%, accuracy 46%), the necessity of developing a new criterion emerged. Since many researchers (Brown, 1994; Sweedler-Brown, 1993; Bachman, 1991; Turner-Upshur, 2002) are of the same mind on the superiority of analytic grading in terms of reliability in testing the written works, it was decided to develop the new criterion analytically (See Appendix H). First, the previous studies on writing-criteria development were searched and among them the study of Sasaki and Hirose (1999) was taken as the starting point. In this study, a questionnaire was devised and 102 language teachers were asked about the possible descriptor groups of a grading standard. The teachers rated 35 different headings and among them about twelve classifying qualities were chosen (organization, content, language, social awareness, vocabulary, appeal to the reader, spelling, punctuation-capitalization, neatness, handwriting, transitions and title).

These qualities were given to the graders of Anadolu University School of Foreign Languages English Preparatory Program to rate according to the importance, and the most popular 8 components, which have to account for the views of minimum 50% of the participants, were taken as the parts of the new analytic instrument. This development procedure lasted for 2 months, and after the final version of the new criterion was formed, all the graders were asked to attend a meeting again. In the meeting, the new criterion and its significance were explained. To make the graders aware of what would be measured, all the steps in the process of the criterion development were told. Next, the criterion was discussed, and then the graders were given 3 different sample papers (See Appendix G) to grade by means of the new criterion. Questions about the new criterion and its wording style were discussed and the scores graders had given were compared simply with the other graders' scores. The results were considerably consistent. At the end of this meeting, which lasted about two and half hours, each grader was again given a file including the 50 papers, the same essay topics, and the new analytic criterion with its grading sheet (See Appendix H and I). Papers were to be graded within a week.

Phase 5: In the last phase, after collecting the results of the first analytic grading, the same files (in different paper orders) were again given to the graders 4 weeks later, so

that not only the inter-rater reliability of the new criterion could be tested, but also the intra-rater reliability which is also related with the grader him/herself could be found. This time, a new questionnaire was also added to the file in order to take the graders' further suggestions about the criterion to clarify its weak or strong points (See Appendix J). The questionnaire was designed as a Likert Scale so as to ease the evaluation of both the criterion and the questionnaires for the graders. The results of the questionnaire were mostly positive except two points emphasized by two of the ten graders. One of them was related to the detailed structure of the analytic grading system. The grader objected to the complexity of the form of the criterion. According to him, the bands and their descriptors were too detailed and most of the descriptors were rather unnecessary. However, the nature of analytic grading involves a number of details. In analytic grading, the student's written work is not considered as a whole product, but regarded as a body that has certain constituents to be graded separately. The details, therefore, should not be considered as a weak side of the analytic grading. Secondly, a grader mentioned the difficulty of keeping the descriptors in mind, which led to various refer-backs to the criterion while grading. However, frequent refer-backs are, in a way, signs of effective instruction and self-discipline. Kunnan (1995) claims that "the more the grader uses the rating scale, the less the 'human factor' involves in language testing" (p: 48). Thus, various refer-backs should not be considered as a weakness of analytic scoring since such refer-backs enable the grader to check the instruction, description or even him/herself regularly throughout the grading.

Phase 6: Finally, because neither of the mentioned complaints have valid justifications, they were not taken into account when re-shaping the grading criterion. Furthermore, the feedback taken from the rest of the graders was encouraging, since the graders' praise for the evaluative degree of the new criterion undoubtedly outweighed the criticism made regarding its practicality. However, to be sure on the reliability of the assigned scores with each criterion, a third grading session was held after 6 months from the last grading.

The same files, each having 50 papers in a different order than the previous grading sessions, were given to the graders. For the 3rd grading a switch design was implemented,

in other words 5 of the graders were assigned to grade with the holistic-analytic instrument while the other 5 were assigned to use the analytic instrument. Doing this, it was aimed to disable the chance factor for each instrument when they are used either first or next. The graders were again given 1 week to grade. After all the markings were done, the files were collected back and the order of the papers was again changed. After a month, the files were given to the graders switching the groups. This time the group who had graded with the holistic-analytic criterion were given the analytic criterion while the other group were assigned to grade with the holistic-analytic criterion. At the end of one-week-grading time, the files were collected and evaluated. The primary reason in conducting the third gradings was to check whether possible variations among the scores were due to the grading systems or the raters themselves. The answer of this question will be discussed thoroughly at the end of the discussion chapter.

3.5.Data Analysis

After the papers had been graded 6 times, 3 different groups of scores were yielded using holistic-analytic and analytic criteria. To clarify their inter/intra-rater reliability degrees and to compare their effectiveness, a number of different techniques were used. In an attempt to find out the inter-rater consistency degrees of 10 graders for each grading criteria, ANOVA (Analysis of Variance) was used. On the other hand, for clarifying the intra-rater reliability degrees t-test was used, rather than ANOVA. The t-test tells whether the variation between two groups is significant so it is quite suitable for the comparison of both grades of a grader given with each criterion. To measure the inter-rater consistency or to test the significance of the relation between groups of scores, "ANOVA was preferred because it performs comparisons for an arbitrary number of factors (Brown, 1996). In other words, ANOVA cannot be efficient to give the reliability itself, but can be used where there are more than two groups of scores to compare. Traub (1994) states that ANOVA works on a single dependent variable such as students' grades, and can be thought of in a practical sense as an extension of the t-test to an arbitrary number of factors (pp:73-75). Furthermore, Pearson's Correlation Coefficient test was implemented on the scores in order to clarify the strength of the relationship between two groups of scores. Traub (1994) explains a correlation coefficient as a value or a number between -1 and 1 that reveals the

degree to which two different groups of variables are linearly related. Since such a degree of correlation may also show the degree of consistency in grading, it was added to the analysis of data within extensive comparisons. Finally, the variance of grades and relationship of the maximum and minimum grades given to the same paper using both criteria was assessed. This issue will also be examined in the following chapter.

CHAPTER IV

RESULTS and DISCUSSION

4.1. Introduction

This study aimed to develop an alternative grading-criterion to the one that is currently being used at Anadolu University School of Foreign Languages English Preparatory Program. It further aims at finding out if the implementation of a new analytic scoring system would result in a considerable increase in the reliability degree. To achieve this, 10 writing instructors graded 50 papers 6 times using two different grading standards. In the first two gradings, a holistic-analytic grading criterion was used, and the papers were graded twice to have valid data for measuring the intra-rater reliability levels. The findings of these gradings revealed a rather low degree of rater consistency. In the next step, different factor analysis tests were implemented on the components of the holistic-analytic criterion, and the diversity among the scores assigned to separate components led to the necessity of designing a new measurement scale. In the next phase, the same pile of papers was graded by the same graders in the same amount of time, but using a new criterion. The analytic criterion was developed for the use of the English Preparatory Program's writing exams, having in mind the findings in the literature and the suggestions of this school's writing program members. It was designed to measure analytically, since many researchers regard such instruments as more reliable than holistic or holistic-analytic instruments. With the new analytic criterion, the papers were again graded twice and the results were compared with the results of the gradings done with the current grading standards of the English Preparatory Program. Six months later, two final grading sessions were held with both criteria by the same raters to see whether the results gathered after the initial gradings were confirmed in the long term. The following section presents these comparisons and their discussion in view of the necessary statistical calculations.

4.1.1. Explanation of Testing Terms Used

It is commonly accepted in all kinds of sciences that before doing an experiment or an empirical study, certain significance levels are determined and the results gathered from a study are evaluated on the basis of these significance levels. Brown (1996) suggests 10 % or 5 % significance levels for social sciences, whereas these who are closely related with certain laws and formulas mostly prefer 1 % significance level. Thus, 5% significance level is maintained as the basis of all calculations in this study in an attempt to reveal that the results of the study are 95 % reliable. Throughout this chapter, 95 % significance level will be written as 0,05 and the value of "p" (p = significance level) in most tables is represented as "sig.". Likewise, the number of graders or the number of total papers is represented with the letter " n ". F- delivery is a value depends on the degrees of freedom for both the numerator (among-groups) and denominator (within-groups). The possibility is connected with the degrees of freedom in the numerator (the number of groups minus one, for a single-classification ANOVA), df is the degree of freedom in the denominator (total n minus the number of groups, for a single-classification ANOVA) (Brown, 1996; pp: 99-101). Since three different gradings were done for each criterion, they will be classified as "Grading 1 ", "Grading 2" and "Grading 3". Finally, the holistic-analytic criterion, which is mainly the matter of discussion in terms of reliability, is represented as the 1st Instrument whereas the analytic criterion is represented as the 2nd Instrument.

4.2 An overview of the 1st instrument in terms of rater consistency after 2 gradings

Since finding out the reliability levels of the 1st Instrument was the initial goal of this study, it was preferred to discuss it first in the analysis of data. To have a rough view of the amount of consistency achieved by the graders with the 1st Instrument, their mean scores after 2 grading sessions were calculated. According to literature more than 10 % of

discrepancy among the raters is not tolerated provided that the same scoring standard is used (Bachman & Palmer, 1996). Table 4.2.1 reflects the mean scores and the amount of consistency among the graders after two gradings using the 1st Instrument.

Table 4.2.1: The means of the graders' overall scores given with the 1st Instrument

Instrument 1		**n**	**Mean**
Grader	1	100	67.70
Grader	2	100	68.24
Grader	3	100	**67.33**
Grader	4	100	70.22
Grader	5	100	69.62
Grader	6	100	76.02
Grader	7	100	74.82
Grader	8	100	68.76
Grader	9	100	74.40
Grader	10	100	**80.35**

(n = the number of papers after two gradings, n=50x2)

Taking the lowest and highest values into consideration, a noteworthy difference of mean scores, ranging from 67.33 to 80.35, was observed. The difference in mean scores is 13.02 which is obviously an important discrepancy when the term "mean scores" is underlined. The achievement level in testing writing skills is measured between 0 (the least successful)- 100 (the most successful) in the mentioned language school, and the score 70 is considered as the minimum level for the learners to achieve. However, the difference ranging from 67.33 to 80.35 is a great diversity of scores; thus, a number of inconsistent results is supposed to affect the learners' achievement levels because of using the 1st Instrument. For a further analysis One-way ANOVA was used to clarify whether the score differences among 10 raters are statistically significant.

Table 4.2.2: ANOVA of the rater scores with the 1st Instrument

ANOVA (a) Grades					
	Sum of Squares	df	Mean Square	F	**Sig.**
Between Groups	17271,704	9	1919,078	5,450	**,000**
Within Groups	348577,780	990	352,099		
Total	365849,484	999			

As Table 4.2.2 shows, the difference among the graders' scores is statistically significant (F= 5.450, df = 9, p = .001). As it suggests, all the raters acted differently and their assigned scores were dissimilar. Since the difference among the graders' mean scores is significant, we, then, looked at whether this significant difference results from the raters' first and second gradings.

Table 4.2.3: The mean scores of the raters' first and second gradings with the 1st Instrument

INSTRUMENT 1	GRADER		Grading	N	Mean
	1	Grades	1	50	66,86
			2	50	68,54
	2	Grades	1	50	**65,56**
			2	50	70,92
	3	Grades	1	50	67,28
			2	50	67,38
	4	Grades	1	50	66,86
			2	50	73,58
	5	Grades	1	50	74,14
			2	50	**65,10**
	6	Grades	1	50	**78,88**
			2	50	73,16
	7	Grades	1	50	74,50
			2	50	75,14
	8	Grades	1	50	70,54
			2	50	66,98
	9	Grades	1	50	75,96
			2	50	72,84
	10	Grades	1	50	78,18
			2	50	**82,52**

In Table 4.2.3, all the graders' 1st and 2nd gradings' mean scores of the 50 papers with the first instrument, are given. At first, the difference between the maximum and minimum mean scores after the first grading can be emphasized. 78.88 was the maximum mean score in the first grading whereas the minimum mean score was 65.6. The difference between the maximum and minimum mean score is 13.32. A 13.32 points difference between the minimum and maximum mean scores may be accepted as a remarkable inconsistency if the term "mean score" is emphasized. In other words, it must be underlined that the score difference is impossible to tolerate in terms of inter-rater consistency since 10% discrepancy is the maximum level tolerated.

After the second grading, the graders' divergence between the maximum and minimum mean scores with the 1st Instrument grew higher. The maximum mean score was found as 82.52 whereas the minimum mean score remained nearly the same as the first grading, 65.10. A noteworthy difference of 17.42 points lessens the reliability of the 1st Instrument to a great extent since 10% discrepancy is the maximum level tolerated and the 17.42 points score difference is above the permissible level. Moreover, the increase in the maximum mean score attracts attention considering the fact that the minimum mean scores remained nearly the same. Taking all these differences into account, it may be concluded that graders tend to give varied scores with the 1st Instrument in each grading. Next, we applied t-test to see whether there is a statistically significant difference between the 1st and 2nd gradings of each rater, because such a measurement would not only let us know whether there are significant differences between raters' own scorings but also help to identify possible statistically proved intra-rater inconsistencies.

With this aim in mind, the differences in the mean scores between the 1st and 2nd gradings were calculated, and considering all the papers that eac rater graded twice t-test was applied to see if there is a statistically difference between each rating. Table 4.2.4 presents the t-test results of the 1st and 2nd gradings of each rater.

Table 4.2.4: T-test of the difference of mean scores of each grader given with Instrument 1 after 2 gradings

INSTRUMENT 1				
GRADER	t-test for Equality of Means			
	t	df	Sig. (2-tailed)	Mean Difference
1	-0,483	98	0,63	**-1,68**
2	-1,375	98	0,172	**-5,36**
3	-0,029	98	0,977	**-0,1**
4	-1,841	98	0,069	**-6,72**
5	2,172	98	**0,032**	**9,04**
6	1,577	98	0,118	**5,72**
7	-0,165	98	0,869	**-0,64**
8	0,97	98	0,335	**3,56**
9	0,695	98	0,489	**3,12**
10	-1,508	98	0,135	**-4,34**

The differences between the mean scores of the 1st and 2nd grading range from –6.72 to 9.04 when all the raters are considered. This wide range of differences in mean scores show that the raters tend to give inconsistent grades. 1st, 2nd, 4th, 7th and 10th graders give lower scores in the second grading while the other graders give higher scores in the second grading. The t-test results reveal that there is a statistically significant difference between the 1st and 2nd gradings of the 5th rater ($t = 5.450$, $df = 9$, $p = .032$, $p < .05$).

To compare the scores of ten graders and find out their correlation levels with the 1st Instrument, Pearson's r product-moment correlation was used. A correlation coefficient, as was mentioned before, measures the strength of a relationship between two variables (Brown, 1996). Thus, to measure the relationship of a group of scores, one needs another group to compare. The first grading with the first instrument was taken as the basis, and the second grading's results were compared with the former one.

Table 4.2.5 : Correlation of the grades given with the 1st Instrument after two gradings.

Correlations			
Instrument 1		**1st Grading**	**2nd Grading**
1st Grading	Pearson Correlation	**1,000**	**,740(**)**
	N	500	500
2nd Grading	Pearson Correlation	**,740(**)**	1,000
	N	500	500

Table 4.2.5 presents the correlation or the level of consistency between the graders' two scorings and there is no statistically significant difference at the .05 confidence level (r = 0.195, df = 498, p =.740).That is, the first and the second gradings of all raters do not show much difference. Despite the general trend that assumes correlation results as a basic indicator of inter/intra-rater reliability, it must be kept in mind that correlation alone would never be enough to measure the reliability of any instrument as it presents only the degree of relation between the two groups of scores given by graders. To conclude, the correlation results gave us the idea that Instrument 1 may present acceptable intra-rater results at the significance level of .05 when implemented on similar situations.

When the reason for such inconsistencies among the graders and their own gradings is questioned, two important points are underlined, the rater and the instrument. For the issue whether the rater is the source of this inconsistency, literature presents a strong counter argument that a rater or the decision maker in a scoring environment should never be seen as the source of the problem. Brown (1996) states that because of the subjective nature of assessing writing, one should always be given the right for subjective scoring to a certain extent. When the tolerated level is exceeded then the validity and reliability of the method or means of grading should be questioned. Bachman & Palmer (1996) also stress the importance of the tool rather than the grader and suggest frequent controls on

measurement tools in language testing since what is always controllable and revisable is the instrument rather than its user.

Since the literature suggests the analysis of scores rather than the raters, the focus to find the source of problem, shifted to the components of the final scores. To see which component differed most in terms of score distributions among the graders would enable us to revise or rearrange the single component and try it again. However, the results of factor analysis tests revealed that more than a renovation, a complete restructuring or development of a new criterion was necessary.

Table 4.2.6: The differences between the maximum and minimum scores and their ratios for each component

Instrument I	**Grading I**		**Grading II**	
Components	Mean difference	Ratio	Mean difference	Ratio
Accuracy	11,10	56 %	7,44	37 %
Essay Org.	20,46	51 %	17,02	43 %
Task Ach.	17,40	44 %	14,28	36 %

The above table includes the means of maximum-minimum score ranges and their percentages among the scores assigned to each component. In the first instrument, there were 3 components: accuracy (20 points), essay organization (40 points) and task achievement (40 points). Of these three, accuracy had a 56 % inconsistency with an 11.10-point range between the highest and lowest mean scores assigned by all graders after the first grading. In the second grading, the percentage of inconsistency decreased to 37 % and the score range was found 7.44, which is also lower than 11.10. The literature presents no certain range of scores between the highest and the lowest values; however, a 10 %

inconsistency among all the graders is tolerated in the component analysis by many experts (Brown, 1996; Traub, 1994). Of the three components, "accuracy" was supposed to be the one on which all the graders would agree having quite similar focuses on grammar rules, spelling etc. However, even "accuracy", the component which attracted the attention most, was found problematic like the other two. Those components "essay organization" and "task achievement" were also estimated problematic since the inconsistency percentages of each component were more than the tolerable amount (10 %). In addition, Traub (1994) also suggests the idea of calculating the standard deviation for each component. He proposes that a component needs revision if its standard deviation was found more than 1.00.

Table 4.2.7: Standard Deviation values of the 1st Instrument components

	Standard Deviation		
	Components	**Grading I**	**Grading II**
Ins. I	Accuracy	3,61	2,50
	Essay Org.	6,69	5,44
	Task Ach.	5,69	4,79

It is clear from the above table that the standard deviation values of each component were quite above the tolerated level (1.00). Among the three components "essay organization" was calculated as the most problematic whereas the component "accuracy" was calculated as the least problematic one. At this point it must be emphasized that questioning why the standard deviation values of "essay organization" were much more than the values of "accuracy" would be a mistake since the allocated scores for both components were different, so the possible score ranges, for each component would differ as well. One step further, One-way ANOVA was also implemented on the scores given to each component (See Appendix M). As was mentioned before, among the given total scores, there was a significant difference among the raters' total scores when the 1st Instrument was the medium of grading (Grading 1: $F = 3680$, $df = 9$, $p = 001$; Grading 2: F

= 3651, df = 9, p= 001). Moreover, there were significant differences among the scores given to each component. What is more, the same findings were held again when the second grading with the 1st Instrument was analyzed. Therefore, it can be claimed that raters assign inconsistent scores to the components of the 1st Instrument leading to a final discrepancy among the total scores after both gradings.

Grading 1:

- Task Ach. F = 2,366, df = 9, p=, 01
- Essay Org. F = 3,699, df = 9, p=, 001
- Accuracy F = 1,779, df = 9, p=, 01

Grading 2:

- Task Ach. F = 2,588, df = 9, p=, 006
- Essay Org. F = 4,158, df = 9, p=, 001
- Accuracy F = 1,988, df = 9, p=, 03

To conclude, the inter-rater consistency with the 1st Instrument was found rather low, and a significant difference was found among the graders' scores using the 1st Instrument. On the other hand, intra-rater consistency of the graders was at an acceptable level (.74). This is, in fact, not really crucial for the institute since gradings on a certain paper are done once by the same rater. Thus, the focus while doing the statistical calculations was on the inter-rater consistency rather then the intra-rater reliability levels. In an attempt to clarify the reason of inconsistency among the raters, a number of factor analysis were implemented on the scores and it was found that developing a new scale would be better since all of the components were erroneous quite above the tolerable level. Therefore, it was inevitable to design a new instrument and test its reliability to see whether it would provide higher reliability levels than 1st Instrument does.

4.3 An overview of the 2nd Instrument in terms of rater-consistency after 2 gradings

The design of the 2nd Instrument was figured in a way that all the raters in this study expressed their ideas and agreed on the final product. Literature agrees on the idea that raters behave consistently if they form a consensus on a grading scale, and they measure successfully if they believe in the criteria. Thus, it is clear that in the development phase a cooperative study for every single component of a grading scale may result in higher consistency levels. In order to test this issue and to see whether a better inter-rater reliability level is achieved when an analytic scale is used, the results gathered after 2 scoring sessions were analyzed in the way that 1st Instrument was analyzed.

Table 4.3.1: The means of the graders' overall scores given with the 2nd Instrument

Instrument 2		n	Mean
Grader	1	100	61.44
Grader	2	100	**60.27**
Grader	3	100	62.60
Grader	4	100	62.31
Grader	5	100	60.43
Grader	6	100	**65.60**
Grader	7	100	64.23
Grader	8	100	64.65
Grader	9	100	63.03
Grader	10	100	63.97

(n = the number of papers after two gradings n=50x2)

As shown in Table 4.3.1, the lowest mean score is 60.27 and the highest mean score is 65.60. Thus, the mean scores range between 60 and 65 indicate that there is a consistency among the raters after both gradings with the 2nd Instrument. A 5.33 points

difference in the mean scores underlines a considerable rater agreement on the papers, and the experts tolerate it since it does not exceed 10 % level of tolerance. Brown (1996) states that it would be more sensible to strive for a lower rater-discrepancy in subjective scoring than to search for a "perfect" rater-discrepancy, since a completely reliable and consistent grading system has not been designed yet. Therefore, comparing the 5.33 points difference of the 2nd Instrument with the 13.02 points-discrepancy of the 1st Instrument, it may be claimed that the difference of mean scores decreased considerably when the 2nd Instrument is chosen as the rating scale (See also Appendix K-L).

For a more detailed analysis on the mean scores given with the 2nd Instrument, the similarity among the raters' scores was checked using ANOVA.

Table 4.3.2 : ANOVA of the rater scores with the 2nd Instrument

ANOVA (a) Grades INSTRUMENT 2					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2884,861	9	320,540	,812	,605
Within Groups	390954,530	990	394,904		
Total	393839,391	999			

(p>0.05)

As Table 4.3.2 shows, with the second instrument, there is no significant difference among the raters' mean scores (F= .812, df = 9, p = .605). In other words, if the mean scores of the graders are compared with each other, it can be stated that the raters tend to measure similarly with the second instrument.

Table 4.3.3: The mean scores of the graders' two different grades by the 2nd Instrument.

Instrument 2				
Grader		**Grading order**	**N**	**Mean**
1	Grades	1	50	61,16
		2	50	61,72
2	Grades	1	50	60,64
		2	50	**59,90**
3	Grades	1	50	62,34
		2	50	62,86
4	Grades	1	50	62,72
		2	50	61,90
5	Grades	1	50	**60,18**
		2	50	60,68
6	Grades	1	50	**66,14**
		2	50	**65,06**
7	Grades	1	50	64,26
		2	50	64,20
8	Grades	1	50	64,64
		2	50	64,66
9	Grades	1	50	62,74
		2	50	63,32
10	Grades	1	50	63,72
		2	50	64,22

Table 4.3.3 shows the mean scores of the 1st and 2nd gradings assigned by each grader with the 2nd Instrument. As it is seen, the maximum mean score was 66.14 whereas the minimum mean score was 60.18 after the first grading. A difference of 5.96 is encouraging, and it may be considered as an acceptable range if the 10 % tolerance interval of literature is recalled. What is more encouraging is that the 5.96 score difference reduces to 5.16 in the 2nd grading. This may be indicative of the fact that a certain level of progress was held in terms of rater-consistency with the use of the 2nd Instrument. Thus, it can be claimed that inter-rater consistency between the graders of writing exams may be higher if the 2nd Instrument is used as the scoring standard.

In the next step, intra-rater consistency level of the 2nd Instrument was analyzed. T-test was applied to see whether there is a significant difference between the two gradings of each rater.

Table 4.3.4: T-test of the difference of mean scores of each grader given with Instrument 1 after 2 gradings

INSTRUMENT 2				
	t-test for Equality of Means			
GRADER	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**
1	-0,142	98	0,887	**-0,56**
2	0,187	98	0,852	**0,74**
3	-0,128	98	0,898	**-0,52**
4	0,225	98	0,823	**0,82**
5	-0,133	98	0,895	**-0,5**
6	0,248	98	0,805	**1,08**
7	0,016	98	0,988	**0,06**
8	-0,006	98	0,996	**-0,02**
9	-0,129	98	0,897	**-0,58**
10	-0,12	98	0,905	**-0,5**

(p>.05)

The mean-score difference of each rater after two gradings with the 2nd instrument revealed that they are quite consistent in their own gradings. Only one of them had a difference more than 1 point while the others were all below 1-point difference. When the significance is considered, it can be said that there are no statistically significant differences between the graders' 1st and 2nd gradings with the 2nd Instrument.

Pearson's r product-moment correlation was implemented on the scores of each rater to see the strength of the relation when the 2nd Instrument is the means of

measurement. The first gradings were again taken as the basis and the results of the 2nd gradings were compared with these.

Table 4.3.5: Correlation of the grades given with the 2nd instrument after two gradings.

Correlations			
		Instrument 2 1st Grading	**Instrument 2 2nd Grading**
Instrument 2 1st Grading	Pearson Correlation	1,000	,975(**)
	N	500	500
Instrument 2 2nd Grading	Pearson Correlation	,975(**)	1,000
	N	500	500

(N = the number of total papers graded by ten graders 50x10=500, correlation coefficients=two different groups are compared to clarify whether there is a relationship between each other; thus, of the two groups, one of them is taken as the basis (with a value of 1.000) each time.)

Table 4.3.5 presents the result that there is no statistically significant difference between the scores given with the second instrument (r = 0.195, df = 498, p = .975). That means, once again, the raters give more or less similar grades in the first and second gradings.

As a last step, Factor analysis was implemented on the components of this criterion. The reason to do this was the need to be sure that the consistency achieved in totals also existing in the pieces, because the nature of analytic grading requires the consistency in the components to reach the total reliability.

Table 4.3.6 The differences between the maximum and minimum scores and their ratios for each component

Instrument II	Grading I		Grading II	
Components	Mean difference	Ratio	Mean difference	Ratio
Organization	2,07	8,28 %	1,93	7,72 %
Content	1,91	7,64 %	1,6	6,38 %
Language	0,91	6,37 %	0,87	6,09 %
Vocabulary	0,69	6,9 %	0,87	8,7 %
Transitions	0,76	7,6 %	0,73	7,3 %
Title	0,0	0 %	0,0	0 %
Punctuation-Capitalization	0,42	8,4 %	0,39	7,8 %
Spelling	0,38	7,6 %	0,33	7.2 %

Table 4.3.6 gives the amount of differences among the raters and their ratio in the sum using the 2nd Instrument As was mentioned before, literature allows maximum 10% difference among the raters in a single component. Having this tolerance in mind, it can be said that none of the components of Instrument 2 is above the tolerated level. Among the components of the second instrument, the discrepancy among the raters occurred most while scoring the punctuation and organization in the first grading while it was the measurement of vocabulary on which the raters agreed less in the second grading. Moreover, the calculation of the standard deviation of the scores assigned to each component of the analytic scale confirmed the idea that each component of the 2nd Instrument supplements the desired level of reliability (See Table 4.3.7).

Table 4.3.7: Standard Deviation values of the components of the 2nd Instrument

Standard Deviation			
	Components	**Grading I**	**Grading II**
Ins. II	**Organization**	0,86	0,735
	Content	0,719	0,61
	Language	0,807	0,776
	Vocabulary	0,713	0,912
	Transitions	0,76	0,701
	Title	0,0	0,0
	Punctuation-Capitalization	0,88	0,821
	Spelling	0,702	0,897

Remembering the fact that literature tolerates maximum 1.00 standard deviation degree, the components of the 2nd Instrument are found to be below the maximum level. Thus, it may be claimed that all the components of the 2nd Instrument meet the standards of statistics and related literature, which is at the same time indicative of the fact that the 2nd Instrument is found reliable not only with the total scores it provides but also with the components it includes.

4.4 Long term grading results of both instruments

The previous 2 sections discuss the results gathered by each instrument used one after another. Most of the findings revealed the superiority of the second instrument and supported the idea that Instrument 1 may not be a reliable tool for grading writing when

used under similar conditions. It was also highlighted that the inconsistency calculated for the 1st Instrument was mainly caused by its components, while 2nd Instrument's components were found to enable the raters assign consistent scores. In attempt to prove that all those findings were not gathered by chance a final grading session was held with each criterion once more.

Table 4.4.1: Mean scores assigned with two instruments after three gradings

Grading Order	Instrument	N	Mean
1	1	500	71,88
	2	500	62,85
2	1	500	71,62
	2	500	62,85
3	1	500	71,48
	2	500	63,11

Table 4.4.1 includes the mean scores of ten graders with each criterion after different grading sessions. The similarity of the mean scores after each grading with the same rating scale may suggest the idea that raters' attitudes in sum remained the same. When all the scores were matched as the ones given with the 1st and the ones given with the 2nd, the similarity can be seen more vividly. No matter which instrument is used, it may be said that the group of raters did not score differently after six months. This may give us the idea that in this six months period the raters remained the same and reflected their performance in just the same way as they did 6 months ago. For a better comparison, the mean scores of the 1st and 2nd gradings were compared with the means of 3rd grading.

Table 4.4.2: Comparison of the mean scores of the 1st and 2nd gradings with the scores of 3rd grading

Group Statistics			
Instrument	**Grading**	**N**	**Mean Scores**
1	1st – 2nd	1000	71,75
	3rd	500	71,48
2	1st – 2nd	1000	62,85
	3rd	500	63,11

(n=1000 represents 1st & 2nd ratings including 50 papers graded by ten raters, n=500 represents the last rating including 50 papers graded by ten raters

It is seen that the graders did not score differently after such a long time. This finding indicates that the source of the problem in gradings is not the human since the mean scores remained nearly the same after a long time. Had the raters been the source of the problem, huge diversities in the last scoring would have been expected after the 3rd grading. The t-test results show that there is no statistically significant difference among the three scoring sessions (See Table 4.4.3). In other words the graders acted nearly the same with their previous gradings since the three columns are almost identical.

Table 4.4.3 : T-test of the comparison of the scores of the 1st and 2nd gradings with the scores of 3rd grading

T-test for Equality of Means				
Instrument	**n**	**t**	**df**	**Sig. (2-tailed)**
I	1500	,255	1498	,799 *
II	1500	-,237	1498	,813 **

(* represents the comparison of the scores of the 1st and 2nd gradings with the scores of 3rd grading with the 1st Instrument, ** represents the comparison of the scores of the 1st and 2nd gradings with the scores of 3rd grading with the 2nd Instrument)

As Table 4.4.3 shows, there are no significant differences between the scores of the 1st and 2nd gradings with the scores of the 3rd grading (Ins.I: t= .255, df= 1498, p= .799; Ins.II: t= -237, df= 1498, p= .813). Thus, it is statistically proved that raters' scores in the long-term remained almost similar in each scoring with both criteria. For the detailed analysis of the scores of each grader, all the scores given with the 1st instrument were compared in the following table.

Table 4.4.4: Comparison of the mean scores of ten graders with the 1st Instrument after three gradings

Instrument 1	Grading order					
	1		2		3	
Grader	Count	Mean	Count	Mean	Count	Mean
1	50	66,86	50	68,54	50	65,38
2	50	65,56	50	70,92	50	69,06
3	50	67,28	50	67,38	50	68,72
4	50	66,86	50	73,58	50	71,42
5	50	74,14	50	65,10	50	66,24
6	50	78,88	50	73,16	50	72,72
7	50	74,50	50	75,14	50	77,36
8	50	70,54	50	66,98	50	70,02
9	50	75,96	50	72,84	50	74,06
10	50	78,18	50	82,52	50	79,82

When the lowest and the highest mean scores are compared in each grading, it can be said that each time the amount of difference remained more than 10 points, which is higher than the acceptable level. In the first grading, it is 13.32, which is surprisingly lower than the 2nd (17.42), and the 3rd grading (14.44). However, it was claimed that the more a scoring guide is used, the better consistency is achieved among the raters. Considering the fact that the lowest mean scores were nearly the same after each grading, it may be concluded that graders are affected by the criterion in a way that they are more tolerable to score higher points if the possible score of a paper is well above the deadline-70. To see

whether the differences among the raters with the 1st Instrument were significant, a one-way ANOVA was implemented.

Table 4.4.5: ANOVA of the rater scores assigned in the 1st, 2nd, and the 3rd gradings with the 1st Instrument

ANOVA (Grades)					
Instrument 1	Sum of Squares	df	Mean Square	F	**Sig.**
Between Groups	25586,096	9	2842,900	8,189	**,000**
Within Groups	517291,773	1490	347,176		
Total	542877,869	1499			

($p<0.05$)

It is clear from the table that there is a statistically significant difference among the mean scores of the raters in each grading ($F= 8.189$, $df= 9$, $p= .001$). In other words, it can be said that all the raters tend to score differently when the scoring standard is the 1st Instrument. On the other hand, when the mean scores of each grader were controlled a much better intra-rater reliability is found with the 2nd Instrument.

Table 4.4.6: Comparison of the mean scores of ten graders with the 2nd Instrument after three gradings

Instrument 2	**Grading order**					
	1		**2**		**3**	
Grader	**Count**	**Mean**	**Count**	**Mean**	**Count**	**Mean**
1	50	61,16	50	61,72	50	62,54
2	50	60,64	50	59,90	50	62,06
3	50	62,34	50	62,86	50	63,54
4	50	62,72	50	61,90	50	63,16
5	50	60,18	50	60,68	50	61,38
6	50	66,14	50	65,06	50	64,42
7	50	64,26	50	64,20	50	64,90
8	50	64,64	50	64,66	50	62,92
9	50	62,74	50	63,32	50	62,04
10	50	63,72	50	64,22	50	64,12

The mean score differences after each grading with the 2nd Instrument show that a better inter-rater consistency was achieved with the 2nd Instrument. The score differences between the highest and the lowest mean scores for each grading are: 5.96 for the 1st 5.16 for the 2nd and 3.52 points for the 3rd grading.

Table 4.4.7: ANOVA of the rater scores assigned in the 1st, 2nd, and the 3rd gradings with the 2nd Instrument

ANOVA (Grades)					
Instrument 2	Sum of Squares	df	Mean Square	F	**Sig.**
Between Groups	3067,667	9	340,852	,883	**,540**
Within Groups	575033,567	1490	385,929		
Total	578101,234	1499			

(p>.05)

As Table 4.4.7 reveals, there is no statistically significant difference among the mean scores of the raters after three gradings (F= .883, df=9, p=.540). Thus, we can conclude that when the 2nd Instrument is used in assessing students writing performance not only in the short but also in the long run, it may give quite consistent results among the raters. As a last step, we tried to find out whether the results of three gradings with the 2nd Instrument showed any correlation.

Table 4.4.8: Correlation of three groups of scores assigned with the 1st Instrument

Correlations				
		Ins. I 1st Grading	Ins. I 2nd Grading	Ins. I. 3rd Grading
Ins. I 1st Grading	Pearson Correlation	**1**	**,740(**)**	**,727(**)**
	N	500	500	500
Ins. I 2nd Grading	Pearson Correlation	**,740(**)**	1	,707(**)
	N	**500**	500	500
Ins. I 3rd Grading	Pearson Correlation	**,727(**)**	**,707(**)**	**1**
	N	**500**	500	500

It can be concluded from the table that certain intra-rater consistency at an acceptable level (.7) is achieved when the 1st Instrument was used (r = 0.195, df = 498, p = .727) . All the correlations reveal that the intra-rater reliability of the 1st Instrument can be stated as a value between .7 and .74, which is more than the suggested level. (.7) by many experts (Brown, 1996; Traub 1994). On the other hand, the correlation results of the 2nd Instrument present much better results than the 1st Instrument.

Table 4.4.9: Correlation of three groups of scores assigned with the 2nd Instrument

Correlations				
		Ins. II 1st Grading	Ins. II 2nd Grading	Ins. II. 3rd Grading
Ins. II 1.st Grading	Pearson Correlation	1	,975(**)	,970(**)
	N	500	500	500
Ins. II 2nd Grading	Pearson Correlation	,975(**)	1	,975(**)
	N	500	500	500
Ins. II. 3 rd Grading	Pearson Correlation	,970(**)	,975(**)	1
	N	500	500	500

The level of similarity between the scores given with the 2nd Instrument was calculated in the above table. It is clear that there is a very strong correlation among the 1st, 2nd and 3rd gradings by the same raters with the 2nd instrument (r = 0.195, df = 498, p = .970). Thus, it may be summarized that the intra-rater reliability of the 2nd Instrument may be significantly high also in the long term grading. When the expected level is remembered (.7) for the intended reliability, a value around .975 is really thriving even in the long run.

To conclude, the findings of the 3rd grading with each criterion proved the fact that the findings of the 1st and 2nd gradings were not coincidental. During those three grading sessions, all the raters performed exactly the same since the thing which was most important in doing so was the application of the same instruments and methods rather than the amount of time passed between the 2nd and 3rd gradings. It would be possible to see different scores and calculate different reliability levels if the source of the consistency were the rater; however having a strong correlation among the scores with both instruments, it may be claimed that the results of the 3rd grading (long term grading)

confirmed the fact that the raters were not the source of variation since they were consistent with their own scoring each time.

4.5 The most problematic papers after all gradings

When a subjective scoring is the matter of concern, a high consistency level is generally not expected. The range of scores among the graders widens if the number of total graders increases. Thus, an amount of difference among the top and down grades was expected. What is more important to emphasize here is not whether there is a difference or not, the major concern here is to see by which instrument the graders performed between the minimum intervals.

Table 4.5.1 : Maximum and minimum score differences assigned with two different instruments in three different gradings

Grading order	Value	Instrument I	Instrument II
Grading I	Min	10	8
	Max	60	30
Grading II	Min	8	5
	Max	60	25
Grading III	Min	14	5
	Max	50	15

The above table shows how the lowest and the highest scores differ among ten raters with both instruments. As Table 4.5.1 shows, with the first instrument even after the 3rd grading the maximum score differences among raters remained almost the same while the minimum difference among the scores increased. On the other hand, Instrument 2

reveals a visible positive change in the level of discrepancy especially after the 3rd grading. A maximum 15 points difference was promising since in subjective scorings to achieve such a consistency among ten raters may be considered an accomplishment. The second instrument was a new scale ,and that 15 points discrepancy may be reduced after a number of further training sessions. Hence, it may be claimed that when compared with the 1st instrument, 2nd instrument provides a greater consistency in the assessment of writing performance.

Up to this section, the mean scores and their variances were compared considering all the papers of the writing test in the same category. Thus the question of whether the calculated levels of rater-reliability remained the same on the most problematic papers, which are quite distinctive in such studies, was still unanswered. Moreover the reaction of the raters to those papers was another important factor that may affect the reliability of a certain instrument to a great extent. With this aim in mind, among the 50 papers used throughout this study the most problematic ones were tried to be identified. To do this, two groups of papers were identified. In the first group, the papers which were graded most inconsistently were classified. Papers 3,5 and 14 were the most problematic papers after 3 gradings done with the 1st Instrument while papers 15.18 and 50 were identified as the most problematic ones with the second instrument. After finding those 6 papers, another concern provided by the literature was put into practice. Experts believe that the difficulty of having reliable results in the papers, which have average success levels or which are quite close to the pass-line (a little below or above), is much more than that of really successful or extremely unsuccessful papers.

Thus, the second group of problematic papers was chosen among the papers which have average success levels. To do this, the papers which were graded between the scores 65 and 75 (the average level of success in the mentioned school) with the first instrument were identified and the matching ones in all three gradings were found. 11 papers including the 2nd, 4th, 8th, 13th, 16th, 17th, 22nd, 25th, 27th, 30th and 31st papers were highlighted. Next, the same thing was applied on the papers graded with the 2nd Instrument ,and papers 16,19,23,24,26,27,30 and 37 were found out. Of al these papers scored between 65 and 75 with each instrument, the matching ones were taken. Paper 16, 27 and

30 were identified as the most difficult papers to grade for both instruments since they were all scored in the danger zone (65-75). Finally 9 papers in sum: 3,5,14,15,16,18,27,30 and 50 were identified as the most problematic ones to grade with both instruments. In the next step, the score ranges between the assigned highest and lowest scores were calculated and compared (See Appendix N).

Table 4.5.2: Most problematic papers for both instruments after 3 gradings

Inst.	Grading order	Paper								
		3	**5**	**14**	**15**	**16**	**18**	**27**	**30**	**50**
I	1	60 (85-25)	40 (65-25)	30 (55-25)	35 (70-35)	36 (91-55)	24 (94-70)	26 (86-60)	27 (85-58)	20 (100-80)
	2	30 (55-25)	60 (85-25)	40 (65-25)	35 (70-35)	36 (91-55)	31 (91-60)	25 (85-60)	28 (83-55)	29 (100-71)
	3	45 (70-25)	40 (65-25)	50 (75-25)	30 (65-35)	40 (95-55)	30 (95-65)	23 (83-60)	35 (85-50)	27 (97-70)
II	1	11 (35-24)	15 (39-24)	12 (37-25)	30 (67-37)	17 (80-63)	25 (76-51)	13 (78-65)	18 (82-64)	12 (92-80)
	2	12 (31-19)	15 (40-25)	12 (35-23)	16 (66-50)	19 (78-59)	25 (77-52)	11 (78-67)	15 (80-65)	15 (95-80)
	3	8 (30-22)	11 (39-28)	11 (37-26)	14 (62-48)	8 (75-67)	15 (73-58)	10 (79-69)	10 (78-68)	5 (91-86)

(Assigned maximum score-minimum score to the paper)

Table 4.5.2 shows the maximum and minimum scores assigned to most problematic papers with both instruments. A general calculation of the score ranges may support the idea that Instrument 2 may provide better results in the most problematic papers comparing with the ones the 1st Instrument provides. Furthermore, when the most striking discrepancies were underlined, it could be seen that most of the score ranges were more than 30 points with the 1st Instrument while only in 1 paper (paper 15) such a score range was seen when the 2nd Instrument was used. In addition, some gradings with the 2nd Instrument were still under the tolerated level (10 % discrepancy); however, with the 1st

Instrument, none of the mentioned papers were graded within the suggested level of consistency. A final analysis was done regarding the reactions of the raters against those problematic papers. To do this with those papers the lowest and highest raters were identified in each grading session considering both criteria (See Appendix N). There were 9 problematic papers and those papers were graded 3 times with both criteria. Cumming (1997) recommends the analysis of the reaction of graders calculating the total number of groups in which the raters acted differently. In problematic papers, the lowest and highest scores were recommended to be taken. Thus, a total of 108 groups of scores (9 papers x 3 gradings x 2 criteria x 2 distinctive figures, the highest and the lowest scores) were held. On those 108 groups of scores, the frequency of each grader being the lowest or the highest rater was examined. The main purpose in doing so was to find out whether any of these raters grades significantly different from the others. The measure the literature suggests is 10 % under normal conditions; however, Traub (1994) tolerates a 20 % (twice more) if the most problematic conditions are studied. Cumming (1997), on the other hand, tolerates a 30 % variation among the graders which would obviously be a more optimistic estimation in such calculations. 20 % variation therefore, was taken as the maximum limit for rater and the frequency of each grader was estimated.

Table 4.5.3: The number of times each grader gave the lowest or the highest scores to the most problematic papers

Grader Frequency				
Grader	Lowest score	Highest score	Total	Ratio
1	11	5	16	**14.8%**
2	7	1	8	**7.4%**
3	6	1	7	**6.48%**
4	6	5	11	**10.1%**
5	11	1	12	**11.11%**
6	2	14	16	**14.8%**
7	0	9	9	**8.33%**
8	3	7	10	**9.25%**
9	6	3	9	**8.33%**
10	2	8	10	**9.25%**

It is obvious from the above table that none of the raters differed more than 20 %, to an extent which can be tolerated according to the findings from the literature. Considering the fact that some raters graded considerably different from the others may lead to the question whether such a difference causes great discrepancies on the total scores. First of all, the difference evident among the raters' overall performance is something expected since the issue was subjective scoring and also it is not over the maximum limit that literature permits. Secondly, it must be underlined that the estimated levels of rater variation are found on extra-ordinary samples and would not be regarded as the final figures of inter-rater reliability. Finally, the scores assigned by grader 1 and 6 in this group, could not be assumed invalid since the percentages reflect the amount of similarity of the assigned scores not the level of true grading. Similarly, the scores by the 3rd grader could not be considered as the best since the table represents just the frequency of most distinguished raters not the one who graded most appropriately.

To conclude, from the results driven from table 4.5.2, grader 3 would never be called the most, grader 1 or 6 would never be called the least consistent graders since literature avoids identifying the raters reliable or not. The possible assumption from that table can be summarized as each of these graders carries a certain amount of variation in scoring and none of those acted significantly different from the other 9 graders regarding the limits literature suggests.

4.6 Discussion

For language testing studies, especially for these which need subjective grading, there are many factors that should be taken into consideration before discussing the issue of reliability. In subjective grading, there are a number of factors like the performance of the learner, the reactions and beliefs of the grader, or the content of the criterion that affect the reliability of the evaluation of a certain test. However, from the researcher's point of view, this study aimed to focus solely on the grading criterion, which should easily be adapted to the intended targets with extensive design and training, instead of resorting to the graders' scoring habits.

In this sense, analytic scoring seems more advantageous since it classifies the expected language mastery into parts and does not allow the grader to measure the student's performance as a whole. Thus, an analytic criterion was designed in order to keep the scorers away from their scoring habits, which seem to be a problem in the holistic-analytic criterion since the goal, to test the final product by looking into pieces, is something extremely hard where subjective scorings take place

To clarify the contradiction between the gradings done with both kinds of techniques, 50 papers were given to 10 graders to score. The papers were graded three times using each grading technique and the results were tested in terms of inter / intra-reliability levels. Before discussing the results of the findings, there are a number of crucial points related with the issue of reliability. First of all, Brown (1996: pp: 192-194) defines reliability as "the consistency of a measurement", and he proposes that reliability itself is not measured but it is estimated. Thus, there is no certain way of measuring reliability, except some statistical tests whose results may be used only to estimate the level of reliability.

Moreover, it must be emphasized that reliability describes final grades, not the graders, subjects or the participants. Thus, a grading criterion cannot be classified as totally reliable or not since the graders and the ones graded will always remain changeable. On the other hand, language-teaching programs need to make their progress clear as to whether they are on the right track by testing their learners' achievement by means of tests. To have sound tests, institutes need certain instruments proved to be reliable to a certain degree in measuring certain abilities in different times. Therefore, the reliability of a grading instrument is hoped to describe the consistency that the instrument ensures in measuring the same performance over time (intra-rater) or by different graders (inter-rater reliability).

Since the initial concern of this study was to determine the inter/intra-rater reliability degrees of the holistic-analytic criterion and to check whether the use of the new analytic scale increases the inter/intra-rater consistency degrees of the criteria used for the evaluation of essays at Anadolu University School of Foreign Languages, the following

section will include the detailed discussion related with the findings driven from different tests to compare the inter and intra-rater reliability degrees.

4.6.1. Inter-rater reliability of both instruments

For a language school, which has hundreds of writing papers that must be graded consistently, the reliability among its raters, which is so called inter-rater reliability, is crucial. To achieve a high consistency among the raters, institutes unfortunately do not have many ways except those, which are still the subject of debate among many language-testing researchers. Three basic issues: grader, grading standards and training sessions attract the attention of researchers looking for a valid way of acquiring consistency. Of these three, 'graders' seem to be the most difficult ones to study, because when the "human factor" is involved in the grading process, no efficient and completely significant measurement can be made.

Next, the training sessions are recommended to be focused, as they are the only arenas where the graders challenge with their ideas and theories very special to them. If a careful observation is carried out during these sessions, one may have valuable ideas to identify the problem in a grading process that affects the level of reliability. What comes later? Given that a researcher succeeded to find the answers of the questions related to reliability, the next problem would be about controlling the graders' scoring processes to keep them away from the troublesome issues that were already determined in former grading sessions. Thus, the last item, grading standards or the grading criterion, seems like the only way that any researcher can hope to have systematic control over the grading process.

Holistic-based criteria have a considerable disadvantage against analytic ones as they attract the attention mainly on the whole while ignoring the consistency through components. Since one can never understand what is praised or punished in a certain exam paper, the grader has the freedom of using his/her own language assessment techniques rather than the established standards, ignoring the fact that when the same paper is graded

by another one according to the common standards, shocking discrepancies are possible to emerge. In the light of these examples, when the inter-rater reliability is the matter of concern, analytic-based criteria look more advantageous than the holistic ones in theory; that is why practice, the other side of the coin, was questioned in this study through empirical research.

To achieve the highest degree of consistency among the graders with different instruments, the sample papers were graded three times to check whether the relationships among the graders were by chance. Furthermore, on 6 different gradings, statistical measurements were implemented in order to have valid results that would support each other while testing the inter-rater reliability. For each criterion, firstly, the mean scores of graders after two gradings were calculated and ranges between the highest and lowest means were pointed out (See Table 4.2.1 and 4.3.1). Secondly, differences among the maximum and minimum scores given to papers after each grading with different instruments were checked and the high rater-consistency was again observed in the results of gradings done with the second instrument (See Table 4.3.2). The next distinctive rater-reliability difference between the two instruments was statistically proved by ANOVA tests.

The significance level for this study was determined as .05 and when the significance levels in both ANOVA tests for each grading instrument are checked, the results again resemble the finding that the graders have significant scoring correlations using the second instrument, whereas the relationship in grading among the graders is not significant when they use the first one .

In addition, the correlation coefficients of both instruments gave the most vivid picture of the superiority of Instrument 2 in terms of rater correlation. Since the strength of a relationship between two groups is emphasized in correlation calculations, the question which grading criterion has better inter-rater reliability degrees is best answered by means of this test. The results revealed that both instruments have a significant degree of rater-correlation, in other words the grades, which all the graders assigned, have strong relationships with the others given to the same papers, but at different times. However,

once the correlation levels were compared, the clear difference between the instruments would be seen. Instrument 2 presents a very high correlation level, which is really encouraging and far more powerful than the first instrument. In fact, Instrument 1's correlation significance was also maintained and its significance degree was found as fairly successful, nevertheless, if the main concern was to find the better means of grading, Instrument 2 was certainly superior to Instrument 1 considering the inter-rater reliability levels.

Finally, the results of the 3rd (long term) grading confirmed the reliability levels calculated through the initial gradings. The fact that nearly the same scores were assigned to 50 papers after 6 months was satisfying since an important degree of consistency was achieved among the results of 3 separate grading. Thus, the long-term results proved the issue that the findings driven from the 1st and 2nd gradings were not found by chance.

4.6.2. Intra-rater reliability of both instruments

As was mentioned before, reliability was defined as the consistency in a measurement approving the fact that when the same measurement is done, the same or very similar results will be repeated (Brown, 1996). It is for sure that in subjective grading sessions, to guarantee the same results from a number of graders after a certain amount of time for the same papers is something really difficult. In language tests such as tests of writing, having similar or consistent results for the same students' performance after some time is quite crucial. Here, the term 'similar' must be emphasized strictly; in other words, the degree of similarity between two scores is indeed very closely related to the level of intra-rater reliability.

What will be the learner's reaction if a grader gives a 35 to a paper and changes his own score into 75 after a certain time in his second grading? What is more, if a student was determined to fail or pass accidentally, just because of his teacher's misjudgments on the student's performance, who will have the responsibility is unknown. Thus, a rater's own reliability or consistency is also important. The way to achieve such consistency may be

the use of the right grading standards, which have proved to supply rater-reliability among its bands. However, it must never be forgotten that there is no totally reliable criterion in the world. Each criterion in printed form or on the internet has a certain level of rater-reliability, but how much tolerance the criterion can bear for intra-rater discrepancies for a certain number of scorings still remains unknown, which is certainly an important factor to consider if the criterion is planned to be used in an extensive language program.

Having an aim of determining each instrument's intra-rater reliability degrees, certain tests were also implemented to clarify which grading standard provides better values of intra-rater consistency. In order to do it, first the t-test of each grader was calculated and the results were compared in different tables (See Tables 4.2.4,4.3.4). After the significance levels of each criterion were found out, it was understood that Instrument 2 revealed better results also in intra-rater consistency levels. Taking the significance levels as the basis for comparison, a considerable difference can be observed between the consistency values that each criterion took. Even those values of significance may be used as the value of each rater's reliability, however, it is clear that the results of t-test would never be enough to a reliability study aimed to find precise answers on the issue of intra-rater reliability of 2 different instruments. Therefore, an alternative test was designed special to that institute's grading system that would let the researcher have different scopes of reliability levels from a totally different confidence interval. First, the terms of failure or success in that school were examined. It was reported that to pass the preparatory class, a student must have at least 70 points from the final exams and should have at least a mean of 70 points when the student's other exams are involved in the evaluation process. In other words, a student who takes 69 as the mean score fails whereas a student who takes 70 as the mean of all scores passes. When the criterion's main usage is remembered, its importance in student's failure or success may be understood better. Both instruments were designed to grade students' writing papers, and writing grades have a weight of 30% in determining the student's grade. Thus, a certain amount of misgrading done with one of those criteria may cause the student's failure or success unexpectedly.

Taking the value 70 as the basis, a student's overall mean grade should never be decreased or increased with +1 or −1 grade; otherwise he/she may fail or succeed by

mistake. In this sense, if a writing grader gives 7 points more/less to a student's writing paper than the grader may cause the student's mean change 1 point up or down. The average was determined as 7 because after the final papers grading the 50% of those scores are taken into account. It means half of those 7 points will be effective and it makes 3.5 points. Next the weight of writing ability in student's success is 30% and makes 1.16 points of those 7 points. Eventually since that 1.16-point difference may cause the student fail or pass, 7 points of difference in the final writing paper was regarded as crucial. In the next step, the graders' all gradings by two different instruments were examined and 7 points difference was searched between the grades given for the same paper. If the difference between the grades was 7 or less than 7, the relation was coded as "1". If the difference between two grades for the same paper was more than 7 it was coded as "0" (for the table, see Appendix O). Later the number of the values 1 and 0 were calculated among the grades given by each instrument, and the mean scores were taken as the new intra-rater reliability measurements for both instruments (See Table 4.6.1).

Table 4.6.1: Primary level of reliability (reliability 7) comparing the two instruments

Grader	INSTRUMENT 1	INSTRUMENT 2
1	0,74	1.00
2	0,42	0,96
3	0,36	0,96
4	0,36	0,88
5	0,26	0,90
6	0,28	0,84
7	0,34	0,98
8	0,34	0,96
9	0,40	0,86
10	0,26	0,86
Final (mean)	**0,376**	**0,92**

After a number of statistical computations, the results of the reliability test were found to be striking. According to this additionally designed reliability test, Instrument 2 was nearly 3 times more reliable than Instrument 1. In addition, when the consistency

levels among the graders were examined carefully, a very important level of consistency was clear. Moreover, the differences among the graders in terms of consistency level were exceptionally indefinite with Instrument 2 whereas the difference was quite striking with Instrument 1. These were the findings on the basis of 7 points difference interval. Later, it was questioned whether the same table would be formed if the reliability levels were changed. In consequence, different reliability levels 0-3 and 10 were applied to the grades of each grader's first and second gradings and a number of interesting results were gathered after those calculations (See Table 4.6.2).

Table 4.6.2: Estimated reliability degrees of both instruments

Group Statistics	Instrument	n	Mean
RELIABILITY 0	1	10	**,0960**
	2	10	**,0800**
RELIABILITY 3	1	10	**,1880**
	2	10	**,5520**
RELIABILITY 7	1	10	**,3760**
	2	10	**,9200**
RELIABILITY 10	1	10	**,6020**
	2	10	**,9820**

(N = the number of graders; mean = the value of various tolerance intervals between the graders' two gradings with different instruments.)

When the reliability level was increased from 7 to 10, the reliability degrees of both instruments increased relatively, the second instrument still being better. When the degree was decreased from 7 to 3, a parallel effect that had been seen at the level of 10 was seen, and the reliability degrees decreased, too. However, Instrument 2 was again better than Instrument 1 regarding the new intra-rater reliability level. Eventually, the most surprising

results were taken when the reliability was lowered to the level of O. In other words, the papers, which were given the same grades after both gradings by the same grader, were examined and Instrument 1 was found more reliable than Instrument 2. The reason of this interesting superiority can be explained as the graders' habit of grading papers with the holistic-analytic instrument in less detailed scores. For instance, a grader who is grading a paper with a holistic-analytic criterion would prefer to give a grade like 75, more than 74 or 76, and a similar situation would be repeated again in the next grading with the same instrument; thus, having more reliable results at the level of O must not be seen as an advantage of the first instrument since the small difference of reliability at the O level was found not significant by the t-test. (See Table 4.6.3).

Table 4.6.3: T-Test for Equality of means as the significance of reliability degrees of both instruments

t-test for Equality of Means				
Reliability Levels	t	df	**Sig. (2-tailed)**	Mean Difference
Reliability 0	,331	18	**,744**	,0160
Reliability 3	-5,564	18	**,000**	-,3640
Reliability 7	-11,408	18	**,000**	-,5440
Reliability 10	-7,551	18	**,000**	-,3800

When the overall significance levels are compared, it is seen that there is no significant difference between the reliability levels of Instrument 1 and 2 at the level of 0 (t=. 331, df= 18, p= .744). This means that except the 0 level, the differences are all significant between the instruments regarding the reliability method of 7 points difference interval, and it can be claimed that there appeared significant differences between the

instruments when the amount of score difference between two gradings was increased. Moreover, the greatest difference among the various reliability levels was taken at the level of 7, which was already determined as the crucial level of confidence. In addition to the significance levels, the t value also resembles how great the variations among the grades given with different instruments were, especially at the level of 7 points interval. One step further, the 7 points confidence interval was applied on the long term results comparing them with the 1st and 2nd gradings done with each instrument. The primary level of reliability for the 1st Instrument was found 0.41 when the 3rd grading's results are compared with the 1st grading, and it was 0.408 when they were compared with the 2nd grading. The second instrument, on the other hand, provided better reliability degrees when the long-term scorings involved. The primary level of reliability was found .941 when the 3rd gradings were compared with the first gradings, and it was .929 when they were compared with the grades of the 2nd grading.

Finally, all the calculations done by different methods led us to the conclusion that in grading student's writing performance, Instrument 2 may provide higher inter/intra-rater reliability values compared with Instrument 1 when a similar grading process like the one in the study is followed. Although both instruments present acceptable intra-rater reliability values at a confidence level of .05, it has been found out that Instrument 2 can give much higher intra-rater reliability values. At this point, an important priority for the Foreign Languages School of Anadolu University in terms of reliability types should be mentioned. The writing papers in this school are not graded more than once by the same grader; instead, two different graders grade the papers one after the other. The inter-rater reliability is thus more important than the intra-rater reliability as the papers are not read by the same rater again. When the inter-rater comparison is taken as the center of focus, the clear difference in terms of rater consistency can be determined among the ten graders with different instruments. In other words, according to the results of a number of statistical tests including ANOVA, t-test and Pearson's correlation, it is found with a significance of .05 that Instrument 2 can give better inter-rater reliability values than Instrument 1. What is more, the maximum score ranges among the scorers are wider with Instrument 1 whereas the score ranges with Instrument 2 are slightly different after three different gradings.

Ultimately, reliability is considered a calculation that may be done by means of certain statistical tests. However, there is no unique statistical computation that may be used to test reliability alone. In the process of data collection, many different values were seen under the heading of reliability, which were regarded as necessary to acquire to call a study reliable. 0.7 and 0.8 were the most common ones recommended for subjective evaluations. However, to identify a certain test or a rating instrument as reliable, one must never forget the necessity of finding a basis to compare with one's data; otherwise, a multi-faced comparison would be impossible. In this study, ten graders' performances were compared from many different aspects because of the lack of a totally consistent and reliable basis. In fact, in the evaluation of writing papers it is pointless to search for a basis to compare the graders since a fixed basis can never meet the ever-changing needs of students' performances. In summary, seeing as developing a grading standard might be the best way of setting one's own standards instead of searching for, then the final product of this study was an analytic grading criterion which is a good deal more reliable than the one which is being used in the Foreign Languages School of Anadolu University at present.

CHAPTER V

CONCLUSION

5.1 Summary of the study

This study aimed to find out the inter/intra-rater reliability levels of the holistic-analytic scale which is presently being used at Anadolu University School of Foreign Languages English Preparatory Program, and investigate if the design and application of an analytic grading system would result in an increase in those reliability degrees. To contribute to the study, ten writing instructors with at least 3 years experience in grading writing accepted to grade 50 writing papers at different success levels derived from the papers of the June 2000 final exams of that school. First, the papers were graded twice by each instructor with the holistic-analytic system, having one-month interval between each grading. After the intra/inter-rater reliability values were found insufficient with this scale, factor analysis tests were applied on the components of the holistic-analytic criterion. As the results suggested all the components of the holistic-analytic criterion to be found erroneous (the proportion of difference for each component was calculated higher, task achievement 40%, essay organization 47%, accuracy 46%, than the permitted amount 10%), it was aimed to design a new analytic grading system.

The new scoring system was checked and tried on 3 sample papers by two coordinators from that language school after a pilot-training session. Later, the writing paper files including 50 papers, in a different order than its former order, were distributed to the graders again and marked once more with the new analytic criterion. In the training session, necessary details about the criterion were given, and three sample papers were graded with the new criterion. After the papers had been distributed, they were graded twice by the same raters. Six months after all these scorings, a third, long-term grading with each instrument was held so as to test the consistency of the results The scores were

then analyzed statistically in terms of inter/intra-rater consistency levels and they were compared with the holistic-analytic grading system's findings.

Finally, all the calculations done by different methods led us to the conclusion that in grading students' writing performance, Instrument 2 (analytic scale) may provide higher inter/intra-rater reliability values (.92) compared with Instrument 1 (.38) (holistic-analytic scale) when a similar grading process like the one in the study is followed. Although both instruments present acceptable intra-rater reliability values at a confidence level of .05, it has been found out that Instrument 2 can give much higher intra-rater reliability values (Ins.II: .970, Ins.I: .740). At this point, an important priority for the Foreign Languages School of Anadolu University in terms of reliability types should be mentioned. The writing papers in this school are not graded more than once by the same grader; instead, two different graders grade the papers one after the other. The inter-rater reliability is thus more important than the intra-rater reliability as the papers are not read by the same rater again. When the inter-rater comparison is taken as the center of focus, the clear difference in terms of rater consistency can be determined among the ten graders with different instruments. In other words, according to the results of a number of statistical tests including ANOVA, t-test and Pearson's correlation, it is found with a significance of .05 that Instrument 2 can give better inter-rater reliability values than Instrument 1. What is more, the maximum score ranges among the scorers are wider with Instrument 1 whereas the score ranges with Instrument 2 are slightly different after three different gradings.

5.1 Conclusion

Literature reveals three main approaches in grading foreign language learners' writing performances (Heaton, 1988; Kunnan, 1995; Raimes, 1983): holistic scoring, analytic scoring and primary-trait method. Of these three, this study aimed to compare the inter/intra reliability degrees of analytic and holistic-analytic grading (which is considered a type of holistic rating) systems.

In fact, most institutes prefer either analytic or holistic based grading systems according to their needs, considering the number of graders, number of students, amount of tolerance for the discrepancies that might appear among the raters. Because of shorter grading times and the practicality of presenting the actual results of the writing papers, holistic systems are commonly preferred among the language schools. However, while measuring the writing abilities of language learners, the 'human factor' and reliability issues are mainly ignored or totally forgotten during the grading processes. It has been widely accepted that if a holistic-based system is implemented in the grading process, the 'human factor', in other words, the grader's personal beliefs and grading habits, interfere much in the evaluation of writing abilities (Sasaki-Hirose, 1999; Bachman, 1991; Schoonen et. Al., 1997; Ruetten, 1994). The involvement of the grader's own grading measures or habits causes deeper inconsistencies among the graders' final scores when the instrument is a holistic-based one. On the other hand, holistic scoring presents a great "time" advantage, whereas analytic scoring requires longer grading times. However, analytic scoring prevents the grader from evaluating the student's performance according to his/her judgments as the standards of grading are much more concrete, having certain bands and descriptors for each writing skill.

Finally, language schools have to consider the superiority of one scale or grading system against the others, regarding not only what they present but also its own evaluation cycle. On one hand, there is a faster and more practical way of grading (holistic-analytic); on the other hand there is a less practical but more detailed and mechanic system (analytic). Since analytic scoring requires more detailed written forms of given grades for each section of the total student's evaluation, graders tend to stay closer to the analytic criterion better than the holistic-analytic system in which great discrepancies were observed. If a discrepancy occurs among the graders, it can easily be found out from the graders' scoring charts. This fact may thus alarm the rater and would not let him/her score according to his or her own standards. However, in a scoring sheet of a holistic-analytic scale such scoring differences do not reflect much about the rater's reactions since the components of a holistic-analytic scale are much wider than the ones in an analytic scale.

In the evaluation of writing, one may not expect to find totally consistent rates among the raters, because subjective scoring systems should always carry a certain inconsistency degree among the grades and the graders, which will always vary according to the institutes' testing policies. This study thus took two alternative models of grading writing as a basis. Moving from this basis, the comparison of the two grading systems were achieved, and it was made clear that, according to the correlation levels, both instruments were reliable. According to the results of three other statistical tests made to clarify both inter and intra-rater reliability values, the second instrument (Analytic criterion) appeared to be much more reliable when compared with the other grading system.

The main concern of this study was not to declare an instrument as reliable while the other is not. As was stated before, reliability is a value between O and 1, and one cannot call an instrument (whatever or however it measures) inconsistent or unreliable, since it will carry a certain amount of reliability. What can be done to understand the reliability levels of each instrument is to test and find the better one. Thus, this study was carried out on the data gathered by means of two different grading styles and the same statistical reliability tests were applied on the scores of writing papers given with different grading criteria by the same graders. Finally, it was found that under the same conditions (graders, grading time and papers to grade) an analytic scoring criterion might lead the graders to have better reliability levels not only among the graders, but also among three distinct scorings of the same rater. The holistic-analytic scoring on the other hand has certain reliability levels for each condition (inter/intra-rater) and the final results reveal that it might be called reliable according to some co-relational statistics based on the intra-rater relationship. Nonetheless, no valid proof was obtained to show that an acceptable degree of inter-rater reliability level could be maintained with the holistic-analytic grading instrument.

The fact that the results of a language test are only as good as the institute's measurements (Brown, 1996) would seem to underline the importance of the reliability of measuring instruments in testing any language skills. Besides clarifying the difference in the level of consistency among the raters, this study also revealed the need to be sensitive

in terms of consistency in grading students' performances. The use of two different grading criteria has publicized extensive difference with respect to the apparent success or failure of writing performances. Language learners should never be sure whether their performances were measured successfully or not; moreover, although they have the right to demand re-gradings for their papers in the case of a possible misgrading, no one can guarantee that his/her paper will be graded appropriately with the same grading instrument for the next time. Therefore, it is the institute's duty to apply the best method to measure the learners' language skills, avoiding the need for the repetition of scorings. To ensure valid and reliable testing methods, a language school thus has to carry the responsibility of applying sound grading methods, which have proven to be reliable after extensive analysis. This need had been the main point of motivation in the application of this study which presented an alternative grading instrument to measure the writing performances of the language learners. This alternative criterion was found to be reliable on certain conditions, but it should never be forgotten that there are no completely reliable grading techniques in testing writing nor will instruments which were found to be reliable always stay reliable, unless the necessary revisions and updating are done periodically.

5.3 Limitations

Despite all the effort to lessen the possible limitations throughout the design of this study, a number of limitations appeared at the end. To begin with, the papers, which were used in all gradings, were chosen considering their final scores given in the year 2000 with the use of the holistic-analytic criterion, which was found to have low reliability levels. Thus, considering the finding that the holistic-analytic instrument was not reliable enough, then, the papers chosen according to the grades given with this instrument may not be classified appropriately. However, at the very beginning of this study the reliability levels of the holistic-analytic criterion were unknown, so it was impossible to avoid such a limitation before choosing and using those papers.

Another limitation can be mentioned on the "norming" issue. Many experts suggest conducting norming sessions if the medium of measurement is a holistic scale. Miller (1997) defines "norming" as the initial stage of a grading session, and it is frequently done in order to achieve a consensus among the groups on how to apply the holistic criterion to certain papers. Such norming sessions were not done while the raters were grading the papers with the holistic-analytic criterion, since those norming sessions are generally accepted as a step of developing a new holistic scale (first step: forming a group, second step: creating criteria, third step: norming to scale, fourth step: orchestrating the session, fifth step: final grading). In this study, a new holistic scale was not developed, it had already been used for many sessions; therefore, conducting a number of norming sessions were not considered crucial.

The third limitation can be emphasized when the origin of the analytic criterion is highlighted. The raters themselves suggested the components of that analytic criterion; thus, it is possible that while grading those 50 papers with the analytic scale they might have acted more carefully and willingly than they did with the holistic-analytic one. The researchers commonly accept taking the raters suggestions into consideration in developing a scale; thus, when compared with the holistic-analytic criterion, the analytic scale may have an advantage in terms of rater approval.

Finally, it would be better if the graders had not graded the same papers six times with the same instruments; thus, most of the graders did not want to accept a new grading because of their hard working conditions. The graders also claimed that most of them were about to remember the content of the papers after the fifth grading. Having a familiarity with the papers after all these sessions would be considered a limitation. In fact, if the aim is to measure both inter/intra-rater reliabilities then it was a must to grade each paper many times. Actually, remembering the content may not affect the degree of reliability if the overall scores are not remembered. To avoid it, the ranking of the papers was changed after each grading; however, the degree to which this measure was successful in preventing the graders from remembering the papers remains still questionable.

5.4 Implications

In terms of the implications of this study, a new analytic criterion which was proved to present more reliable results than a holistic-analytic instrument (if the grading conditions, materials and grades are the same) could be mentioned as the most important final product. Perhaps the most difficult, time-consuming and also expensive task in designing and conducting any language-testing program is developing an instrument to standardize grading. The general intention in such cases is to search for an already-existing instrument which may save the day; however, quite often these grading instruments cannot be found, or the ones available need major revisions before using them in a performance-grading session. Furthermore, the problem of reliability and validity appear if those major revisions are applied to the grading standards which were declared to be reliable in their original forms. Taking these problems into consideration, this study presents a very useful and applicable way of grading writing papers by means of an analytic instrument which was found to be reliable for the use of Anadolu University Language School. After implementing the necessary reliability tests, the criterion may also be suggested for the use of similar language schools in their evaluation programs on the ability of writing. However, it should be highlighted that this suggested criterion can be said to be reliable under the circumstances tested; it may lack external-validity in terms of reliability results.

What is more, this study may encourage other studies that would focus on the issue of reliability in testing the other language skills. In an attempt to attract the attention of other researchers on the issue of reliability, this study also presents the importance of consistent grading, seeing that a simple misgrading may cause serious unwanted effects on a student's future career. Taking the issue of "fairness" as a basic necessity for all language programs, this study may then motivate language teachers and test-designers to check their grading measurements of all language skills including writing.

REFERENCES

Alderson, C., **Bands and Scores in Alderson and North 1991**, 1991. In Nunn, R., "Designing rating scales for small-group interaction", ELT Journal Volume 54/2, Oxford University Press, 2000.

Alderson, C., Clapham, C., Wall, D., **Language Test Construction and Evaluation.** Cambridge, U.K: Cambridge University Press, 1995. In Connor, U. & Mbaye, A., "Discourse Approaches to Writing Assessment", Annual Review of Applied Linguistics, 22, 263-278, 2002.

Alderson, J. C., "Judgments in Language Testing. Paper presented at the 12th Annual Language Testing Research Colloquium, San Francisco, C.A, p:672, 1990. In Bachman, L.F., "What does Language Testing have to offer?". TESOL Quarterly, Vol.25, No.4, Winter 1991.

Bacha, N., "Writing evaluation: what can analytic versus holistic essay scoring tell us?, Retrieved from http://www.elsevier.com/locate/system, System 29, 371-383, 2001.

Bachman, L.F., "What does language testing have to offer?", TESOL Quarterly, Vol. 25, No.4, Winter 1991.

Bachman, L.F. & Palmer, A.S., **Language Testing in Practice: Designing and Developing Useful Language Tests**. Oxford University Press, 1996.

Boughey, C., "Learning to write by writing to learn: a group-work approach", ELT Journal Volume 51/2, April 1997. Oxford University Press, 1997.

Brown, H.D., **Principles of Language Learning and Teaching**. San Francisco State University, Prentice Hall Regents, p:254, 1994.

Brown, J.D., "Do English and ESL Faculties rate writing samples differently?", TESOL Quarterly, 25, 587-603, 1991.

Brown, J.D., **Testing in Language Programs**. Prentice Hall Regents, 1996.

Carrel, P.L., "The effect of writers' personalities and raters' personalities on the holistic evaluation of writing". Assessing Writing, Volume 2, Issue 2, 153-190, 2002.

Clapham, C. & Corson, D., **Language Testing and Assessment. Encyclopedia of Language Education**, Vol. 7. Kluwer Academic Publishers, 1997.

Connor-Linton, J., "Looking behind the curtain: What do L2 composition ratings really mean?", TESOL Quarterly 29(4), 762-765, 1995.

Connor, U. & Mbaye, A. "Discourse approaches to writing assessment", **Annual Review of Applied Linguistics**, 22, 263-278, 2002.

Cumming, A., "Expertise in evaluating second language compositions". Language testing 7. 31-51, 1990. In Schoonen, R., Vergeer, M. & Eiting, M., "The assessment of writing ability: expert readers versus lay readers", Language Testing, 14(2), 157-184, 1997.

Cumming, A., "The testing of writing in a second language", Kluwer Academic Publishers, 1997. (eds) Clapham, C. & Corson, D., **Language Testing and Assessment. Encyclopedia of Language Education,** Vol. 7. Kluwer Academic Publishers, 1997.

Davies, A., **Dictionary of Language Testing**. Cambridge University Press. Cambridge, Studies in Language Testing 7, 1999.

Diederich, P.B., French, J.W. and Carlton, S.T., "Factors in Judgements of writing ability", Research Bulletin 61-15. Princeton, N.J.: Educational Testing Service, 1961. In Sasaki, M., Hirose, K., "Development of an analytic rating scale for Japanese L1 writing", Language Testing, 16(4), 457-478, 1999.

Doshisha, K.K., "Measurement of Validity and Reliability", Retrieved from: http://www.ilc 2.doshisha.ac.jp./users/kkitoo/library/article/test/design.html, Doshisha Women's College, Kyoto, 2000.

Elbow, P., "Ranking, evaluating, and liking: Sorting out three forms of judgment. College English, 55, 187-206, 1993. In Song, B. & Caruso, I., "Do English and ESL Faculty Differ in Evaluating the Essays of Native English-Speaking and ESL Students?", Journal of Second Language Writing, 5(2), 163-182, 1996.

Gannon, P., **Assessing Writing: principles and practice of marking written English.** Edward Arnold, 1985. In Heaton, J.Bl, **Writing English Language Tests.** Longman Handbooks for Language Teachers, 1988.

Hamp-Lyons, L., "Rating nonnative writing: the trouble with holistic scoring.", TESOL Quarterly 29(4), 753-758, 1990.

Harmer, J., **The Practice of English Language Teaching.** New Edition, Longman, 1991.

Harris, M.& McCann, P., **Assessment**, Handbooks for the English Classroom. Heinemann English Language Teaching, 1994.

Heaton, J.B. **Writing English Language Tests.** Longman Handbooks for Language Teachers. Longman Group UK Limited, p:135, 1988.

Hedge, T., **Writing**. Resource books for teachers, edited by Alan Maley. Oxford: Oxford University Press, p:163, 1988. In Bilash, O.S.E., "Planning for Writing Instruction in a Middle-Years Immersion/Partial Immersion Setting. Foreign Language Annals, 31, No.2, 1998.

Homburg, T.J., "Holistic evaluation of ESL Compositions: Can it be validated objectively?", TESOL Quarterly. Vol. 18, No:1, March 1984.

Hughes, A., **Testing for Language Teachers.** Glasgow: Cambridge University Press, 1989. In Oruç. N., Evaluation of the reliability of two grading systems for writing

assessment at Anadolu University Preparatory School, Unpublished M.A. Thesis, Bilkent University, July 1999.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F. and Hughey, J.B., "Testing ESL composition: a practical approach", Rowley, M.A.: Newbury House, 1981. In Sasaki, M. & Hirose, K., "Development of an analytic rating scale for Japanese L1 writing" Language Testing, 16(4), 457-478, 1999.

Janopoulos, M., "Writing Across the Curriculum, Writing Proficiency Exams, and the NNS College Student", Journal of Second Language Writing, 4(1), 43-50, 1995.

Johns, A.M., "Interpreting an English Competency Examination", Written Communication. 8, 379-401, 1991. In Ruetten, M, K. "Evaluating ESL Students' Performance on Proficiency Exams", Journal of Second Language Writing, 3(2), 85-96, 1994.

Kameen, P.T., "Syntactic skill and ESL writing quality", **Learning to Write: First Language/Second Language.** (eds) Freedman, A.; Pringle, I. & Yalden, J. Selected Papers from the 1979 CCTE Conference, Ottowa, Canada , pp:162-170, 1987.

Kroll, B., "Assessing writing abilities". Annual Review of Applied Linguistics, 18, 219-240, 1998. In Connor, U. & Mbaye, A., "Discourse approaches to writing assessment", Annual Review of Applied Linguistics, 22, 263-278, 2002.

Kunnan, A. J., **Test taker characteristics and test performance: A structural modelling approach.** University of Cambridge, 1995.

Lien, A.J., **Measurement of Learning**. Wisconsin State University, 1971.

Lloyd-Jones, R., "Tests of writing ability", University of Iowa, p:155, 1989.(eds) Tate, 6., **Teaching Composition**. Texas Christian University, 1991.

McDaniel, B. A., "Raiting versus equity in the evaluation of writing. Paper presented at the 36th Annual Conference on College Composition and Communication",

Minneapolis, 1985. In Ruetten, M.K., "Evaluating ESL Students' Performance on Proficiency Exams". Journal of Second Language Writing. 3(2), 85-96, 1994.

McGirt, D., "The effect of morphological and syntactic errors on the holistic scores of native and non-native compositions". Unpublished master's thesis, University of California, 1984".. In Sweedler-Brown, C.O., "ESL Essay evaluation: The influence of sentence-level and rhetorical features". Journal of second language writing, 2(1), 3-17, 1993.

Miller, M., "Alternative grading models: Holistic grading". Retrieved from http://www.cstw.ohio-state.edu/tutor/holist.htm, 1997. In Oruç, N., Evaluation of the reliability of two grading systems for writing assessment at Anadolu University Preparatory School, Unpublished M.A. Thesis, Bilkent University, July 1999.

Nunan, D., **Language Teaching Methodology**. A textbook for teachers. Prentice Hall, 1998.

Nunn, R., "Designing rating scales for small-group interaction", ELT Journal Volume 54/2, Oxford University Press, April 2000.

Oruç, N., **Evaluation of the reliability of two grading systems for writing assessment at Anadolu University Preparatory School**, Unpublished master's thesis, Bilkent University, July 1999.

Perkins, K., "On the use of composition scoring techniques, objective measures and objective tests to evaluate ESL writing ability". TESOL Quarterly 17(4), 1983. In Bacha, N., "Writing evaluation: what can analytic versos holistic essay scoring tell us?", Retrieved from http://www.elsevier.com/locate/system, System 29 , 371-383, 2001 .

Pollitt, A. & Murray, N.L., "What raters really pay attention to", **Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC). Cambridge and Arnhum**. (eds.) Kunnan, A.J., Cambridge: Cambridge University Press, pp:74-91, 1996.

Porto, M., "Cooperative writing response groups and self-evaluation", ELT Journal Volume, 55/1, Oxford University Press, p:40, January 2001.

Purpura, J. E., **Learner Strategy Use and Performance on Language Tests: A structural equation modelling approach**. Cambridge University Press, 1999.

Raimes, A., **Techniques in Teaching Writing**. Oxford University Press, p:3, 1983.

Reid, J.M., Teaching ESL Writing. Prentice Hall. New Jersey, 1993. In Bacha, N., "Writing evaluation: what can analytic versus holistic essay scoring tell us?", Retrieved from: http://www.elsevier.com/locate/system, System 29, 371-383, 2001.

Ruetten, M.K., "Evaluating ESL Students' Performance on Proficiency Exams", Journal of Second Language Writing, 3(2), 85-96, 1994.

Sakyi, A. A., "Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions", University of Toronto, 2000. (eds) Kunnan, A.J., **Fairness and validation in language assessment:** Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida, 2000.

Sasaki, M. & Hirose, K., "Development of an analytic rating scale for Japanese L1 writing "Language Testing, 16(4), 457-478, 1999.

Schallek,J., "What is writing?", Retrieved from: http://www.syr.edu.elfire.com/me3/Schallek/index.html,on 07.05.2002, p:2, 1999.

Schoonen, R., Vergeer, M. & Eiting, M., "The assessment of writing ability: expert readers versus lay readers", Language Testing, 14-2, 157-184, 1997.

Scott, R., "Changing Teachers' Conceptions of Teaching Writing: A collaborative study", Foreign Language Annals, 28, No.2, 1995.

Song, B. & Caruso, I., "Do English and ESL Faculty Differ in Evaluating the Essays of Native English-Speaking and ESL Students?", Journal of Second Language Writing, 5(2), 163-182, 1996.

Sorenson, S., "The advantages of writing in Language Learning", Retrieved from: http://www.ed.gov/data-bases/Eric digest#62, Eric Clearinghouse, 1992.

Sweedler-Brown, C.O., "ESL Essay Evaluation: The influence of sentence-level and Rhetorical Features", Journal of Second Language Writing, 2(1), 3-17, 1993.

Thorndike, R.L. & Hagen, E. **Measurement and Evaluation in Psychology and Education**, 2nd ed. New York, 1961. In Lien, A. J., **Measurement of Learning.** Wisconsin State University, 1971.

Traub, R.E., **Reliability for the Social Sciences.** Theory and Applications. Volume 3, Sage Publications, pp:73-75, 1994.

Turner, C.E. & Upshur, J.A., "Rating scales derived from student samples: Effects of the scale Marker and the Student Sample on Scale Content and Student Scores", TESOL Quarterly, Vol. 36, No.1, Spring 2002.

Turrisi, P., **An Introduction to Writing Across the Curriculum**, Colorado State University, p:2, 2000.

Upshur, J.A. & Turner, C.E., "Constructing rating scales for second language tests", Oxford University Press, ELT Journal Volume 4971, January 1995.

Walker, F., "Evaluating Student Writing: Methods and Measurement", Retrieved from: http://www.ed.gov.databases/ERIC-Digests/ed 315785.html, 1988. In Sorenson, S., "The advantages of writing in Language Learning", Retrieved from: http://www.ed.go./databases/Eric digests #62, Eric Clearinghouse, 1992.

Weigle, S.C., "Effects of training on raters of ESL compositions", Language Testing II, 197-223, 1994. In Schoonen, R., Vergeer, M. & Eiting, M., "The assessment of writing ability: expert readers versus lay readers", Language Testing, 14(2), 157-184, 1997.

APPENDICES

Page

APPENDICES

APPENDIX A

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

(Taken from the June 2000 Final Exams)

DURATION: 75 Min.

Write an opinion essay either agreeing or disagreeing on ONE of the statements below.

1. Television destroys communication among friends and families.
2. A dormitory is the best place for university students to live.
3. Parents are / are not the best teachers.

APPENDIX B

STUDENTS' WRITING EXAM PAPERS (Taken from the June 2000 Final Exams)

Paper

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

1

PARENTS AND CHILDREN

Every children like to parents. Children, long time live in the parents. Children, teach to all of them rules next to parents. Which good, which bad ... All of them, teach to of parents a lot of thing.

First, are parents have to children interest to school education. The school, teach to anything of the child, but not to next more than parents.

Second, the parents teach to have got child to live, to eat, to walk ... The parents to have got child, he's or she's friends teach to respect.

Third, the parents have got to child, teach of to be socialty. To the other people strong of the communicate. The parents, to child of trust win and child, trust herself or himself.

As a result, one child teach to lots of thing of the parents. Child, teach to school or to friends, but that is not enough. Parents are the best teachers for children

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

2

MY TEACHER'S

People are communication in the world. Everybody have a teacher. They are teach a lot of everythink. Therefore, parents are the best teachers, namely parents are teach good speak, want to be good friends and want to be clean.us.

First,Parents are teach speaking, because they are want to be good speak us. Accordig to the statistic, good speak is cause parents, since we teach a lot of everythink. For example, I am stand up my friends. They are good speaking, but I am not good speaking, while they are kidding me.

Second,Parents are want to be good friends us, because good friends are need us. Prof. Dr. Murat Yılmazsoy, " Friends are more important for people." For example, I am ill. My friend is going to hospital. Than,He is going to drugstore, and take drug. Finally I am good.

Finally,Parents are want to clean us, therefore they are take new wear, new shoes, new belt and new tie. Prof. Dr. Erdem Güvenç says, "To be clean is important for people life." Parents are very working, since they are happy for us. For example,My sister is wearing dirty. My parents angry, as they want to be clean wearing.

In brief,People a lot of teach us, but Parents are important for us, because They are teach life us.and we to be succes in the world life. I love my parents.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

3

DORMİTORY AND HAUSE

A dormitory is the bad place. Students lesson don't study,Because is very crowded and don't relax. A dormitory is very cheap more than house but darmitory don't live.

The student be dormitory unsuccesful. The why don't lesson study. Either a why don't sleep and don't relax. The dormitory very pullutiton. The life only one place home

The house be very expensive. Havewhere is dormitory don't life. The home very contable because relax, clear. The students lessan study. İn my opinion dormitory don't live.

Hause is very good place. The dormitory is very uncantable place. In the short darmitory is very bd place.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

4

-DORMİTORİES_

All students want to win a üniversite, but they can't win a universite in their town, therefore they cango to other town. When students went other a üniversite, üniversite has got can be best dormitory. However students mustn't live a dormitory. In short dormitories aren't best place for üniversity studentsto live. There are very important three cause.

First, students don't study very well,Because students haven't got theirself a room. A lot of students can live in a room, so very noise can be in room. For example, a student can't study his exam, therefore student can't pass.

Second, students don't go outside everytime. They must let from their dormitory, whereas students want to be free. For example, a student don't go his friends at midnight.

Finally, student don't know responsiable or the life isn't explained by student. For example, there a lot of homework, so student learns the life. If students live in dormitories,They don't learn the life.

In brief, dormitories aren't comfortable for student. A student must find someone and students must live a home.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

5

Advantages

The dormitory was building for students. The student to advantage dormitory live for example dormitories in the campus doesn't transportation problem. The student doesn't give for money transportation. That help the students.

A person's behaviours, develops in dormitory, because the student meets much new people. The student learn new information.This informations develop the student's intelligent.

The student doesn't for television much studies lesson. When a student was while working lesson helps friends. The student to advantage lesson. The other doesn't advantage

We opinion students stay dormitor much luck, because that students has got up advantage. That expect a dormitory is the best place for university students live

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

6

The Anadolu University have a lot of students therefore students therefore students have to staying problems. Generally students haven't many and the university is opened the door for students. Besides, if student's have to lots of many students stay at special dormitory. For example for a lot of dormitory; Anadolu University dormitory Dumlupınar dormitory and special dormitory And than student dormitory.

Firstly; Anadolu University dormitory have to 500.000 students therforethe live very hard for students because the dormitory have any bedroom and students can't wash our clothes. Students must live with this problems Sometimes the students can't student with this problems.

Second; The Dumlupınar dormitory very far to Anadolu University. Students aren't arrive at the time to school. Beside must to stay at the same room. For example; 6 students therefore the live very difficult for student. One student want to listen music other student angry to friend. Something wasn't way to the dormitory.

Third; The special dormitory is ideal student dormitory. The live very easy for students. Because; Dormitory is cleaned by survey. All day have to special food. You want to live alone. But the dormitory close the doors at 11.00 o'clock, only problem is it.

In short; if you have lots of many, you could stay at the best place. If you haven't many, you cauldn't want to live.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES 7
WRITING FINAL EXAM

The dormitory are the best placee.but something together there are not valeybol football places. That for students not get used dormitory therefore very very boring. Therefor a lot of student going to home because home is very good, but not cold. Every students not live a dormitory because They don't like but somethink students for not money Therefor a dormitory very sheep.

The first ! There are a lot of dormitory but There are diffrent. Students dormitory and for forigen costomer lojman and pancion, our lojman is very vanderful places because There are in room is television, kendisin and minibar. The University dormitory is good but sometimes, in roommate is very best friendly, There are dormitory is very dsplin There for students a lot of going to home in room not listen to music an watching television but out garden playing football valeybol an tenis. The best type university students because, University students have to study the University is to diffirent.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

8

_ Television not destroys communication _

People come face to face al ot of problems in ancidant time first problem communication. They had been inventing television before they listed radio. Radio not usefull communication machine for they, so that they seeked another communication machine and the television was invented by people 65 years ago so that they use television in al ot of communication place.there are thee important

The first, television is very important for communication. If television use for a good aim it is very important of this communication, for exampal when apeople watche television if the television are given good programs, they takes good things and they are usefull people.

Second television are important in family life People learn al ot of things of television this is family life My friend says " if television isn't a lot of people not know in family life" beause nowadays al ot of people's mather and father are working and child not see family.

finaly television is important for culture when people watche television they meet diffren people and contrys and they want to go this contrys.in this contry meet another people for exampel Coşkun Aral's program.

in short television nessary for communication for people so that they meet another people.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

9

COMMUNİCATİON

T.V is found by macine engineeer at 1860 Many people watchs the TV in spare time, but the T.V not good everytime for our Specially, sometimes we don't speak friend and families. Since, the T.v brake between our communication friends and families What we can do? When we has got in spare time, we can spor, listen to music and walk with our friends and families.

First; the spor. The spor is not necessary only healty; Many people makes spor with friends, and they are enjoyale. For example; you think a team, they are calm not play, they are speaking, playing and such. It's very important for human. Besides, the spor useful for our healty.

Second, listen to music. When we listen to music to relax. As if we go to the different world. We can speak our friends and families. Christina Altock, who is a writer, says; "You must listen to music everytime because it's necessary for relax speak." Certainly, it's true.

Third; the walk. The walk is necessary, everytime. Since, when we walking, we can more than think, and good say something our friend and families. According to resaerch, %95 people can't say something, when walking to say. We can do. Besides the walking useful for healty. May be it's not enough, but good.

In brief; we can watch the T.V, but we must speak with everybody. Since, they are important for our life. When we listen to music, play the spor; and walking, we can strong communication. We not forget never, If they not to be at our life, we are worry.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

10

A dormitory is the best place for university student to live.

AMONG A DORMITORI

A dormitory is university students for economical. Student is nessecery a dormitory. Student is for big problem. Student is a latta friends.

Exampel Anadolu University school is small a dormitory. Students for not usefull. Student is roommed very crowdet. Because student is not relax.

Students for a dormitory to live usefull. Students is soclial and friendly. Students is for the economical. Students is for to live dificult hause. Becaus students is expensive.

Students is for new big a dormitory to nessecery. Students is relax begam. Student is crhative begam.

Popels is helped students. They are our the light

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES 11
WRITING FINAL EXAM

YANKS IS A STUDENT TİME

Yanks win a University time everything begin new Yanks live a place and familys leave Yonks go to other centurys and Yanks is a students. Students have a dormitory where as home. Students live dormitorys for students; have 3 avendaj

First students have a lot of frends therefor they have happy and they are not sing students have a roommates

Second students study more very and life confortable because frends help frends however student live home frends no help. Thirty Dormitorys cheap for students student live dormitory therefor eletiry many water many ....no problem

Finally students have two luck Home whereas dormitory in my opinion Dormitory is good place and usufull for students

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

12

in Turkey new technology more important television. Many people way of life seeing and listening many different and important television.

at first; this is age of new technology and interesting life; besides tereble street, new big centrum mad people, goverment works.possitive and negative a world communication television.

at Second; in Turkey watch TV.developed good program and stand up. This program is has been time of a day of day and long years funny with enjoy time.

at finaly; many bad are has been beautiuful world take of kinds conversation and different way of life the world communication for good technology television.

in world healty life with friends and families beautiuful a communication. There, for more important technology communication television.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES **13**
WRITING FINAL EXAM

First of all a dormitory isn't the best place for university students to live. Because, Dormitories very dirty in the our country. Students shouldn't want to live in dormitories. Dormitories are very bad in my country. Students have a lot of problems in dormitories. Such as Dormitories have very dirty, eat problem, very bad friendships.

Day by day dormitories are very dirty in my country. Students live six person one room in Eskişehir Yunus Emre Dormitory. As a result; Rooms are very fast dirting. Last year One students haven't got air him room. So this person died.

The other reasen is food problems. Students have got a few food menu. They haven't good food. They are giving not clear food. There are a lot of very small animals in their food. Last week seven persons went in hospital. Because; They were loving of small animals.

Another reason is In terms of bad friendships problems. Students aren't making friendships in dormitory. They are fighting. Last day they was very big fight. Accordingly;A lot of students went in hospital.

In brief, students shouldn't live in dormitory every time. They can go to live in houses. As a result they will be free persons and haven't got any problems.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES **14**
WRITING FINAL EXAM

A dormitory is the best place for university students to live.
First of all many students came to university first times they are a forigan In this contry and they are don't know there.
The students need to stay somewhere. The first choos in a dormitory for living. It's very advantages place for student.
It's not very confortable but goot place for a student. This place it is not expensive. If they want to stadi lesan, it have a smol libery.if they wont to enjoy there have a play place.
I think this is a very good.
Finaly I like all of the student came hear

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

15

THE EDUCATİON SYSTEM

The education is the best way for devoloping countries in the world; therefore, the country is starded the smallest family member, if the country used developing system, parents should be besd way and teacher

The most and first reason the children are learned by their family.because the children lived their family undil adult term

The most and second reason; people have to be obtimisticfor their children; therefore, family members have to take learning course for depent on their children,

The most and third reason; about how to use parents for give a something eachother in the house, for example; last year, all of the newspaper wrote about this action some father's son speak so fluently and write; but, child was four years old, therefore all of the speaker asked him; how did he learn his son, he said one sentence, everytime they live together

The final reasan; ıf the people had successfully they would live their family; because, one finger is a something, two finger is a everthing.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

16

DORMİTORİES AND STUDENTS

The most of the students stay in a dormitory during the their university life, but several students dislike from the dormitory, because of several reasons. The dormitory is not the best place for university students life

The first and the most important reason is dormitory's condition. Because some students don't habit dormitory's room, toilet, and bed. For examples dormitory's room are crowded, small and unreasonable for students.

The second reason is meal. Students' meals are made by others people. That people thinks only their money. Because the meals are bad smelling and bad loking because of that bad materials. When the students don't eat dormitory's meals, they abendon from dormitory.

The thirt reason is dormitory's rules. All of the dormitories have a lot of rules. For instance ıf the students come late to dormitory, they don't into dormitory. Because this is rule. An other example ıf the student fight with others students, they are threw from the dormitory.

As a result, a lot of students are staying in a dormitory. They dislike dormitories and they think leave from the dormitories.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

17

DORMİTORY LİFE

Dormitory life can be best place for the university students; for example, country dormitories. These dormitories are bigger than the special dormitories. Country dormitories are cheaper than the other special dormitories have single room, however country dormitories have more than four persons room.

Firstly, many students choose dormitory for their university life; for example, one student from Antalya goes to Ankara for his university, After that, he is poor and he can't live in house, so that, he choose to live in a dormitory,the most important reason house is more expensive than the dormitory

Secondly, if the person isn't poor, he want to live single in his life, so that, he want to live single in his life, so that, he want to live in a special dormitory. Especially, special dormitories are being in the city center. Special dormitories are chosed by rich students.

Thirdly, if student can't be study easily in dormitories he can choose one of the most expensive dormitory. Many dormitories have got studying room, so that, many students can choose country dormitories,However, some special dormitories haven't got any studying room.

Finally, one student who is easy going wants to have many friends, so that, he wants to live in big dormitory like a country dormitory. One student can grow his ability to live in a dormitory, so that, dormitory life can help student life.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES **18**
WRITING FINAL EXAM

IS IT DANGEROUS?

The tecnologhy has been developing so fastly. And it changes our lifestyle. A lot of new things, invention and process have been found for humanbeing everyday. But sometimes ıt doesn't go well, and out of the way. Because many tecnological things capture humanity.such as television destroys communication among friends and families for several reasons.

The first and most important reason is that hero person, new and the best things are dissepeared by television so people effect that who watched them. For example action films and magazines programe are mention about beautiful life.despite can not be real.

Second reason is that we lost our manner logic like an animals. For instance while we eat something with our families front of television, do not talk about anything about us, so just look television.

Final reason is that if wewatch soap opera much more our manners changes so we imitate them. Both we lost our personality and character. As a result we sperate to our friends. Because person don't like it and we can not be among them.

İn brief we shouldn't watch Television too much since television effect us a lot of ways. It never emancipate.to us. Therefore we shouldn't make much of television. We shouldn't make much of television.we should use only about communication and give a news.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES **19**
WRITING FINAL EXAM

Students Living Area

A dormitory is the best place for university students to live. Students who study in university have accomodation problem. A few students rent an apartment, on the other hand, a lot of students live in dormitory. Most of the students haven't enough money and experience, in addition, they have difficulty living in an apartment with friends,Therefore, they prefer a dormitory to an apartment.

The students coming countryside aren't plenty of money, because their parents work on farm and farmwork doesn't get enough money. The life of conditions requires them to live in a dormitory. They aren't able to do anything to live in an other place. The students who are in the same conditions are happy with each other. They have the same problem and they tolorant each other. When when the sky becames dark, the students living in dormitory put out their music instrument and they sing songs. Although, they have rather money they like to live in their dormitory.

An other circumstance is life experience. A lot of students live with their parents. They haven't also any idea to live alone. They can't even boil an egg, furthermore the housework is very difficult forexample; washing dishes, washing clothes, cleaning the floor. So it is the best way living in a dormitory.

Three or four students rent an apartment, because the rent of an apartment is high for one student. At first, nothing doesn't go wrong. The students arrange the houseworks. In particularly, at that time they have an argument. Some houseworks is more difficult than the others. An other problem is for the students coming guests. Some students like coming guests most of the students not. Therefore, it is difficult to live in an apartment with friends, although you are close friend.

We consider the other choice the students had better live in a dormitory. Living in dormitory cheaper and more exciting than the others

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

20

A part from Education Problems

Education is a big problem with it's difficulties One of the difficulties is where to stay! A lot of students move to different place sto complete their education. Some of students prefer to stay in a dormitory but also it brings some problems and it has some difficulties. We can state the problems, students meet, in three part.

First, students meet with a hunger problem. Non of dormitories give lunch or dinner. A student living in a dormitory must eat outside. By this way student spend the big part of his money.

Second one is the phsyocology that students stay in. For example, living in a room with six people may cause some problems; one of them wants to study opposite of this other one wants to listen music. People should do what they want but if you are living in a dormitory and if you live with a lot of people sometimes you can't.

The last and important one is the pressure in dormitories. Getting in or going out from dormitory is already a problem. Three week before. A friend of mine came to my house at midnight because he couldn't get in, the security did not let him get in. In other words you are not completely free in a dormitory.

In brief, students should stay somewhere. It may be a dormitory or a house. Before living in a dormitory students should consider the ways and the acts they meet. Hunger problem, phsyocolgic problem and pressure can effect students in wrong way. We came here for our education not for swimming in problems.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

21

TELEVİSİON DESTROYS COMMUNİTY

Television is important part, which musn't be do, people's life nowadays. From 1960's to twenty first century television has become more popular than last situation. People spend their time in front of TV set; furthermore, it destroys comminication among friends and families. It is not only family or friendship problem but also social problem.

The first and most important part of destroying communication begin from friends Before the television, friends were always together. They went everywhere as a group. Because the television, comminication among friends has broken. First, person who waste his time in front of TV set; don't interest his friends. Then he don't understand his friends. Finally their friendship is broken because of the watching television.

Another part of destroying communication among family. Like every family, people came their house after their work or school then ate a dinner together in the past. In contrast, these days people have eaten their meal in front of TV set. This is a kind of caos. This caos begun watching football match after add soup operas, discussion programs, cartoons and magazines. Day to day parents have become irresponsible people. They don't interest their house and children's problem. After some time every person in family be a different from each other.

The third part of destroying commucation begin in social. After watching television, every person hate everything. They always want to watch their favorite programs. A person need help while other person watching television, he can't do anything about his work because other one don't do something to help him. This situation is same in streets, markets or train stations. Finally, every person in social don't communicate each other.

In brief, since television has been part of our life communication has destroyed. As a result people in family, friendship or social, shouldn't be more interested television than their life's other thing.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

22

- LIFE IN A DORMITORY –

Most of the students think about where we should stay at the university. Actually, they have no chance. Generally, they stay in a dormitory. It has got some advantages and disadvantages. Therefore, you learn to share something with your friends, you know a lot of different people, but you can meet with bad people at the same time.

All the students are far away from their families, and of course they need the others' help. If your money has finished, you can borrow your friend. It's not only for money, but also for problems. All the things decrease when they share.

There are many people in different citiesin the dormitories. It means, you 'll learn different cultures, religions, foods ...ect. At the same time, you know the people day by day. You learn how to behave them.

We don't forget that, there are a lot of bad people around us. They will be in the dormitory, too, but the only thing that we should do not to talk about everything until we know them.

As a result, living in a dormitory as well as living with your family. They have only some responsibilities, that's it. Sometimes, your friends may be more closer than your family. The most wonderful thing in dormitories, you will have nice friendships and memories.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

23

A lot of university students choose dormitories to live. So that reason a lot of students live in dormitories Because dormitories have several advantages for university students.

First advantage; Dormitories are the cheapest place for university students to live. Because University students don't earn money and don't spend much money. They must think their financial state. They don't pay bill if They live in dormitory. These event is very important for student's financial state. So that reason University students who live in dormitory have advantage about money.

Second advantage University students who live in dormitory have regular life. Because Dormitories have some rules. If students don't accept these rules, They don't live in dormitory.if students live in home they haven't any rule and don't control their life so that reason Dormitonies are very important for University students life.

Last advantages, University students who live in dormitories meet a lot of new friends easily. So that reason They who come from other city habit new city quickly. They who live in dormitory have chance to study lesson with friends.if These students don't understand lesson, They can ask their friends. Every dormitory have small hospital. They are gone these hospitals by their friends when they who live in dormitory be ill. So that reason They who live in dormitory have a lot of chance about meet new friends.

So that reasons Who live in dormitory is very lucky than another students. Because dormitories have very advantages for university students to live So that reason a lot of university students want to live in dormitory.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

24

DORMITORY LIFE

There are a lot of students all over the world. These students have a lot of need, and their need changes according to their school life. One of the most important of these need is place which students can live during university. Especially, this place must have evry kind of chance for university. Therefore, a dormitory can be enough because it has a lot of students as each other, enough study place, and a power which breaks students up such a dormitory.

firstly, in dormitories, the most known way is theirs being crowded which cause friendship, love among students. For example, a person who won university exam feels strange when he/she first came to other city, and he/she need a friend as her/him. the best place which he/she can find this friend is a dormitory.

.second, it is a fact that a student without enough study place never can be succeful. Whenever student want to study he/she can find every kind of metarial, for example students can not find another place which has a lot of people who can teach everything to student except dormitory when they need to learn about a subject.

.third one, A student can hel each other limitedly. So, All the students need other thing which can help more, while students is within more difficult chonce. This effect is government. For example, After an illness or accident dormitory does necessary thing. Everytime, a person who can help can be found in dormitory.

In short, A dormitory has every kind of chance for a student, and is the best one for living during the university life, and if a student needs to choose a place she/he must choose dormitory.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

25

Television destroys communication among friends and families.

Before the television invented, people sat in cafés or their houses among friends in evenings. Children played games and develped new ones in nature or in their neghborhoods. People prefered reading a book etc. Now all of these people sit down in front of the television without talking anyone and doing nothing.

It isn't long time ago people had have more friends than now. They were meeting somewhere with friends and talking with them about their problems or happynesses. Although people prefer watching television programs and other people talking without they talk with their husbands, wifes children or friends now.

Furtermore about children specialists says children who watch T.V for long times, lose their imaginations. Moreover they haven't got any friend and most of them aren't successful in school.

Other important thing is about people hobies and inteligents. They had would like going teaather, collecting something, makeing , sport and reading book etc. Unfortunately they don't now. They lose communication between their friends. They prefer watching TV to talking their families.

As a result TV. Destroyed not only communication between people but also be lazy them. Therefore TV is the worst enemy of communication between friends and families.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

26

LIVE IN THE DORMITORY

In the world and in Turkey some collage's students live different own towns. So they stay somewhere, for examplethey stay in dormitory, in house, in relstive's house, etc. they don't usually choose and want to live dormitory because it has got some disadvantagous, for instance studying, sleeping, eating are big problems in dormitory.

First problem is lesson's studies. Because somebody says " We haven't got good classrooms, desks, etc." They want to only this small needness, but government doesn't give it to them. For example this year I stayed in government's dormitory, and I've got some like this problems, one more thingour classroom is very big and everybody smokes in it, but I and some my friends hate of cigarettes smoke, so we couldn't do anything, and we studied on our beds in our rooms.

Second and big problem is sleeping for students. About 6 or 4 students stay in one room, and they don't know theirself. One body likes to sleep early, so another ones like to sleep late. However one person doesn't do anything, and he or she starts to sleep lately. "What is this person crime. This year I saw the similarly problems in our dormitory. Moreover this problem occurred somewhere because nobody respect to their friend or friends. They didn't learn when sleeping and get up.

Last another problem is eating. In dormitory they don't find good meals everytime, so they go to restaurants, cafes for eating. In addition they take little meal who give to big money. Some bosses know to this problem in dormitories; consequently they sales expensively. In this year one student eating, drinking outdoors. He or she gave about two million for them in one day.

In brief, they have got same problems in dormitory, so they live in dormitory very diffucult but those problems solve very easy. Government wants to solve these problems, first opening the new and undirty classrooms, then decreasing students numbers in one room, next cooking delicious meal. In short, I think they solve very easy and clearly. You should see them quickly, please!

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

27

PARENTS

Parents are the teachers for children. I agree this idea, because children learn something about everything from their parents before the school. Also parents always teach something while their children are going to school.

Firstlty parents teach something to their children about real life. Real life is different from books. Schools always teach something, but they aren't teach real life.

Another important thing is about help to people. Of course schools teach very important things but parents teach very very important things about life. For example, children learn to help people in their family. Because parents want to grow good children to their country. Therefore they must be carefully about it.

Last important thing is, love animals and people. Children learn to love animals in their parents, because parents know if a person doesn't like animals, the person won't like people, too. So parents learn these things to their children. Also we can say many of thing about parents.

In short, parents's aims are different from schools. Schools want to learn something about lessons but.family want to grow a person who knows everything about real life. So I thing parents are the best teachers for their children, because parents teach many things to their children.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

28

OUR PARENTS

Learning is very important. Who is the best teachers in our life? I think fathers and mothers are very important for teaching.

Some of the people say we can learn everything from school teachers, but we start to learn since we were born. I know our school teahers are important, but our parents are always in our life, everywhere, everytime!!

My life is a very good example, I think everyone's life is a very good example themselves. When we were born, we didn't know anything, but later, our parents teach us eating, running, walking, smiling, speaking, ... etc., and we learn thinking to them. Thinking is very important for our life we must think everything when we start something.

Everyone has a different characters because of our different life style. If I saw to tell a lie from my parents, I learned to tell a lie. I think our parents are best teacher for teaching although some of their sons / daughters don't want to learn good things for them.

In brief, parents are the best teachers. We learned the most important things from our parents.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

29

- THE BEST PLACE TO LİVE -

University students, who study in another city, can live in a dormitory or in a house with his friends. Most of the students are thinking to live in a house is better than to live in a dorm, in the other hand families are thinking that a dormitory is the best place for students to live.

Dormitories are very crowded places, therefore they are sometimes too loud and if you want to study for your exam, it is very hard to study. You have a lot of roommates. You want to sleep, but the other don't. Your friends can't visit you in the dormitory and you can't go outside after eleven o'clock. They are some bad opinions for dormitories.

In the other hand, if you rent a house with your best friends, you have your own room. Your friends can visit you and you can go outside every time. You can work alone in your room and watch on your comfortable chair.

However these good thiings, to live in a house is more difficult, because you have a lot of responsibilities. You must pay the electric and water bills. You must talk with the apartment manager and with your neighbors, clean your room, make something to eat, wash the dishes...

In short living in a house is more difficult, because of the responsibilities, but it is more comfortable. I don't agree with the families and therefore I choose the house life

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

30

Diffulties of Dormitory

I have been lived in a dorm for five years, therefore I know all of time diffuties of living in a dormitory. Consequently, I don't agree dormitory is the best place for university students to live.

At first, there are many students lives in dormitory so there are many problems to each of them. They are fighting because of very simple problems

Second, you can't do any thing, what time do you want. For example; you are hungry, so you go to canteen to eat something, but you have to wait. You need to go WC, you are going there, again you have to wait, if you are tired, you will sleep and rest but you can't rest because of noisy, also you can't study. As a resut you lost line.

Then, you have be careful. Two months ago, I forgot my waluet in the roon, on my bed and I went to the garden, suddenly I remember it and I start to run to the room, but I didn't find my waluet. In addition, some of my friend's mobile phone were stolen like this.

These are only some of difficulties that I want to say. I left to the dormitory a month ago. I living in my home now. It is better than to live in a dorm, more comfortable, more clean, and more cheap. Also I learn houselife. Consequently, I agree houses are the best places for university students to live.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

31

LIVE IN DORMITORY

Dormitory is a place which students live among their university life. It always builts near the university, because of transportation.

Live in dormitory is very easy for students. It is very cheap and students don't pay much money for transportation. They spent their money for food and drink, of course if they don't smoke cigarette and drink alchol.

Live in dormitory's advantages aren't only spend a few money but also you can learn how can you live without your family, also you can introduce a lot of friends and you doesn't forget your college friends.

Although you know about live in dormitory's advantages, you don't want to live in there, you can live in house. It isn't bad idea, but that time you must spend much money. First of all you must buy some furniture and you must know cooking, because you miss your hometown food, also you pay electricity and water fee, so live in home is very expensive and very difficult for students.

Turkies life standarts aren't good. A lot of family need money. Because of it, so me family don't give too much money for their children, so dormitory is the best place for university students to live.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

32

Staying is a big problem in students life. Students have different alternatives to stay. Students usually choose dormitories to live. Therefore its reasons are: being cheap, isn't far away to school and having a lot of friends.

Money is very important in students life. Consequently students also choose cheap places to live. For example; a house costs 50 millioun, a dormitory costs 6 millioun. In this reason students chose the dormitories.

Going to somewhere is too difficult in the cities. If your home were far away to school, you could be late to school. In this reason students choose the dormitories to live and theyy don't be late to school.

Students can find a lot of friends in dormitories. They can do a lot of things with their friends. For example if they lived in a home, they would be alone and their lifes would be boring. Therfore, students choose the dormitories to live.

In short, living a dormitory is the best way for the students. Students have a lot of reasons to choose the dormitories. In this reasons, dormitories have a special role in students life and it makes our life easy.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES 33
WRITING FINAL EXAM

_Parents are the best teachers.

When person born s/he doesn't know anything about the world and the life. She can't speak, can't walk, can't do anything. People who responsible from bring children upare mothe and father. Parents should be carful to these sabjects when bring up their children. They should know when, how and why they bring their children up we can order them.

Firstly, the time is very importan to teash a child.parents from borth their child should begin to teach, because the children are able to learn from their borth.perents should saparet time to teach.

Secondly, parents should intrest with children. They should teach them a good things and do not shout and do not talk bad thing infront of them. Parents should keep their children far away from bad friends.

Thirdly, parents should bring up the children in the best way, because these children in the future they will be teachers to their children.

Shortly, person when s/he born like an empty box. If parents fulled it with good things they can be the best and their children after they grew up the will success, on the other hand if they don't full it with good things they can't be the best teachers.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

34

Most of the university students leave their hometowns and start to live in strange towns for their education. Unless they have no relations living in those towns, they face too many problems. The biggest and the most important problem is accomodation for these students. Renting a flat or living in a dormitory are the choices both of which have its own problems. In my opinion, a dormitory is not the best place for university students to live for three reasons.

The first reason to support this idea is that a dormitory does not serve the equal comfort of aflat. In a flat, a student does not have to share his bathroom, kitchen with another student, but in a dorm he does. In addition, cleaning, eating are other problems a students has in a dormitory. A student can not do whatever and whenever he wants to do. Furthermore, it is forbidden to have a television which lets student to get bored in a dormitory.

The second of the reasons is the problems that a student face with because of living together with other students. For example, I live in a dormitory and share my room with other six students. Because each department has its own exam dates, we sometimes argue about studying in the room late in the night which destroys others' sleeping. Also, our hometowns and lifestyles are very different, we have troubles understanding and sharing our ideas.

The third reason is the rules about the dormitory. In a dormitory, there are clerks whose duty is to control students. Of course, rules are necessary and vital when there is a lot of young people living together, but some of them are very silly and let us to ask questions about them. For instance, why we can not have a TV, why we can not change our beds' place, why it is forbidden to bring food in the dormitory and so on. Always they answer " This is the rule and everyone will obey."

In short, a dormitory is not the best place for university students to live according to the problems I mentioned. I would rather living in a flat with my best frend and enjoy the comfort of it than being crazy in the dormitory. Students living in a flat should be aware of how lucky they are.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES 35
WRITING FINAL EXAM

BAD SIDES OF A DORMITORY

Nowadays, university exam is een as a key of finding a proper job in Turkey. Therefore, thousands of young people enter this exam every year. Unfortunately, it's not a end of every hing; on the contrary, beginning of evrything. Generally, teenagers have to live their home in order to have a good education at university in a different city. So, they choose a dormitory to live as a best place for them. However, in my opinion, living in a dormitory has lots of disadvantage.

First disadvantage of living in a dormitory is crowded rooms. You have to share your room with other people who you don't know. That brings lots of problems. For instance, your roommate can play his /her guitar or sing songs loudly when you are trying to sleep. At that time, you really want to be alone as in your room at home!

Beside living in a crowded room, your studying hours is probably .nterrupted by others. Because, you are not the only person who lives there. For example, in the studying room anyone else can speak loudly to memorise his/her homework. Moreover, the following morning you can enter an important exam without any working. It may cause lots of unwanted events.

The final disadvantage of living in a dorm is uncomfortable life conditions. Dormitories are not designed like our beautiful houses. You may feel yourself in an hospital or in a jail because of their out-look. You can't make many things individually in these places. For example, you can't be alone while you are working, sleeping, having bath, eating and etc...

In short, a dormitory is not proper place for the university students. It brings lots of problems besides their daily life in university. Moreover, it's disadvantages can effect students's success at their department. So, finding a different place to live is a best idea for the students.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

36

The Dificulties of Staying in Dormitory

For some students; spending their university lifes I dormitories is the worst thing in the world. Because when you are a university student, you want to be free, have your own room and meet your friends in your own place. But in a dormitory you have to do things in time, share a room with four or six people and also you are never allowed to invite your friends from outside the dormitory. In addition to these; you just have a bed and if you compare a university dormitory with a special one or a home, you can recognise.some important details abot these difficulties.

First of all; it is not healthy sharing a room with many people. Because you can not sleep whenever you want, there will be always some of them chatting, listening some music and things like that. Then in some places, hundreds of students use the same bath, same toilets and maybe the same plumber. These kind of problems cause many illnesses.

On the other hand; living in dormitories costs moe money than a special dormitory or home. Whenever you want to go somewhere you have to pay for transport and if it is a busy day for you, it costs more than you guess. If you don't want to eat in your dormitory all the time because of the taste of food, it means you need extra money for each meal. Futhermore; it is not cheap as people think.

Moreover; in order to think about health or money, the worst thing is feeling yourself alone. Because your family is not there your friends can not come in to your dormitory and you have to live with people you don't know. So it looks like just a prison.

In short, students generally don't think that dormitories are not the best place to stay. Because they are full of problems and they never help students to achieve some things.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

37

Best Teachers

People are the creatures that always open to learn somethings during their lifes. The more the new informations they get, the easier the life becomes for them. A baby makes his head full with the behaviors he has seen or stories he has listened from his family, neighbours, friends etc. He writes the same informations that he chooses from his enviroment to his brain, and don't forget them. He uses these even he becomes a adult. In my opinion children get most of their knowledge in their families; therefore, their parents are not only their parents but also their best teachers.

The first reason of my thinking in this way is, parents are the first people we meet, and we never runaway from. They are always wih us all our lifes and they are the only people who want the best for us. We can ask them whatever we want to learn; moreover, we are sure that they wii teach us the more correctly and suficiently.

Next nobody else, except parents, make us big pressure. In school teachers frightened us with marks but we know bad marks can be turned to good marks. In home there is osmething different, you don't have any mark system. Instead of marks you have punishments, eyes which follow your all behaviours and so on... It is the worst thing to have to live with parents who realize your bad behaviours and the parents with angry eyes. In this situation you want to change yourself immediately, and learn how to be a good person.

The third reason parents know us better than the others. They guess our reactions and they know how to behave us. They knowour personalities and our system of learning. They are the best observers to give us the informations of our whole lifes.

In brief, the people learn the most important informations from the people who they meet first in their lives and people who they spent a lot of time together in their childhood; their parents. They are our best teachers in our lifes.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

38

LIVING IN A DORMITORY

Accomodation is a big problem for university students. Dormitories are offered as a solution generally, but it is not a good solution. On the contrary it's it's a bigger problem to accommodate in a dormitory instead of being solution. In a dormitory you'll encounter many problems.

First, living in a dormitory limits your freedom. You have to obey more rules than you do at home. You have to tidy your room as they want. You mustn't be late for going in. According to me these are limiting freedom.

Second, you may have many health problems in a dormitory, because you live with many people. It's not an easy thing to control the clearness of a dormitory and epidemics may occur in a dormitory, once I stayed in bed for a week because of an infection.

The last thing which may cause a problem in a dormitory is time. I consider living in a dormitory as wasting time. It's difficult to make a plan as you want and apply your plan properly. You also share your time like you do many things. Especially for a university student planning their times is very important because of studying. It's the most irritating point retaining from studying. I lived in a dormitory for two terms, approximately nine months. In this amount of time I studied in studying room rarely because the studying room was very small and wildy which makes me reluctant to study. So I lost many things in nine months.

In brief, while you are trying to decrease the number of the problems in university life, living in a dormitory increasing the number of problems. Regarding these problems, dormitory is not a good place to live in for a university student.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES 39
WRITING FINAL EXAM

DORMITORY LIFE

Probably, many people lived in a dormitory during his or her education life. It is very easy to say " Oh, it is not a problem, only I need is a bed to sleep." But the reality is that it is very hard to live in dormitories for several reasons.

First of all, the standart of living in a dormitory is very low. It sems good from outside, but it's very miserable from inside. Most of the students waste time in the garden and they only go to dormitory when need to sleep. The way they use the dormitory generally makes them lazy and because of this students spend most of their time in café's or pubs.

Secondly, the place that you have to live is very boring and dirty. Normally you can see instincts both in your room and bed. Always you have to sleep with them, also with your-generally 3 or 5- roommates. Therefore, because of bad roommates, lack of sleep may cause you missing lessons or passing you bad mid terms.

Admittedly, after all this drawbacks you need bathing and clean clothes. In my opinion, the worst part of living in a dormitory is that. Bathing with strangers may seem very odd to you but it's very usual in dormitory. You may think that these things only happen in movies but this is the reality of dormitories.

To sum up, being a student in Turkiyr has many difficulties. Dormitories are only one of them. Maybe private dormitories have better conditions but neither universities nor government does anything to provide better conditions for normal ones. If you don't have rich parents or a good status you will probably test it by yourself in your university life !!!

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

40

Human or Machine

Twentieth century is atime that people became lonely and lonelier. It has lots of causes. Industry revolution, developments in architecture, transportation and communication can be given as examples. Developments in communication is one of the most important of their all. Invention and spreading of television is the milestone of destruction of communication among friends and families.

Before the television people used to spend more time with their families, because they didn't have many other choices. It is very well known that, the tightly knit structures of families and friendships have faded away.

Furthermore, people faced new probblems with becoming lonely. For example, asocialisation and purchasing very much. The rates indicate that the more we watch the lonelier we became and the lonelier we became the more we consume.

It is now accepted that the people who watches TV very much, loses their contact with their nearby enviroment. And also they desire more than they have. They want families and friends just like they have seen on the screen. Because black can be shown pink on TV.

To sum up, we must quit watching that magic, silver screen and see what or who we have. It is not too late to get the things which were stolen by TV culture back. That will be the day we find our friends and familes.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

41

BENEFITS OF DORMITORIES

When students go to a new city for university, there are lots of changes occur to them. For example, they should separate from their families, friends and home. At that time, they need a new place to live and dormitories help to students. I think a dormitory is the best place for university students to live.

First of all, the students should study hard for their exams and look after themselves. For example, they should pay attention to eating, because if they didn't eat in a good way, they could get ill easily. However, the dormitories give healthy meals to the students.

Secondly, if a person think that the students could live in a house, a new problem of cleaning the houses will occur, but in the dormitories the students don't have to clean their rooms, because government does it before them.

Furthermore, the students need friends, but if they live in a house, they could only have one or two more friends, but in the dormitories, there are lots of students and if a student wasn't in a good friendship with another, he(or she) could change his (or her) friend.

In short, ther are several advantages of living in the dormitories as having prepared-meals, cleaned-rooms and lots of friends. In my opinion, a dormitory is the best place for university students to live.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

42

THE ADVANTAGES OF DORMITORY LIFE

A dormitory is a place that students stay together. In every dormitory, different number of students stays in a room. Even you can stay on your own in a room in private dormitoires. A dormitory is the best place for university students to live for several reasons.

Firstly, disciplineis an important foctor in dormitories. There are a lot of rules that you should obey. If the rules weren't, the students would do whatever they want, so this would cause a conflict. Obeying rules teaches a lot to students. In the future their lives become tidier. Also, it helps them to be succesful in their lives.

Secondly, staying in dormitory is easier than sharing a flat with friends. You don't have to prepare your food, cook, do washing, pay bills (water, electric, telephone ...etc.). the only thing that you should do in dormitory is studying. These things don't take your time. ,

Furthermore, you can have lots of friends by staying in dormitory. Meeting many people is relevant for the future life. It helps you to get a general idea about people. Also, it is good during dormitory life. You never get bored. You can spend your time with them. They help you in your bad days. You can share both your happiness and your sadness together.

In short, staying in dormitory is the best choice for university students. It gives us a lot of possibilities. In my opinion they should prefer dormitory life for these reasons.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

43

DORMITORIES JUST LIKE HOMES

Dormitories have been more attractive for especially university students in recent years. As students' wishes increase day by day, managers put money on that business willingly. In my opinion, a dormitory is the best place for university students to live for several reasons.

First, dormitories are places easy to live. In dormitories, you don't have to think about How I can pay my bills every month. Some dormitories also prepare breakfast and dinner so that you can't worry about it. For instance, in my dormitory, there are some workers in order to prepare meal. They have also responsibilities for cleaning the rooms, even preparing tea that is always ready to drink.

Second, you may have a chance to meet lots of people. You can learn to share the things you have. Because almost every people have a different character, you may learn to know people and How to behave in crowd.

The third and most important reaso is that workers' sincere behaviours and managers' behaviours showing that students comfort is more important than money make students feel just as they were living at home. Like my dormitory, it's good to know that there are some people who are ready to help and behave just a they were the members of your family.

In short, if you think all the possibilities that you face, dormitories are good places to live. Instead of living at home with your friends, you can prefer to live in dormitory for living in comfort and feeling as you were at home.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

44

Living in Dormitory

Ther are a lot of students who attend university after passing the university exam. They face some problems. One of these problems is where they can live. Some of them prefer to live in a dormitory while others prefer to stay in a house. Living in a dormitory has severel advantages.

Firstly, students can meet a lot of people who come from different countries and have different cultures. They can learn more information about people. Owing to this information, they can choose their friends easily. When I begun to live in dormitory I didn't know many students who are now my friends.

Secondly, students can learn living alone. Before coming to university, many students live with their families, and families always help them. After living home, they will be alone and they will have to live without their families. Because of this, they will get knowledge about life and face difficulties. Then, these difficulties will make students experience.

The last advantage is freedom. Some students think that they aren't free when they are living with their families. In a dormitory, they will have freedom and they can do whatever they want and go out whenever they want. Besides these, students can stay in their friends' houses as much as they want. Nobody can interfere students about this.

To sum up, there are a lot of uses of living in a dormitory. Students can learn more information about people, living alone and difficulties of life. In addition to these, they will be free so, students should prefer to live in dormitory.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES 45
WRITING FINAL EXAM

LİFE İN A DORMİTARY

A dormitory is the second home of a student. It is not as comfortable as home, but it is a good place to live. There are a lot of advantages of living in a dormitory so it's the best place for university student to live.

The first advantage of living in a dormitory is freedom. Student who stays in a dormitory has more freedom than a student who lives with his family in a home. Moreover, a dormitory student learns how to use money. It is very useful for his future life.

Another advantage of living in a dormitory is its condition for studying. Ther are a lot of rooms for students who want to study. You can study when you want. They are open twenty-four hours. You can also study with your friends. They are always ready to help you. In addition, there is no television in the dormitory, so TV can't effect you.

The final advantage of living in a dormitory is friendship. You can have a lot of friends in dormitory. You can do different activities with them because your friends have different hobbies. Some of them may like sports. Some of them may like music. As a result, you can do different activities with different friends. In addition, they are always ready to help you. They share your problems. They are your second family.

Finally, Living in a dormitory has a lot of advantages so it is wonderful to live in there. In short, a dormitory is the second home of a student.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

46

A dormitory is the best place for university students to live.

Life is difficult for the students in Turkey. Accomodation is basic problem for the students who have to live their own home. Thus, government creates opportunity to stay in dormitories for the students. There are abundant dormitories in our country. Furthermore it has a lot od advantages to stay in a dormitory for several reason.

First of all, families want their children to stay in a secure place; in my opinion state's dormitories are the best places about security in all of other private dormitories. I stayed Yunus Emre dormitory and I saw that they give importance security of all students. For example, they take students' significiant every evening.

The second one is having good friends and abilities at choosing friends. I think, if I didn't stay in the dormitory, I wouldn't meet a lot of people. Now, I can know a person when I looked at him or her eyes. I have got a lot of good friends;Although, we know each other for a year.

Another reason is economical problems. A dormitory is more reasonable than a home. If you don't want to make food, you can eat in a canten. Moreover, there are a lot of facilities for doing sports. Therefore, you can use opportunities without paying money.

In brief, I believe in that a dormitory is the best place for university students to live. The university students should stay in a dormitory to know the life better. It is difficult to get used to the way of the life in a dormitory but it is wonderfull after getting used to it. You feel that your self-confident increases.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

47

Dormitory life is very important for university students. Although living in dormitory has bed sides, it forms and protects students charactheristics. And also there are a lot of good sides about living in dormitory.

First of all, one of the best thing which dormitory gives us is learning about common life. Besides, this dormitory period makes us a real mature in time. Sharing good, meal, even a bed, drives student learn the real life behind the veil.

Then, friendship is very important in dormitory. If you become acquinted with others, you can solve any kind of difficulty. Morever, you hardly feel blue, because in every lack of something- missing family, examinations- someone will be always beside you.

After that, you can easily study your lessons. When the examinations come nearer, everybody goes to studying room. So, this causes you to study regularly. And also this brings you great effort. When you see a horde in studying room, you fear about not to find a seat.

In brief, dormitory life is necessary for students, especially for university students. All parents must agree with me, because the reality of dormitories, is undeniably obvious. In an other word, dormitories are a part of life. Only thing you have to do is sending your children to them. Leave the other things to dormitories.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES **48**
WRITING FINAL EXAM

Living in a dormitory

A domitory is the best place to live. Dormitoties are usefull for new students who come from different cities. Every student wants to live in a comfortable environment. But, it is so expensive for alone students. Because of such reasons, most students prefer to live in a dormitory which has some advantages.

The first advantage is saving money. Dormitory fee is reasonable for alone students. They may easily save and collect extra money for their requirements.

The second is friendship. In a dormitory, a huge number of students stay. Students may find a lot of friends and make good friendships.

The last advantage is sharing help. That living alone far from family is difficult for a student. Students can help each other for some daily requirements, such as washing, cooking, and ironing.

In conclusion, dormitories have several advantages for alone students. Living in a dormitory may be a good experience and a perfect preparation for real life.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

49

A NEW WORLD

A dormitory is the best place for university students. Living in a home more difficult than living in a dormitory. Because if you live in a home you are responsible for many things; for instance, cleaning, paying bills, cooking etc. On the other hand living in a dormitory is more enjoyable and easier than living in a house. Living in a dormitory has many benefits for a university students.

First, a university student needs money and he gets that money from his family. If you live in a dormitory, belongs to government, you give only 6000000 TL. for each month. However, government pays back 4000000 TL you for your breakfast. So you don't need much money. Also you can eat your launch and dinner at dormitory too.

In addition, if you give not much money to live and stay somewhere, you can save your money and buy what you want. You can buy more book or you can go to cinema more regularly than you did before. You will have a chance for shopping, if you like to do.

Furthermore, dormitory is the most enjoyable place for university students. You can have many friends from diffrent country. If you like meeting new people or if you like to know different people, you will have chance. You will stay with 6 people in the same room, so it is a good chance to develop your communication.

Consequently, living in a dormitory has many benefits for a university student. It is cheap, enjoyable and diffrent. It is a new world which full of diffrent kinds of people. So, a university student should stay in a dormitory for a while and should see the new world.

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

50

LIVING IN A DORMITORY

There are lots of students who study at the universities which are not in their home towns all over the world. Some of these students live in dormitories, and some of them live in flats. In my opinion, a dormitory is not the best place for university students to live for some reasons.

It is difficult to live in a dormitory because there lots of students in a dormitory. A student has to share her room with some other students. Because of the crowd, a person can't sleep and relax very well. For example, students talk, laugh and do something very thoughtlessly at every time of the day. Because of this noisy atmosphere, it is almost impossible to study well. I think that crowd is the most serious problem that a university student has in a dormitory.

A dormitory may be very very dirty. In my opinion, dirt is a big problem in a dormitory. Because dirty places can affect someone's health very badly. According to me cleaners of the dormitories are not doing their jobs very well. Places that lots of people live in should be cleaned more carefully, but they behave as if they are unaware of this reality.

There aren't electronic devices that a person needs everytime of his life in a dormitory. For example, there is not a refrigerator. Because of that problem students can't keep their food fresh, so they have to eat something in restaurants. There is not a television which is an important mean of communication.

To sum up, a dormitory is not a good place for a university student to live for several reasons. A university student should live in better places in order to live a comfortable and a good life.

Appendix C

FINAL EXAM GRADING STANDARDS
(Holistic-analytic criterion, adapted from Appendix D)

TASK ACHIEVEMENT

40-38	The content is relevant with the topic and there is no irrelevant content. The main idea in each paragraph is supported by clear and appropriate evidence/examples.
30-	The content is relevant with the topic, but there may be some irrelevant information. Some main ideas are supported by appropriate evidence/examples.
20-	Most of the content is not relevant to the topic. Most of the main ideas are not supported by appropriate evidence/examples.
10-	The content is not relevant. None of the ideas presented are supported by appropriate evidence/examples.

ESSAY ORGANISATION

40-38	The paragraphs of the essay are clearly and logically organised. The text is organised into a clear introduction, body and conclusion.
30-	The paragraphs of the essay are not logically organised. Some part of the introduction, body and/or conclusion is incomplete.
20-	The paragraphs of the essay are not organised. One of the introduction, body and/or conclusion paragraphs is missing.
10-	There is no distinct introduction, body and conclusion.

ACCURACY OF WRITTEN SKILLS

20-18	Few and minor grammar errors. The use of vocabulary is clear and effective with few inaccuracies.
15-	More grammatical errors in general, a few major errors, which do not interfere with understanding. The use of vocabulary is clear but not well developed/varied, still with few inaccuracies.
10-	The number and quality of the errors make understanding difficult. The use of inaccurate vocabulary frequently confuses the reader.
5-	Grammatical errors are so frequent that some portions of the essay are incomprehensible. The use of inaccurate vocabulary makes some portions of the essay incomprehensible.

PENALTY : -10 FOR NOT ANSWERING THE QUESTION

APPENDIX D

(Original Holistic Criterion)

Anadolu University, Eskişehir
English for Nonnative Speakers Preparatory Program
Final Exam Grading Standards

5 The essay demonstrates clear competency though there may be a few errors

The content is relevant with the topic and there is no irrelevant content.
The main idea in each paragraph is supported by clear and appropriate evidence/examples.
The paragraphs of the essay are clearly and logically organised.
The text is organised into a clear introduction, body and conclusion.
Few and minor grammar errors.
The use of vocabulary is clear and effective with few inaccuracies.

4 The essay demonstrates competency

Each paragraph has a main idea.
The content is relevant to the topic, but there may be some irrelevant information.
The main idea in each paragraph is supported by appropriate evidence/examples.
The text is organised into introduction, body and conclusion.
The paragraphs of the essay are logically organised.
More grammatical errors in general, a few major errors, which do not interfere with understanding.
The use of vocabulary is clear but not well developed/varied, still with few inaccuracies.

3 The essay demonstrates minimal competency

Some paragraphs do not have a main idea.
The content is generally relevant to the topic, but not all parts of the essay are.
Some main ideas are supported by appropriate evidence/examples.

Some of the introduction, body and conclusion may be incomplete.
The paragraphs of the essay are not logically organised.
More major errors that sometimes interfere with understanding.
The use of inaccurate vocabulary sometimes confuses the reader.

2 The essay shows a developing competency

Most of the paragraphs do not have a main idea.
Most of the content is not relevant to the topic.
Most of the main ideas are not supported by appropriate evidence/examples.
Some part of the introduction, body and/or conclusion is missing/incomplete.
The paragraphs of the essay are not organised.
The number and quality of the errors make understanding difficult.
The use of inaccurate vocabulary frequently confuses the reader.

1 The essay demonstrates incompetency in writing

None of the paragraphs has a main idea.
The content is not relevant.
None of the ideas presented are supported by appropriate evidence/examples.
There is no distinct introduction, body and conclusion.
The paragraphs of the essay are not coherent.
Grammatical errors are so frequent that some portions of the essay are incomprehensible.
The use of inaccurate vocabulary makes some portions of the essay incomprehensible.

Appendix E

(Grading sheet for holistic-analytic criterion)

Grader's Name:

Paper No.	Task Ach.	Essay Org.	Acc.of Wr.Sk.	Penalty - 10 pts	Total	**Paper No.**	Task Ach.	Essay Org.	Acc.of Wr.Sk.	Penalty - 10 pts	Total
1						**26**					
2						**27**					
3						**28**					
4						**29**					
5						**30**					
6						**31**					
7						**32**					
8						**33**					
9						**34**					
10						**35**					
11						**36**					
12						**37**					
13						**38**					
14						**39**					
15						**40**					
16						**41**					
17						**42**					
18						**43**					
19						**44**					
20						**45**					
21						**46**					
22						**47**					
23						**48**					
24						**49**					
25						**50**					

APPENDIX F

Suggestions for the Design of a Final Grading Criterion

Dear colleague,

This questionnaire is prepared in order to obtain your suggestions to design an alternative grading criterion for the final writing exams of our school. The aimed criterion will be analytic in which a separate score will be awarded for each quality of the written work. Therefore, your ideas about the following prompts will enlighten the crucial details, the evaluative parts, the descriptors, bands and their weightings to form the necessary parts of the overall grading criterion. Please read each item carefully, check the appropriate box according to your ideas about grading writing papers and follow the instructions to give the necessary details. While filling the percentage gaps please remember that the criterion will evaluate the writing quality on 100 % as the total success; thus, your percentages should be 100 as a total at the end of counting your prompts.

1. Grading **organisation** is essential for the final exams. **Yes** ☐ **No** ☐

(If your answer to question **1 is "yes"** go to question **2**, if it is **"no"** go to question **4**.)

2. ..... % weighting is enough for grading organisation. (Please write your appropriate amount)

3. Organisation part should include ..... evaluative bands. (Please state the number of bands.)

4. Grading **content** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **4 is "yes"** go to question **5**, if it is **"no"** go to question 7.)

5. ..... % weighting is enough for grading content. (Please write your appropriate amount)

6. Content part should include ..... evaluative bands. (Please state the number of bands.)

7. Grading **language** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question 7 **is "yes"** go to question **8**, if it is **"no"** go to question **10**.)

8. ..... % weighting is enough for grading language. (Please write your appropriate amount)

9. Language part should include ..... evaluative bands. (Please state the number of bands.)

10. Grading **social awareness** * is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **10 is "yes"** go to question **11**, if it is **"no"** go to question **13**.)

11. ..... % weighting is enough for grading social awareness. (Please write your appropriate amount)

12. Social awareness part should include ..... evaluative bands.(Please state the number of bands.)

13. Grading **vocabulary** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **13 is "yes"** go to question **14**, if it is **"no"** go to question **16**.)

14. ..... % weighting is enough for grading vocabulary. (Please write your appropriate amount)

15. Vocabulary part should include ..... evaluative bands. (Please state the number of bands.)

16. Grading the issue **"appeal to the readers"**** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **16 is "yes"** go to question **17**, if it is **"no"** go to question **19**.)

17. ..... % weighting is enough for grading **"appeal to the readers"**. (Please write your appropriate amount)

18. **"Appeal to the readers"** part should include ..... evaluative bands. (Please state the number of bands.)

19. Grading **title** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **19 is "yes"** go to question **20**, if it is **"no"** go to question **21**.)

20. ..... % weighting is enough for grading title. (Please write your appropriate amount)

21. Grading **spelling** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **21 is "yes"** go to question **22**, if it is **"no"** go to question **23**.)

22. ..... % weighting is enough for grading spelling. (Please write your appropriate amount)

23. Grading **punctuation and capitalisation** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **23 is "yes"** go to question **24**, if it is **"no"** go to question **25**.)

24. . ..... % weighting is enough for grading punctuation and capitalisation. (Please write your appropriate amount)

25. Grading **neatness** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **25 is "yes"** go to question **26**, if it is **"no"** go to question **27**.)

26. ..... % weighting is enough for grading neatness. (Please write your appropriate amount)

27. Grading **handwriting***** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **27 is "yes"** go to question **28**, if it is **"no"** go to question **29**.)

28. ..... % weighting is enough for grading handwriting. (Please write your appropriate amount)

29. Grading **transitions** is essential for final exams. **Yes** ☐ **No** ☐

(If your answer to question **29 is "yes"** go to question **30**, if it is **"no"** go to question **31**.)

30. ..... % weighting is enough for grading transitions.(Please write your appropriate amount)

31. A **penalty** should be given if an unrelated topic was written. **Yes** ☐ **No** ☐

(If your answer to question **31 is "yes"** go to question **32**, if it is **"no"** go to question **33**.)

32. ..... % weighting is enough for the penalty.

33. A **penalty** should be given if there is too much or too little written work. **Yes** ☐ **No** ☐

(If your answer to question **33 is "yes"** go to question **34**, if it is **"no"** go to question **35**.)

34. ..... % weighting is enough for the penalty.

The following is the checklist of your score weighting. First, in relation with your previous suggestions, check the ones you decided to include in the criterion, next write your own percents and check them in order to obtain 100 % as a total :

a. ☐ Organisation ..... %

b. ☐ Content ..... %

c. ☐ Language ..... %

d. ☐ Social awareness ..... %

e. ☐ Vocabulary ..... %

f. ☐ Appeal to the readers ..... %

g. ☐ Title ..... %

h. ☐ Spelling ..... %

i. ☐ Punctuation and capitalisation ..... %

j. ☐ Neatness ..... %

k. ☐ Handwriting ..... %

l. ☐ Transitions ..... %

Total ..... 100 %

Further Suggestions : .....

.....

.....

.....

.....

.....

Thanks for your contribution and valuable co-operation!

* **Social awareness** means to be aware of what is going on and what has been going on in the world or around so far. The quality of social awareness in writing papers is the ability of expressing this kind of awareness.

** **Appeal to the readers** represents the quality of attracting or arousing the interest of readers to a written work.

*** **Handwriting** is the type of writing typical of a person done with the hand. The quality of handwriting is related with its clarity and ease of reading.

Appendix G

Sample Paper 1 (Taken from the June 2000 Final Exams)

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

DORMITORY AND STUDENTS

Dormitories build for students, because a lot of students stay in dormitories, in this reason there are a lot of reason for staying in dormitory, so a dormitory is the most important and the best place for university students to live.

One group of reason which is more important for poor students is cheap dormitories. These dormitories have got of goverment ,and they built for university students. Also, these dormtories are more cheap, so a lot of poor students can stay in these dormitories, because poor students usually don't stay in houses for example, recently students pay 6.000.000 TL. to Government dormitories for a month ,but a student pays 100.000.000 TL to his or her landlord for a month. In this reason Government dormitories are more requirment for poor students.

Another gorup of reason is the most important for friendships, because a student has a lot of roommates in dormitory, so a student has alot of friendships. In addition, a dormitory helps for describing people personality. for example one people stays in cumhuriyet girl dormitory and that person has 12 roommates. This example is important that's person, because person both was lonely and hadn't got any friend, therefore that person had a lot of friends in dormitory.for these reasons dormitories are the beat place for friendships.

The final group of reason is more safely for students to live, and students learn to be responsable in dormitories, espacialy Goverment dormitories are safely for students. For example two men and one woman wait for students in a dormitory everynight, so students don't frghten at night. Another example, students must be at 11 o'clock in a dormitory. Students learn a lot of thing in dormitories.

In short, dormitories are everything for poor students, friendships and safely life, I think each students must stay in a dormitory.

Sample Paper 2

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

Advantages

The dormitory was building for students. The student to advantage dormitory live for example dormitories in the campus doesn't transportation problem. The student doesn't give for money transportation. That help the students.

A person's behaviours, develops in dormitory, because the student meets much new people. The student learn new information.This informations develop the student's intelligent.

The student doesn't for television much studies lesson. When a student was while working lesson helps friends. The student to advantage lesson. The other doesn't advantage

We opinion students stay dormitor much luck, because that students has got up advantage. That expect is the best place a dormitory for university students live

Sample Paper 3

ANADOLU UNIVERSITY
SCHOOL OF FOREIGN LANGUAGES
WRITING FINAL EXAM

PARENTS

The most important thing in someone's life is educating themselves, but there are so many styles of education some of which are bad and some are good. Since its birth, everybody faces these styles. Moreover, so many people think that real life is always the best teacher for them. However, in my opinion parents are the best teachers for several reasons.

The first and most important reason is that nobody else except that a family can give a person his own character. In addition, everybody needs a few years just after their bith to acquire their own characters, because these years are the times that babies learn themselves. For example, when each of a twins lives in different lifestyles, they show different characteristic behaviours ,although they have almost the same characters just after the time of their birth.

The second reason is that parents always educate their children in a good way. This means that parents never teach them having bad habits ,saying bad words, behaving in a way to other people, or so on. Also, no only they give them good habits, they protect them from the bad effects of the environment as well. For example, if a child learns swearing, his parents do whatever they can in order to stop this bad behaviour.

The final reason is that, because the parents know their children better than anyone else, they know the best solution for their problems. When the children have a problem, others can't see the details of events. Just because their parents know their psychological situations, they can help them to overcome with these problems. Also they can teach them how to see the situations in a broader sense.

In brief, although we can mention these specific situations, we should always learn our parents' opinions about us during our whole lives, because noone can understand us better then they can. They will always keep our priorities in front so that they will always teach us the best, therefore they are the best teachers.

Appendix H

GRADING CRITERION for FINAL EXAMS

(New analytic criterion)

ORGANISATION

25-19 Pts A clearly stated thesis answering the question, all the body paragraphs (3) are related to the thesis, true use of a funnel introduction and a conclusion for the thesis.

18-12 Pts A thesis answering the question, of the 3 body paragraphs one of them is missing or not related to the thesis, minimal use of introduction and/or conclusion.

11-5 Pts A thesis partly answering the question, of the 3 body paragraphs two of them are missing or not related to the thesis, problems in the lay out of conclusion or introduction.

4-0 Pts A thesis that is unclear or not answering the question, all the body paragraphs are missing or not related, either introduction or conclusion is missing.

CONTENT

25-19 Pts Completely unified and coherent paragraphs, totally relevant with the given topic, focused on the task, appropriate support and logical analysis, no lapses in development.

18-12 Pts Mostly unified and coherent paragraphs, mostly relevant with the given topic, focused on the task, some attempt at logical analysis and support of the proposal, minor lapses in development.

11- 5 Pts Some coherence problems in one or more paragraphs, some irrelevant information, inadequate support and/or problems in logical analysis, major lapses in development.

4- 0 Pts Coherence is lacking, not focused on the task, most of the written work is not logical and irrelevant with the given topic.

LANGUAGE

15-12 Pts Some errors, which generally do not interfere in meaning. Effective control of sentence structure, verb formation, agreement and tenses. Effective control of articles and pronouns.

11- 8 Pts Errors which sometimes interfere in meaning. Some control of sentence structure, verbs, formation, agreement and tenses. Some control of articles and pronouns.

7- 4 Pts Frequent errors that often interfere with understanding. Problems in sentence structure, verbs, formation, agreement and tenses. Inadequate control of articles and pronouns.

3- 0 Pts The paper is full of major and repeated errors. Many unclear sentences. Little or no control of sentence structure and verbs.

VOCABULARY

10- 8 Pts Variety and accuracy in word choice, correct word formation.
7- 4 Pts Reasonable use of vocabulary, some control of word formation.
3- 0 Pts Noticeably simple, limited and misused vocabulary with many problems in word formation.

TRANSITIONS

10-8 Pts Effective use of transitions and other coherence devices like: also, therefore, for instance, indeed, nevertheless, although, furthermore, finally...
7-4 Pts Average use of transitions and other coherence devices which may be missing in some parts of the essay.
3-0 Pts Minimal or no use of transitions, the used ones may be incorrect.

TITLE ..... **5 pts**
PUNCTUATION-CAPITALISATION ..... **5-0 pts**
SPELLING ..... **5-0 pts**

PENALTY: A totally unrelated topic ..... -30 pts

Appendix I (Grading sheet for analytic criterion)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Organisation																									
Content																									
Language																									
Vocabulary																									
Transitions																									
Title																									
Punc- Cap.																									
Spelling																									
Penalty –30																									
TOTAL																									

Continued on Page 173

	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Organisation																									
Content																									
Language																									
Vocabulary																									
Transitions																									
Title																									
Punc- Cap.																									
Spelling																									
Penalty –30																									
TOTAL																									

Appendix J

Evaluation of the Final Grading Criterion

Dear colleague,

This questionnaire is prepared in order to obtain your suggestions to check the usability of the alternative criterion for the final writing exams of our school. Your ideas about the following prompts will enlighten the crucial details to focus in further studies; thus, they are quite valuable. You are supposed to read each item carefully and check the one that best suits to your opinions about the new grading standards regarding its weak and strong points through the grading process. The following represent the meanings of the symbols used in the table:

SD: Strongly Disagree **D:** Disagree **U:** Undecided **A:** Agree **SA:** Strongly Agree

Thanks for your contributions!

	SD	D	U	A	SA
1. The bands of the ***organisation part*** are successfully planned.					
2. The first band of the organisation part is reasonable and thorough.					
3. The second band is clear and compatible with the first band.					
4. The third band is clear and compatible with the previous bands.					
5. The fourth band is clear and compatible with the other 3 bands.					
6. The bands of the ***content part*** are sufficient and easy to follow.					
7. The first band of content is clear and quite fulfilling.					
8. The second band is clear and compatible with the first band.					
9. The third band is clear and compatible with the previous bands.					
10. The fourth band is clear and compatible with the other 3 bands.					
11. The bands of the ***language part*** are extensive and comprehensive.					
12. The first band of the language part is logical and comprehensible.					
13. The second band is reasonable and compatible with the first band.					
14. The third band is clear and compatible with the previous bands.					
15. The last band is apparent and compatible with the previous bands.					
16. The bands of the ***vocabulary part*** are appropriate and clear.					
17. The point ranks of the vocabulary part are properly designed.					
18. The bands of the **transition part** are reasonable and clear.					
19. The point ranks of the transition part are properly designed.					
20. To mark the ***Title*** is something necessary for grading a paper.					
21. ***Punctuation and capitalisation*** are relatively important to grade.					
22. Grading ***spelling*** is essential for writing tests.					
23. The ***penalty*** and its quantity is appropriate for such an exam.					
24. The criterion in general is quite handy and easy to follow.					
25. The grading time for each paper is not too long.					
26. The descriptors and their contents are suitable to our programme.					

Further Suggestions :.....

.....

.....

Appendix K

Table 6: The grades given with the 1st Instrument in three grading sessions by ten raters.

INSTRUMENT 1																													
Paper	Grader 1			Grader 2			Grader 3			Grader 4			Grader 5			Grader 6			Grader 7			Grader 8			Grader 9			Grader 10		
	Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	55	60	65	35	55	55	55	65	60	35	55	35	68	45	45	45	55	55	45	60	55	65	55	50	25	45	25	60	73	55
2	65	55	60	60	70	70	55	60	60	60	65	55	40	65	55	65	55	50	45	65	66	88	65	75	75	68	70	78	91	90
3	25	25	25	25	35	30	45	35	50	35	55	55	65	35	35	85	45	70	55	45	55	25	35	30	25	45	45	70	50	45
4	50	50	50	70	78	70	65	70	70	60	78	70	68	60	65	70	73	70	50	70	71	73	65	70	86	50	55	75	83	83
5	45	50	40	25	35	35	45	30	40	30	25	25	50	35	35	65	85	65	35	45	35	55	25	57	45	25	45	65	55	60
6	50	50	60	60	40	55	60	35	50	60	70	60	53	55	55	75	45	50	65	60	65	50	70	70	45	60	50	83	75	78
7	45	45	40	35	25	30	45	30	30	35	25	30	55	25	25	55	35	55	45	25	25	25	30	30	75	25	25	40	50	50
8	70	70	60	50	70	70	65	50	60	75	70	70	35	60	45	65	58	60	45	70	73	70	55	60	78	70	70	83	75	83
9	40	45	45	60	50	60	60	50	60	50	60	60	60	50	55	50	58	50	55	68	68	38	70	70	58	68	60	63	86	65
10	45	45	35	25	35	35	55	25	30	25	55	65	35	25	35	45	25	25	35	25	35	60	25	30	25	35	25	40	45	35
11	45	45	35	35	45	45	45	50	50	35	45	45	45	25	25	55	45	25	75	45	50	45	25	25	25	35	35	75	60	45
12	45	45	35	35	55	45	45	30	40	25	55	55	25	35	25	55	35	45	65	35	35	45	25	25	45	35	45	60	45	60
13	70	70	70	60	73	65	70	70	70	75	70	70	78	55	55	78	70	55	50	86	85	88	70	80	86	83	80	80	91	91
14	45	45	45	25	35	35	45	50	40	25	60	55	35	25	35	35	45	75	55	30	50	45	50	40	30	25	25	55	65	55
15	45	70	45	60	55	55	50	55	50	40	55	55	50	35	35	70	55	50	45	60	65	65	60	65	35	50	50	65	70	65
16	75	82	75	70	78	70	70	75	75	55	70	70	83	55	55	58	70	60	80	91	91	86	75	78	88	83	80	91	91	95
17	45	65	60	60	50	55	60	65	70	60	60	60	55	45	50	75	68	60	65	65	65	60	70	65	73	86	83	91	83	90
18	75	78	75	70	68	83	75	75	70	70	78	70	83	60	65	70	85	75	78	91	91	78	75	76	94	86	88	83	91	95
19	75	75	78	65	75	71	65	60	60	75	83	75	94	75	80	93	75	85	71	91	91	75	70	73	91	96	96	91	83	85
20	70	70	75	75	83	75	83	60	83	70	78	75	86	75	70	85	85	80	76	96	93	98	75	78	96	88	86	75	91	90

Continued on page 176

21	75	83	60	83	91	85	45	75	58	65	78	75	83	70	70	85	75	85	78	83	84	65	83	81	98	86	97	91	83	95
22	75	83	70	75	65	65	65	50	65	78	78	80	78	65	50	75	55	70	70	83	85	85	65	70	88	68	85	73	91	75
23	70	75	75	83	80	83	95	75	75	78	86	78	83	70	70	83	70	70	75	91	88	81	70	83	88	91	90	50	91	90
24	75	83	75	60	83	70	70	83	75	75	96	83	55	60	65	88	85	80	78	60	78	85	75	88	98	91	93	60	83	65
25	75	65	66	50	70	65	60	65	70	60	68	60	65	60	60	83	90	80	91	86	85	40	70	45	86	75	86	98	96	95
26	83	81	75	91	88	86	75	96	75	78	73	83	75	70	70	98	70	70	90	91	94	90	73	91	78	83	80	91	93	93
27	75	75	75	60	78	60	65	75	70	75	68	68	86	60	75	68	85	78	86	83	83	70	70	70	86	75	75	75	83	75
28	45	55	50	60	60	55	55	65	60	65	73	75	68	50	50	78	78	70	80	65	75	55	60	65	48	55	55	88	75	85
29	45	40	40	50	76	65	55	65	65	75	63	75	68	55	60	78	58	70	96	65	65	50	60	55	58	65	65	83	65	65
30	65	65	50	75	70	70	65	55	65	65	83	75	65	60	60	83	75	80	85	83	85	78	70	80	58	75	73	83	75	83
31	65	65	60	50	65	55	55	60	65	65	78	73	75	65	75	86	70	70	78	83	78	65	50	60	75	75	70	75	83	81
32	60	60	60	65	70	75	70	75	78	75	88	83	94	75	78	96	93	95	91	91	93	78	70	78	96	78	95	96	83	96
33	45	45	45	60	65	60	55	60	55	65	60	60	73	60	70	88	50	65	70	65	75	55	55	55	65	68	60	65	83	60
34	91	91	91	93	94	96	91	96	98	94	96	94	98	94	94	98	100	95	100	88	98	98	65	95	100	80	80	90	98	98
35	91	91	91	91	94	96	75	88	75	86	91	88	80	94	91	93	95	90	75	98	96	96	90	98	98	96	98	83	96	83
36	83	91	75	75	75	78	65	83	83	86	94	95	86	91	85	93	65	65	91	91	91	78	70	73	91	96	93	91	83	90
37	78	75	83	86	83	88	91	88	90	75	83	83	94	83	88	98	85	75	95	94	91	96	93	94	98	88	96	94	98	93
38	75	75	81	86	86	90	83	96	88	75	91	75	94	88	94	96	88	95	91	98	93	94	83	91	100	80	100	94	98	98
39	70	75	73	76	75	78	65	91	90	94	94	90	94	76	86	93	86	90	70	94	91	68	83	80	98	90	94	94	100	91
40	65	55	65	78	86	80	90	75	83	91	91	93	68	90	88	98	90	90	78	94	93	78	85	83	75	88	78	94	98	100
41	78	75	83	75	78	80	75	91	91	83	83	85	94	96	90	74	98	90	80	70	81	75	80	80	86	86	83	83	96	94
42	86	91	83	78	86	75	75	88	86	94	83	83	96	94	90	70	91	70	98	75	93	94	86	88	94	94	93	83	96	93
43	86	91	94	83	86	90	83	88	84	87	63	75	98	80	88	78	96	95	96	85	81	75	86	80	100	96	96	83	96	95
44	94	94	83	86	91	96	91	96	96	86	94	94	98	90	88	96	88	95	100	91	96	95	90	95	78	96	97	83	96	95
45	86	78	75	63	94	68	91	83	83	91	94	93	96	88	86	96	98	90	98	94	91	68	90	88	88	88	88	83	88	83
46	91	94	83	94	93	90	65	83	83	83	83	83	96	91	93	96	91	90	90	91	91	83	86	88	94	88	91	75	91	70
47	78	78	83	86	91	86	83	75	78	83	83	80	94	88	88	86	94	90	76	90	88	70	73	70	98	86	95	83	88	88
48	78	86	80	78	68	68	95	55	78	73	75	75	98	75	75	98	86	88	90	83	83	73	70	75	78	73	75	73	80	78
49	86	78	78	94	97	94	75	83	86	65	83	83	94	83	94	98	86	90	96	75	88	75	70	70	100	88	98	78	91	75
50	94	94	94	94	98	97	83	71	70	83	94	79	98	94	96	100	100	95	98	94	91	80	88	85	100	86	86	94	96	94

Appendix L

Table 7: The grades given with the 2nd Instrument in three grading sessions by ten raters.

INSTRUMENT 2																													
Paper	Grader 1			Grader 2			Grader 3			Grader 4			Grader 5			Grader 6			Grader 7			Grader 8			Grader 9			Grader 10		
	Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	40	41	44	44	42	39	52	50	47	43	46	40	37	44	39	50	52	50	41	41	43	52	52	47	41	40	43	48	48	49
2	59	57	59	58	57	55	59	57	57	57	59	53	50	52	48	60	63	57	54	57	59	55	62	59	50	53	57	50	50	56
3	24	24	27	24	19	27	27	29	28	35	27	30	28	29	24	32	26	25	26	27	24	32	31	30	25	21	22	32	26	28
4	66	61	63	57	59	61	56	58	66	65	57	59	60	59	64	55	68	56	66	69	68	61	62	59	56	58	60	60	68	67
5	39	40	35	31	35	29	27	33	33	34	28	33	26	29	32	24	31	33	27	33	31	37	33	39	29	31	31	26	25	28
6	52	53	47	48	44	43	37	44	38	44	42	38	50	45	45	41	43	40	44	50	46	33	40	36	33	39	38	48	44	46
7	20	21	27	20	18	23	19	21	24	22	20	27	24	25	29	21	21	25	22	22	23	27	26	27	22	20	27	24	18	20
8	48	45	44	45	43	41	39	43	37	40	39	46	36	40	41	40	38	35	43	48	49	46	41	40	31	37	35	44	47	45
9	39	38	39	41	36	39	41	45	44	51	44	47	31	39	37	37	32	39	40	43	46	44	40	39	36	43	39	35	31	38
10	21	24	22	20	22	28	26	27	28	33	29	33	31	28	24	22	24	28	24	25	29	30	31	25	24	22	20	20	22	22
11	27	27	34	34	30	32	34	34	33	29	32	33	34	29	37	36	38	33	34	37	33	38	40	39	38	32	35	40	36	38
12	19	19	20	21	17	25	26	28	20	26	22	27	24	23	18	24	22	25	24	27	23	28	25	20	17	21	23	21	23	28
13	65	61	60	58	52	60	55	54	61	65	60	64	57	55	61	65	67	70	64	63	68	67	68	64	57	63	67	62	67	61
14	29	27	32	25	24	26	25	24	31	37	34	37	26	28	30	30	23	31	37	32	35	36	35	32	25	28	30	28	23	29
15	50	51	49	58	56	50	37	50	48	51	57	49	60	59	62	63	66	58	67	61	58	57	52	55	56	61	60	53	58	56
16	61	59	67	70	67	72	76	72	75	73	68	73	68	62	68	75	70	71	80	78	74	76	73	70	63	69	69	75	71	74
17	61	61	62	44	50	53	67	67	61	57	60	51	48	56	52	48	53	56	58	58	62	55	61	60	48	54	55	66	59	62
18	77	77	70	58	54	60	60	59	66	52	57	61	51	56	58	66	63	60	73	71	67	76	69	73	52	52	59	65	65	66
19	66	68	65	62	64	70	69	65	67	74	70	63	62	67	66	70	72	68	66	66	67	63	68	62	59	64	62	67	72	66
20	80	77	79	70	75	76	80	81	79	67	70	77	71	71	78	86	80	85	77	81	88	81	77	82	73	75	80	78	83	80
21	77	77	77	70	78	73	83	86	81	76	79	71	75	74	78	79	74	80	77	73	78	80	76	76	68	75	69	70	83	76
22	58	57	64	48	48	59	70	72	69	62	68	65	59	54	60	50	59	58	64	65	62	62	60	67	57	66	60	55	65	59
23	58	58	60	64	60	59	68	67	68	62	58	65	54	59	61	70	68	66	69	73	70	75	70	69	57	60	61	70	67	67
24	68	70	65	73	66	67	62	68	63	69	74	73	74	73	70	83	75	72	71	69	69	76	67	68	64	70	64	65	75	72

Continued on page 178

25	51	51	58	66	59	61	50	57	59	58	55	63	57	68	58	65	70	66	61	60	67	59	62	58	63	66	60	71	70	68
26	63	65	65	72	70	66	71	67	73	65	68	70	63	66	66	77	74	70	72	73	69	79	70	61	62	71	60	75	72	70
27	65	67	71	70	74	73	67	69	72	73	71	76	75	70	71	78	74	79	70	71	74	73	77	69	74	75	73	71	78	70
28	57	55	56	54	60	57	63	60	65	67	60	65	58	53	57	67	62	60	52	56	59	55	62	58	60	66	56	52	58	56
29	59	59	61	51	58	61	61	61	62	56	61	58	60	61	64	61	60	58	61	63	62	61	59	60	61	60	63	59	60	57
30	64	65	68	66	69	73	70	72	75	69	65	72	73	78	71	82	80	78	74	76	77	74	75	70	77	70	70	72	72	75
31	51	54	56	53	50	53	49	47	53	48	50	56	49	48	55	50	56	60	57	54	51	63	59	59	54	58	60	50	55	57
32	56	55	61	58	57	56	60	59	62	66	58	62	54	58	56	63	63	58	72	68	64	73	66	57	58	63	60	65	70	64
33	36	43	42	36	41	40	34	38	41	55	51	48	39	44	49	40	48	50	45	44	40	48	51	40	39	40	44	42	45	48
34	89	91	94	85	86	92	91	92	87	86	85	86	86	84	87	91	95	90	87	86	90	90	85	89	93	93	90	89	91	87
35	88	91	89	92	86	94	79	81	86	86	84	87	81	84	86	90	91	87	79	86	85	90	89	87	92	96	90	90	91	91
36	75	75	75	74	76	75	70	70	77	72	74	76	81	73	76	85	79	81	83	86	81	75	80	80	85	75	78	70	78	80
37	74	76	74	73	78	76	77	77	80	62	74	73	64	71	73	76	77	80	83	80	77	74	78	76	76	77	77	78	80	81
38	81	84	85	83	81	83	94	93	91	74	82	81	89	87	82	90	86	91	84	81	86	79	86	83	89	87	86	82	84	80
39	72	78	80	71	72	78	72	69	75	80	81	76	75	75	75	88	79	77	77	80	84	86	83	80	86	79	76	86	74	80
40	82	82	79	74	80	79	84	84	78	84	88	81	75	78	76	91	87	84	75	81	82	79	84	84	88	85	80	74	82	75
41	78	79	82	77	75	83	80	79	84	79	72	79	76	79	76	85	87	83	87	79	82	79	76	77	88	83	85	87	83	82
42	83	87	81	83	86	84	80	79	83	84	90	85	87	80	83	92	80	80	89	84	87	80	82	80	93	83	80	87	85	88
43	79	82	84	80	79	84	87	87	80	84	82	81	79	75	80	93	88	85	83	84	89	79	83	85	88	87	79	86	80	84
44	83	84	85	87	79	81	85	81	87	79	78	86	84	81	83	91	87	88	84	80	85	83	85	82	97	89	85	89	82	86
45	79	81	82	82	81	86	87	83	86	78	79	80	81	76	77	94	85	87	87	84	82	79	84	79	96	90	79	91	87	84
46	74	74	75	75	76	78	80	79	84	81	79	85	81	76	84	87	82	80	77	70	77	78	78	81	84	87	84	84	81	83
47	71	71	71	82	76	79	79	75	71	75	76	81	75	75	71	88	79	79	81	81	80	76	79	72	88	76	78	86	82	82
48	77	71	72	84	80	79	73	76	73	75	74	69	73	74	69	85	85	79	81	77	78	77	80	73	81	84	75	77	79	74
49	75	78	80	77	70	76	88	80	81	88	75	80	75	84	76	82	85	79	77	75	76	76	78	80	95	86	80	80	74	80
50	92	95	90	84	80	89	91	94	88	88	83	87	83	80	86	89	86	91	87	82	89	80	83	88	89	86	88	91	92	91

Appendix M

Table : ANOVA for the significance of component scores among 10 graders with the 1st Instrument,

ANOVA							
			Sum of Squares	df	Mean Square	F	Sig.
Grading 1	Total	Between Groups	11530,352	9	1281,150	3,680	,000
		Within Groups	170595,960	490	348,155		
		Total	182126,312	499			
	Task_Ach.	Between Groups	1554,792	9	172,755	2,366	,013
		Within Groups	35776,680	490	73,014		
		Total	37331,472	499			
	Essay_Org.	Between Groups	3007,832	9	334,204	3,699	,000
		Within Groups	44266,280	490	90,339		
		Total	47274,112	499			
	ACCURACY	Between Groups	357,952	9	39,772	1,779	,017
		Within Groups	10955,920	490	22,359		
		Total	11313,872	499			
Grading 2	Total	Between Groups	11544,792	9	1282,755	3,651	,000
		Within Groups	172161,480	490	351,350		
		Total	183706,272	499			
	Task_Ach.	Between Groups	1607,152	9	178,572	2,588	,006
		Within Groups	33810,120	490	69,000		
		Total	35417,272	499			
	Essay_Org.	Between Groups	2845,688	9	316,188	4,158	,000
		Within Groups	37259,240	490	76,039		
		Total	40104,928	499			
	ACCURACY	Between Groups	296,928	9	32,992	1,988	,039
		Within Groups	8131,720	490	16,595		
		Total	8428,648	499			

Appendix N

Table 9: Most problematic 9 papers after three gradings

		Instrument I			Instrument II		
		Grading Order			Grading Order		
Paper	Grader	1	2	3	1	2	3
3	1	**25,00**	**25,00**	**25,00**	**24,00**	24,00	27,00
	2	25,00	35,00	30,00	24,00	**19,00**	27,00
	3	45,00	35,00	50,00	27,00	29,00	28,00
	4	35,00	**55,00**	55,00	**35,00**	27,00	30,00
	5	65,00	35,00	35,00	28,00	29,00	24,00
	6	**85,00**	45,00	**70,00**	32,00	26,00	25,00
	7	55,00	45,00	55,00	26,00	27,00	24,00
	8	25,00	35,00	30,00	32,00	**31,00**	**30,00**
	9	25,00	45,00	45,00	25,00	21,00	**22,00**
	10	70,00	50,00	45,00	32,00	26,00	28,00
5	1	45,00	50,00	40,00	**39,00**	**40,00**	35,00
	2	**25,00**	35,00	35,00	31,00	35,00	29,00
	3	45,00	30,00	40,00	27,00	33,00	33,00
	4	30,00	**25,00**	**25,00**	34,00	28,00	33,00
	5	50,00	35,00	35,00	26,00	29,00	32,00
	6	**65,00**	**85,00**	**65,00**	**24,00**	31,00	33,00
	7	35,00	45,00	35,00	27,00	33,00	31,00
	8	55,00	25,00	57,00	37,00	33,00	**39,00**
	9	45,00	25,00	45,00	29,00	31,00	31,00
	10	65,00	55,00	60,00	26,00	**25,00**	**28,00**
14	1	45,00	45,00	45,00	29,00	27,00	32,00
	2	**25,00**	35,00	35,00	**25,00**	24,00	**26,00**
	3	45,00	50,00	40,00	25,00	24,00	31,00
	4	25,00	60,00	55,00	37,00	34,00	**37,00**
	5	35,00	**25,00**	35,00	26,00	28,00	30,00
	6	35,00	45,00	**75,00**	30,00	**23,00**	31,00
	7	**55,00**	30,00	50,00	**37,00**	32,00	35,00
	8	45,00	50,00	40,00	36,00	**35,00**	32,00
	9	30,00	25,00	**25,00**	25,00	28,00	30,00
	10	55,00	**65,00**	55,00	28,00	23,00	29,00
15	1	45,00	70,00	45,00	50,00	51,00	49,00
	2	60,00	55,00	55,00	58,00	56,00	50,00
	3	50,00	55,00	50,00	**37,00**	**50,00**	**48,00**
	4	40,00	55,00	55,00	51,00	57,00	49,00
	5	50,00	**35,00**	**35,00**	60,00	59,00	**62,00**
	6	**70,00**	55,00	50,00	63,00	**66,00**	58,00
	7	45,00	60,00	**65,00**	**67,00**	61,00	58,00
	8	65,00	60,00	65,00	57,00	52,00	55,00
	9	**35,00**	50,00	50,00	56,00	61,00	60,00
	10	65,00	**70,00**	65,00	53,00	58,00	56,00
16	1	75,00	82,00	75,00	61,00	**59,00**	**67,00**
	2	70,00	78,00	70,00	70,00	67,00	72,00
	3	70,00	75,00	75,00	76,00	72,00	**75,00**

Continued on page 181

	4	**55,00**	70,00	70,00	73,00	68,00	73,00
	5	83,00	**55,00**	**55,00**	68,00	62,00	68,00
	6	58,00	70,00	60,00	75,00	70,00	71,00
	7	80,00	91,00	91,00	**80,00**	**78,00**	74,00
	8	86,00	75,00	78,00	76,00	73,00	70,00
	9	88,00	83,00	80,00	**63,00**	69,00	69,00
	10	**91,00**	**91,00**	**95,00**	75,00	71,00	74,00
18	1	75,00	78,00	75,00	75,00	**77,00**	70,00
	2	70,00	68,00	83,00	58,00	54,00	60,00
	3	75,00	75,00	70,00	60,00	59,00	66,00
	4	70,00	78,00	70,00	52,00	57,00	61,00
	5	83,00	**60,00**	**65,00**	**51,00**	56,00	**58,00**
	6	**70,00**	85,00	75,00	66,00	63,00	60,00
	7	78,00	91,00	91,00	73,00	71,00	67,00
	8	78,00	75,00	76,00	**76,00**	69,00	**73,00**
	9	**94,00**	86,00	88,00	52,00	**52,00**	59,00
	10	83,00	**91,00**	**95,00**	65,00	65,00	66,00
27	1	75,00	75,00	75,00	**65,00**	**67,00**	71,00
	2	**60,00**	78,00	**60,00**	70,00	74,00	73,00
	3	65,00	75,00	70,00	67,00	69,00	72,00
	4	75,00	68,00	68,00	73,00	71,00	76,00
	5	86,00	**60,00**	75,00	75,00	70,00	71,00
	6	68,00	**85,00**	78,00	**78,00**	74,00	**79,00**
	7	86,00	83,00	**83,00**	70,00	71,00	74,00
	8	70,00	70,00	70,00	73,00	77,00	**69,00**
	9	**86,00**	75,00	75,00	74,00	75,00	73,00
	10	75,00	83,00	75,00	71,00	**78,00**	70,00
30	1	65,00	65,00	**50,00**	**64,00**	**65,00**	**68,00**
	2	75,00	70,00	70,00	66,00	69,00	73,00
	3	65,00	**55,00**	65,00	70,00	72,00	75,00
	4	65,00	**83,00**	75,00	69,00	65,00	72,00
	5	65,00	60,00	60,00	73,00	78,00	71,00
	6	83,00	75,00	80,00	**82,00**	**80,00**	**78,00**
	7	**85,00**	83,00	**85,00**	74,00	76,00	77,00
	8	78,00	70,00	80,00	74,00	75,00	70,00
	9	**58,00**	75,00	73,00	77,00	70,00	70,00
	10	83,00	75,00	83,00	72,00	72,00	75,00
50	1	94,00	94,00	94,00	**92,00**	**95,00**	90,00
	2	94,00	98,00	**97,00**	84,00	**80,00**	89,00
	3	83,00	**71,00**	**70,00**	91,00	94,00	88,00
	4	83,00	94,00	79,00	88,00	83,00	87,00
	5	98,00	94,00	96,00	83,00	80,00	**86,00**
	6	100,00	**100,00**	95,00	89,00	86,00	91,00
	7	98,00	94,00	91,00	87,00	82,00	89,00
	8	**80,00**	88,00	**85,00**	**80,00**	83,00	88,00
	9	**100,00**	86,00	86,00	89,00	86,00	88,00
	10	94,00	96,00	94,00	91,00	92,00	**91,00**

Appendix O

Table 9: Factor Analysis Charts

Post Hoc Tests

Multiple Comparisons
LSD

Grading	Dependent Variable	(I) GRADER	(J) GRADER	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
1	Total	1	2	1,30	3,732	,728	-6,03	8,63
			3	-,42	3,732	,910	-7,75	6,91
			4	,00	3,732	1,000	-7,33	7,33
			5	-7,28	3,732	,052	-14,61	,05
			6	-12,02(*)	3,732	,001	-19,35	-4,69
			7	-7,64(*)	3,732	,041	-14,97	-,31
			8	-3,68	3,732	,325	-11,01	3,65
			9	-9,10(*)	3,732	,015	-16,43	-1,77
			10	-11,32(*)	3,732	,003	-18,65	-3,99
		2	1	-1,30	3,732	,728	-8,63	6,03
			3	-1,72	3,732	,645	-9,05	5,61
			4	-1,30	3,732	,728	-8,63	6,03
			5	-8,58(*)	3,732	,022	-15,91	-1,25
			6	-13,32(*)	3,732	,000	-20,65	-5,99
			7	-8,94(*)	3,732	,017	-16,27	-1,61
			8	-4,98	3,732	,183	-12,31	2,35
			9	-10,40(*)	3,732	,006	-17,73	-3,07
			10	-12,62(*)	3,732	,001	-19,95	-5,29
		3	1	,42	3,732	,910	-6,91	7,75
			2	1,72	3,732	,645	-5,61	9,05
			4	,42	3,732	,910	-6,91	7,75
			5	-6,86	3,732	,067	-14,19	,47
			6	-11,60(*)	3,732	,002	-18,93	-4,27
			7	-7,22	3,732	,054	-14,55	,11
			8	-3,26	3,732	,383	-10,59	4,07
			9	-8,68(*)	3,732	,020	-16,01	-1,35
			10	-10,90(*)	3,732	,004	-18,23	-3,57

Continued on the following page

4	1	,00	3,732	1,000	-7,33	7,33
	2	1,30	3,732	,728	-6,03	8,63
	3	-,42	3,732	,910	-7,75	6,91
	5	-7,28	3,732	,052	-14,61	,05
	6	-12,02(*)	3,732	,001	-19,35	-4,69
	7	-7,64(*)	3,732	,041	-14,97	-,31
	8	-3,68	3,732	,325	-11,01	3,65
	9	-9,10(*)	3,732	,015	-16,43	-1,77
	10	-11,32(*)	3,732	,003	-18,65	-3,99
5	1	7,28	3,732	,052	-,05	14,61
	2	8,58(*)	3,732	,022	1,25	15,91
	3	6,86	3,732	,067	-,47	14,19
	4	7,28	3,732	,052	-,05	14,61
	6	-4,74	3,732	,205	-12,07	2,59
	7	-,36	3,732	,923	-7,69	6,97
	8	3,60	3,732	,335	-3,73	10,93
	9	-1,82	3,732	,626	-9,15	5,51
	10	-4,04	3,732	,280	-11,37	3,29
6	1	12,02(*)	3,732	,001	4,69	19,35
	2	13,32(*)	3,732	,000	5,99	20,65
	3	11,60(*)	3,732	,002	4,27	18,93
	4	12,02(*)	3,732	,001	4,69	19,35
	5	4,74	3,732	,205	-2,59	12,07
	7	4,38	3,732	,241	-2,95	11,71
	8	8,34(*)	3,732	,026	1,01	15,67
	9	2,92	3,732	,434	-4,41	10,25
	10	,70	3,732	,851	-6,63	8,03
7	1	7,64(*)	3,732	,041	,31	14,97
	2	8,94(*)	3,732	,017	1,61	16,27
	3	7,22	3,732	,054	-,11	14,55
	4	7,64(*)	3,732	,041	,31	14,97
	5	,36	3,732	,923	-6,97	7,69
	6	-4,38	3,732	,241	-11,71	2,95
	8	3,96	3,732	,289	-3,37	11,29
	9	-1,46	3,732	,696	-8,79	5,87
	10	-3,68	3,732	,325	-11,01	3,65
8	1	3,68	3,732	,325	-3,65	11,01
	2	4,98	3,732	,183	-2,35	12,31
	3	3,26	3,732	,383	-4,07	10,59

Continued on the following page

			4	3,68	3,732	,325	-3,65	11,01
			5	-3,60	3,732	,335	-10,93	3,73
			6	-8,34(*)	3,732	,026	-15,67	-1,01
			7	-3,96	3,732	,289	-11,29	3,37
			9	-5,42	3,732	,147	-12,75	1,91
			10	-7,64(*)	3,732	,041	-14,97	-,31
		9	1	9,10(*)	3,732	,015	1,77	16,43
			2	10,40(*)	3,732	,006	3,07	17,73
			3	8,68(*)	3,732	,020	1,35	16,01
			4	9,10(*)	3,732	,015	1,77	16,43
			5	1,82	3,732	,626	-5,51	9,15
			6	-2,92	3,732	,434	-10,25	4,41
			7	1,46	3,732	,696	-5,87	8,79
			8	5,42	3,732	,147	-1,91	12,75
			10	-2,22	3,732	,552	-9,55	5,11
		10	1	11,32(*)	3,732	,003	3,99	18,65
			2	12,62(*)	3,732	,001	5,29	19,95
			3	10,90(*)	3,732	,004	3,57	18,23
			4	11,32(*)	3,732	,003	3,99	18,65
			5	4,04	3,732	,280	-3,29	11,37
			6	-,70	3,732	,851	-8,03	6,63
			7	3,68	3,732	,325	-3,65	11,01
			8	7,64(*)	3,732	,041	,31	14,97
			9	2,22	3,732	,552	-5,11	9,55
	Task_Ach.	1	2	1,78	1,709	,298	-1,58	5,14
			3	,08	1,709	,963	-3,28	3,44
			4	,66	1,709	,700	-2,70	4,02
			5	-3,22	1,709	,060	-6,58	,14
			6	-3,52(*)	1,709	,040	-6,88	-,16
			7	-1,80	1,709	,293	-5,16	1,56
			8	-,86	1,709	,615	-4,22	2,50
			9	-3,14	1,709	,067	-6,50	,22
			10	-2,62	1,709	,126	-5,98	,74
		2	1	-1,78	1,709	,298	-5,14	1,58
			3	-1,70	1,709	,320	-5,06	1,66
			4	-1,12	1,709	,513	-4,48	2,24
			5	-5,00(*)	1,709	,004	-8,36	-1,64
			6	-5,30(*)	1,709	,002	-8,66	-1,94
			7	-3,58(*)	1,709	,037	-6,94	-,22

Continued on the following page

	8	-2,64	1,709	,123	-6,00	,72
	9	-4,92(*)	1,709	,004	-8,28	-1,56
	10	-4,40(*)	1,709	,010	-7,76	-1,04
3	1	-,08	1,709	,963	-3,44	3,28
	2	1,70	1,709	,320	-1,66	5,06
	4	,58	1,709	,734	-2,78	3,94
	5	-3,30	1,709	,054	-6,66	,06
	6	-3,60(*)	1,709	,036	-6,96	-,24
	7	-1,88	1,709	,272	-5,24	1,48
	8	-,94	1,709	,583	-4,30	2,42
	9	-3,22	1,709	,060	-6,58	,14
	10	-2,70	1,709	,115	-6,06	,66
4	1	-,66	1,709	,700	-4,02	2,70
	2	1,12	1,709	,513	-2,24	4,48
	3	-,58	1,709	,734	-3,94	2,78
	5	-3,88(*)	1,709	,024	-7,24	-,52
	6	-4,18(*)	1,709	,015	-7,54	-,82
	7	-2,46	1,709	,151	-5,82	,90
	8	-1,52	1,709	,374	-4,88	1,84
	9	-3,80(*)	1,709	,027	-7,16	-,44
	10	-3,28	1,709	,056	-6,64	,08
5	1	3,22	1,709	,060	-,14	6,58
	2	5,00(*)	1,709	,004	1,64	8,36
	3	3,30	1,709	,054	-,06	6,66
	4	3,88(*)	1,709	,024	,52	7,24
	6	-,30	1,709	,861	-3,66	3,06
	7	1,42	1,709	,406	-1,94	4,78
	8	2,36	1,709	,168	-1,00	5,72
	9	,08	1,709	,963	-3,28	3,44
	10	,60	1,709	,726	-2,76	3,96
6	1	3,52(*)	1,709	,040	,16	6,88
	2	5,30(*)	1,709	,002	1,94	8,66
	3	3,60(*)	1,709	,036	,24	6,96
	4	4,18(*)	1,709	,015	,82	7,54
	5	,30	1,709	,861	-3,06	3,66
	7	1,72	1,709	,315	-1,64	5,08
	8	2,66	1,709	,120	-,70	6,02
	9	,38	1,709	,824	-2,98	3,74
	10	,90	1,709	,599	-2,46	4,26

Continued on the following page

		7	1	1,80	1,709	,293	-1,56	5,16
			2	3,58(*)	1,709	,037	,22	6,94
			3	1,88	1,709	,272	-1,48	5,24
			4	2,46	1,709	,151	-,90	5,82
			5	-1,42	1,709	,406	-4,78	1,94
			6	-1,72	1,709	,315	-5,08	1,64
			8	,94	1,709	,583	-2,42	4,30
			9	-1,34	1,709	,433	-4,70	2,02
			10	-,82	1,709	,632	-4,18	2,54
		8	1	,86	1,709	,615	-2,50	4,22
			2	2,64	1,709	,123	-,72	6,00
			3	,94	1,709	,583	-2,42	4,30
			4	1,52	1,709	,374	-1,84	4,88
			5	-2,36	1,709	,168	-5,72	1,00
			6	-2,66	1,709	,120	-6,02	,70
			7	-,94	1,709	,583	-4,30	2,42
			9	-2,28	1,709	,183	-5,64	1,08
			10	-1,76	1,709	,304	-5,12	1,60
		9	1	3,14	1,709	,067	-,22	6,50
			2	4,92(*)	1,709	,004	1,56	8,28
			3	3,22	1,709	,060	-,14	6,58
			4	3,80(*)	1,709	,027	,44	7,16
			5	-,08	1,709	,963	-3,44	3,28
			6	-,38	1,709	,824	-3,74	2,98
			7	1,34	1,709	,433	-2,02	4,70
			8	2,28	1,709	,183	-1,08	5,64
			10	,52	1,709	,761	-2,84	3,88
		10	1	2,62	1,709	,126	-,74	5,98
			2	4,40(*)	1,709	,010	1,04	7,76
			3	2,70	1,709	,115	-,66	6,06
			4	3,28	1,709	,056	-,08	6,64
			5	-,60	1,709	,726	-3,96	2,76
			6	-,90	1,709	,599	-4,26	2,46
			7	,82	1,709	,632	-2,54	4,18
			8	1,76	1,709	,304	-1,60	5,12
			9	-,52	1,709	,761	-3,88	2,84
	Essay_Org.	1	2	,52	1,901	,785	-3,21	4,25
			3	-,56	1,901	,768	-4,29	3,17
			4	-,52	1,901	,785	-4,25	3,21

Continued on the following page

			5	-2,00	1,901	,293	-5,73	1,73
			6	-6,28(*)	1,901	,001	-10,01	-2,55
			7	-3,36	1,901	,078	-7,09	,37
			8	-2,06	1,901	,279	-5,79	1,67
			9	-4,46(*)	1,901	,019	-8,19	-,73
			10	-6,72(*)	1,901	,000	-10,45	-2,99
		2	1	-,52	1,901	,785	-4,25	3,21
			3	-1,08	1,901	,570	-4,81	2,65
			4	-1,04	1,901	,585	-4,77	2,69
			5	-2,52	1,901	,186	-6,25	1,21
			6	-6,80(*)	1,901	,000	-10,53	-3,07
			7	-3,88(*)	1,901	,042	-7,61	-,15
			8	-2,58	1,901	,175	-6,31	1,15
			9	-4,98(*)	1,901	,009	-8,71	-1,25
			10	-7,24(*)	1,901	,000	-10,97	-3,51
		3	1	,56	1,901	,768	-3,17	4,29
			2	1,08	1,901	,570	-2,65	4,81
			4	,04	1,901	,983	-3,69	3,77
			5	-1,44	1,901	,449	-5,17	2,29
			6	-5,72(*)	1,901	,003	-9,45	-1,99
			7	-2,80	1,901	,141	-6,53	,93
			8	-1,50	1,901	,430	-5,23	2,23
			9	-3,90(*)	1,901	,041	-7,63	-,17
			10	-6,16(*)	1,901	,001	-9,89	-2,43
		4	1	,52	1,901	,785	-3,21	4,25
			2	1,04	1,901	,585	-2,69	4,77
			3	-,04	1,901	,983	-3,77	3,69
			5	-1,48	1,901	,437	-5,21	2,25
			6	-5,76(*)	1,901	,003	-9,49	-2,03
			7	-2,84	1,901	,136	-6,57	,89
			8	-1,54	1,901	,418	-5,27	2,19
			9	-3,94(*)	1,901	,039	-7,67	-,21
			10	-6,20(*)	1,901	,001	-9,93	-2,47
		5	1	2,00	1,901	,293	-1,73	5,73
			2	2,52	1,901	,186	-1,21	6,25
			3	1,44	1,901	,449	-2,29	5,17
			4	1,48	1,901	,437	-2,25	5,21
			6	-4,28(*)	1,901	,025	-8,01	-,55
			7	-1,36	1,901	,475	-5,09	2,37
			8	-,06	1,901	,975	-3,79	3,67

	9	-2,46	1,901	,196	-6,19	1,27
	10	-4,72(*)	1,901	,013	-8,45	-,99
6	1	6,28(*)	1,901	,001	2,55	10,01
	2	6,80(*)	1,901	,000	3,07	10,53
	3	5,72(*)	1,901	,003	1,99	9,45
	4	5,76(*)	1,901	,003	2,03	9,49
	5	4,28(*)	1,901	,025	,55	8,01
	7	2,92	1,901	,125	-,81	6,65
	8	4,22(*)	1,901	,027	,49	7,95
	9	1,82	1,901	,339	-1,91	5,55
	10	-,44	1,901	,817	-4,17	3,29
7	1	3,36	1,901	,078	-,37	7,09
	2	3,88(*)	1,901	,042	,15	7,61
	3	2,80	1,901	,141	-,93	6,53
	4	2,84	1,901	,136	-,89	6,57
	5	1,36	1,901	,475	-2,37	5,09
	6	-2,92	1,901	,125	-6,65	,81
	8	1,30	1,901	,494	-2,43	5,03
	9	-1,10	1,901	,563	-4,83	2,63
	10	-3,36	1,901	,078	-7,09	,37
8	1	2,06	1,901	,279	-1,67	5,79
	2	2,58	1,901	,175	-1,15	6,31
	3	1,50	1,901	,430	-2,23	5,23
	4	1,54	1,901	,418	-2,19	5,27
	5	,06	1,901	,975	-3,67	3,79
	6	-4,22(*)	1,901	,027	-7,95	-,49
	7	-1,30	1,901	,494	-5,03	2,43
	9	-2,40	1,901	,207	-6,13	1,33
	10	-4,66(*)	1,901	,015	-8,39	-,93
9	1	4,46(*)	1,901	,019	,73	8,19
	2	4,98(*)	1,901	,009	1,25	8,71
	3	3,90(*)	1,901	,041	,17	7,63
	4	3,94(*)	1,901	,039	,21	7,67
	5	2,46	1,901	,196	-1,27	6,19
	6	-1,82	1,901	,339	-5,55	1,91
	7	1,10	1,901	,563	-2,63	4,83
	8	2,40	1,901	,207	-1,33	6,13
	10	-2,26	1,901	,235	-5,99	1,47
10	1	6,72(*)	1,901	,000	2,99	10,45

Continued on the following page

		2	7,24(*)	1,901	,000	3,51	10,97
		3	6,16(*)	1,901	,001	2,43	9,89
		4	6,20(*)	1,901	,001	2,47	9,93
		5	4,72(*)	1,901	,013	,99	8,45
		6	,44	1,901	,817	-3,29	4,17
		7	3,36	1,901	,078	-,37	7,09
		8	4,66(*)	1,901	,015	,93	8,39
		9	2,26	1,901	,235	-1,47	5,99
ACCURACY	1	2	-1,00	,946	,291	-2,86	,86
		3	,06	,946	,949	-1,80	1,92
		4	-1,38	,946	,145	-3,24	,48
		5	-1,86(*)	,946	,050	-3,72	,00
		6	-2,22(*)	,946	,019	-4,08	-,36
		7	-2,48(*)	,946	,009	-4,34	-,62
		8	-,96	,946	,311	-2,82	,90
		9	-1,50	,946	,113	-3,36	,36
		10	-2,22(*)	,946	,019	-4,08	-,36
	2	1	1,00	,946	,291	-,86	2,86
		3	1,06	,946	,263	-,80	2,92
		4	-,38	,946	,688	-2,24	1,48
		5	-,86	,946	,364	-2,72	1,00
		6	-1,22	,946	,198	-3,08	,64
		7	-1,48	,946	,118	-3,34	,38
		8	,04	,946	,966	-1,82	1,90
		9	-,50	,946	,597	-2,36	1,36
		10	-1,22	,946	,198	-3,08	,64
	3	1	-,06	,946	,949	-1,92	1,80
		2	-1,06	,946	,263	-2,92	,80
		4	-1,44	,946	,128	-3,30	,42
		5	-1,92(*)	,946	,043	-3,78	-,06
		6	-2,28(*)	,946	,016	-4,14	-,42
		7	-2,54(*)	,946	,007	-4,40	-,68
		8	-1,02	,946	,281	-2,88	,84
		9	-1,56	,946	,100	-3,42	,30
		10	-2,28(*)	,946	,016	-4,14	-,42
	4	1	1,38	,946	,145	-,48	3,24
		2	,38	,946	,688	-1,48	2,24
		3	1,44	,946	,128	-,42	3,30
		5	-,48	,946	,612	-2,34	1,38

Continued on the following page

			6	-,84	,946	,375	-2,70	1,02
			7	-1,10	,946	,245	-2,96	,76
			8	,42	,946	,657	-1,44	2,28
			9	-,12	,946	,899	-1,98	1,74
			10	-,84	,946	,375	-2,70	1,02
		5	1	1,86(*)	,946	,050	,00	3,72
			2	,86	,946	,364	-1,00	2,72
			3	1,92(*)	,946	,043	,06	3,78
			4	,48	,946	,612	-1,38	2,34
			6	-,36	,946	,704	-2,22	1,50
			7	-,62	,946	,512	-2,48	1,24
			8	,90	,946	,342	-,96	2,76
			9	,36	,946	,704	-1,50	2,22
			10	-,36	,946	,704	-2,22	1,50
		6	1	2,22(*)	,946	,019	,36	4,08
			2	1,22	,946	,198	-,64	3,08
			3	2,28(*)	,946	,016	,42	4,14
			4	,84	,946	,375	-1,02	2,70
			5	,36	,946	,704	-1,50	2,22
			7	-,26	,946	,783	-2,12	1,60
			8	1,26	,946	,183	-,60	3,12
			9	,72	,946	,447	-1,14	2,58
			10	,00	,946	1,000	-1,86	1,86
		7	1	2,48(*)	,946	,009	,62	4,34
			2	1,48	,946	,118	-,38	3,34
			3	2,54(*)	,946	,007	,68	4,40
			4	1,10	,946	,245	-,76	2,96
			5	,62	,946	,512	-1,24	2,48
			6	,26	,946	,783	-1,60	2,12
			8	1,52	,946	,109	-,34	3,38
			9	,98	,946	,301	-,88	2,84
			10	,26	,946	,783	-1,60	2,12
		8	1	,96	,946	,311	-,90	2,82
			2	-,04	,946	,966	-1,90	1,82
			3	1,02	,946	,281	-,84	2,88
			4	-,42	,946	,657	-2,28	1,44
			5	-,90	,946	,342	-2,76	,96
			6	-1,26	,946	,183	-3,12	,60
			7	-1,52	,946	,109	-3,38	,34

Continued on the following page

			9	-,54	,946	,568	-2,40	1,32
			10	-1,26	,946	,183	-3,12	,60
		9	1	1,50	,946	,113	-,36	3,36
			2	,50	,946	,597	-1,36	2,36
			3	1,56	,946	,100	-,30	3,42
			4	,12	,946	,899	-1,74	1,98
			5	-,36	,946	,704	-2,22	1,50
			6	-,72	,946	,447	-2,58	1,14
			7	-,98	,946	,301	-2,84	,88
			8	,54	,946	,568	-1,32	2,40
			10	-,72	,946	,447	-2,58	1,14
		10	1	2,22(*)	,946	,019	,36	4,08
			2	1,22	,946	,198	-,64	3,08
			3	2,28(*)	,946	,016	,42	4,14
			4	,84	,946	,375	-1,02	2,70
			5	,36	,946	,704	-1,50	2,22
			6	,00	,946	1,000	-1,86	1,86
			7	-,26	,946	,783	-2,12	1,60
			8	1,26	,946	,183	-,60	3,12
			9	,72	,946	,447	-1,14	2,58
2	Total	1	2	-2,38	3,749	,526	-9,75	4,99
			3	1,16	3,749	,757	-6,21	8,53
			4	-5,04	3,749	,179	-12,41	2,33
			5	3,44	3,749	,359	-3,93	10,81
			6	-4,62	3,749	,218	-11,99	2,75
			7	-6,60	3,749	,079	-13,97	,77
			8	1,56	3,749	,678	-5,81	8,93
			9	-4,30	3,749	,252	-11,67	3,07
			10	-13,98(*)	3,749	,000	-21,35	-6,61
		2	1	2,38	3,749	,526	-4,99	9,75
			3	3,54	3,749	,345	-3,83	10,91
			4	-2,66	3,749	,478	-10,03	4,71
			5	5,82	3,749	,121	-1,55	13,19
			6	-2,24	3,749	,550	-9,61	5,13
			7	-4,22	3,749	,261	-11,59	3,15
			8	3,94	3,749	,294	-3,43	11,31
			9	-1,92	3,749	,609	-9,29	5,45
			10	-11,60(*)	3,749	,002	-18,97	-4,23
		3	1	-1,16	3,749	,757	-8,53	6,21

Continued on the following page

	2	-3,54	3,749	,345	-10,91	3,83
	4	-6,20	3,749	,099	-13,57	1,17
	5	2,28	3,749	,543	-5,09	9,65
	6	-5,78	3,749	,124	-13,15	1,59
	7	-7,76(*)	3,749	,039	-15,13	-,39
	8	,40	3,749	,915	-6,97	7,77
	9	-5,46	3,749	,146	-12,83	1,91
	10	-15,14(*)	3,749	,000	-22,51	-7,77
4	1	5,04	3,749	,179	-2,33	12,41
	2	2,66	3,749	,478	-4,71	10,03
	3	6,20	3,749	,099	-1,17	13,57
	5	8,48(*)	3,749	,024	1,11	15,85
	6	,42	3,749	,911	-6,95	7,79
	7	-1,56	3,749	,678	-8,93	5,81
	8	6,60	3,749	,079	-,77	13,97
	9	,74	3,749	,844	-6,63	8,11
	10	-8,94(*)	3,749	,017	-16,31	-1,57
5	1	-3,44	3,749	,359	-10,81	3,93
	2	-5,82	3,749	,121	-13,19	1,55
	3	-2,28	3,749	,543	-9,65	5,09
	4	-8,48(*)	3,749	,024	-15,85	-1,11
	6	-8,06(*)	3,749	,032	-15,43	-,69
	7	-10,04(*)	3,749	,008	-17,41	-2,67
	8	-1,88	3,749	,616	-9,25	5,49
	9	-7,74(*)	3,749	,039	-15,11	-,37
	10	-17,42(*)	3,749	,000	-24,79	-10,05
6	1	4,62	3,749	,218	-2,75	11,99
	2	2,24	3,749	,550	-5,13	9,61
	3	5,78	3,749	,124	-1,59	13,15
	4	-,42	3,749	,911	-7,79	6,95
	5	8,06(*)	3,749	,032	,69	15,43
	7	-1,98	3,749	,598	-9,35	5,39
	8	6,18	3,749	,100	-1,19	13,55
	9	,32	3,749	,932	-7,05	7,69
	10	-9,36(*)	3,749	,013	-16,73	-1,99
7	1	6,60	3,749	,079	-,77	13,97
	2	4,22	3,749	,261	-3,15	11,59
	3	7,76(*)	3,749	,039	,39	15,13
	4	1,56	3,749	,678	-5,81	8,93

Continued on the following page

			5	10,04(*)	3,749	,008	2,67	17,41
			6	1,98	3,749	,598	-5,39	9,35
			8	8,16(*)	3,749	,030	,79	15,53
			9	2,30	3,749	,540	-5,07	9,67
			10	-7,38(*)	3,749	,050	-14,75	-,01
		8	1	-1,56	3,749	,678	-8,93	5,81
			2	-3,94	3,749	,294	-11,31	3,43
			3	-,40	3,749	,915	-7,77	6,97
			4	-6,60	3,749	,079	-13,97	,77
			5	1,88	3,749	,616	-5,49	9,25
			6	-6,18	3,749	,100	-13,55	1,19
			7	-8,16(*)	3,749	,030	-15,53	-,79
			9	-5,86	3,749	,119	-13,23	1,51
			10	-15,54(*)	3,749	,000	-22,91	-8,17
		9	1	4,30	3,749	,252	-3,07	11,67
			2	1,92	3,749	,609	-5,45	9,29
			3	5,46	3,749	,146	-1,91	12,83
			4	-,74	3,749	,844	-8,11	6,63
			5	7,74(*)	3,749	,039	,37	15,11
			6	-,32	3,749	,932	-7,69	7,05
			7	-2,30	3,749	,540	-9,67	5,07
			8	5,86	3,749	,119	-1,51	13,23
			10	-9,68(*)	3,749	,010	-17,05	-2,31
		10	1	13,98(*)	3,749	,000	6,61	21,35
			2	11,60(*)	3,749	,002	4,23	18,97
			3	15,14(*)	3,749	,000	7,77	22,51
			4	8,94(*)	3,749	,017	1,57	16,31
			5	17,42(*)	3,749	,000	10,05	24,79
			6	9,36(*)	3,749	,013	1,99	16,73
			7	7,38(*)	3,749	,050	,01	14,75
			8	15,54(*)	3,749	,000	8,17	22,91
			9	9,68(*)	3,749	,010	2,31	17,05
	Task_Ach.	1	2	-1,16	1,661	,485	-4,42	2,10
			3	-,04	1,661	,981	-3,30	3,22
			4	-2,24	1,661	,178	-5,50	1,02
			5	1,66	1,661	,318	-1,60	4,92
			6	-2,94	1,661	,077	-6,20	,32
			7	-2,14	1,661	,198	-5,40	1,12
			8	,84	1,661	,613	-2,42	4,10

Continued on the following page

	9	-,36	1,661	,829	-3,62	2,90
	10	-4,58(*)	1,661	,006	-7,84	-1,32
2	1	1,16	1,661	,485	-2,10	4,42
	3	1,12	1,661	,501	-2,14	4,38
	4	-1,08	1,661	,516	-4,34	2,18
	5	2,82	1,661	,090	-,44	6,08
	6	-1,78	1,661	,285	-5,04	1,48
	7	-,98	1,661	,556	-4,24	2,28
	8	2,00	1,661	,229	-1,26	5,26
	9	,80	1,661	,630	-2,46	4,06
	10	-3,42(*)	1,661	,040	-6,68	-,16
3	1	,04	1,661	,981	-3,22	3,30
	2	-1,12	1,661	,501	-4,38	2,14
	4	-2,20	1,661	,186	-5,46	1,06
	5	1,70	1,661	,307	-1,56	4,96
	6	-2,90	1,661	,082	-6,16	,36
	7	-2,10	1,661	,207	-5,36	1,16
	8	,88	1,661	,597	-2,38	4,14
	9	-,32	1,661	,847	-3,58	2,94
	10	-4,54(*)	1,661	,007	-7,80	-1,28
4	1	2,24	1,661	,178	-1,02	5,50
	2	1,08	1,661	,516	-2,18	4,34
	3	2,20	1,661	,186	-1,06	5,46
	5	3,90(*)	1,661	,019	,64	7,16
	6	-,70	1,661	,674	-3,96	2,56
	7	,10	1,661	,952	-3,16	3,36
	8	3,08	1,661	,064	-,18	6,34
	9	1,88	1,661	,258	-1,38	5,14
	10	-2,34	1,661	,160	-5,60	,92
5	1	-1,66	1,661	,318	-4,92	1,60
	2	-2,82	1,661	,090	-6,08	,44
	3	-1,70	1,661	,307	-4,96	1,56
	4	-3,90(*)	1,661	,019	-7,16	-,64
	6	-4,60(*)	1,661	,006	-7,86	-1,34
	7	-3,80(*)	1,661	,023	-7,06	-,54
	8	-,82	1,661	,622	-4,08	2,44
	9	-2,02	1,661	,225	-5,28	1,24
	10	-6,24(*)	1,661	,000	-9,50	-2,98
6	1	2,94	1,661	,077	-,32	6,20
		Continued on the following page				

			2	1,78	1,661	,285	-1,48	5,04
			3	2,90	1,661	,082	-,36	6,16
			4	,70	1,661	,674	-2,56	3,96
			5	4,60(*)	1,661	,006	1,34	7,86
			7	,80	1,661	,630	-2,46	4,06
			8	3,78(*)	1,661	,023	,52	7,04
			9	2,58	1,661	,121	-,68	5,84
			10	-1,64	1,661	,324	-4,90	1,62
		7	1	2,14	1,661	,198	-1,12	5,40
			2	,98	1,661	,556	-2,28	4,24
			3	2,10	1,661	,207	-1,16	5,36
			4	-,10	1,661	,952	-3,36	3,16
			5	3,80(*)	1,661	,023	,54	7,06
			6	-,80	1,661	,630	-4,06	2,46
			8	2,98	1,661	,073	-,28	6,24
			9	1,78	1,661	,285	-1,48	5,04
			10	-2,44	1,661	,143	-5,70	,82
		8	1	-,84	1,661	,613	-4,10	2,42
			2	-2,00	1,661	,229	-5,26	1,26
			3	-,88	1,661	,597	-4,14	2,38
			4	-3,08	1,661	,064	-6,34	,18
			5	,82	1,661	,622	-2,44	4,08
			6	-3,78(*)	1,661	,023	-7,04	-,52
			7	-2,98	1,661	,073	-6,24	,28
			9	-1,20	1,661	,470	-4,46	2,06
			10	-5,42(*)	1,661	,001	-8,68	-2,16
		9	1	,36	1,661	,829	-2,90	3,62
			2	-,80	1,661	,630	-4,06	2,46
			3	,32	1,661	,847	-2,94	3,58
			4	-1,88	1,661	,258	-5,14	1,38
			5	2,02	1,661	,225	-1,24	5,28
			6	-2,58	1,661	,121	-5,84	,68
			7	-1,78	1,661	,285	-5,04	1,48
			8	1,20	1,661	,470	-2,06	4,46
			10	-4,22(*)	1,661	,011	-7,48	-,96
		10	1	4,58(*)	1,661	,006	1,32	7,84
			2	3,42(*)	1,661	,040	,16	6,68
			3	4,54(*)	1,661	,007	1,28	7,80
			4	2,34	1,661	,160	-,92	5,60

Continued on the following page

		5	6,24(*)	1,661	,000	2,98	9,50
		6	1,64	1,661	,324	-1,62	4,90
		7	2,44	1,661	,143	-,82	5,70
		8	5,42(*)	1,661	,001	2,16	8,68
		9	4,22(*)	1,661	,011	,96	7,48
Essay_Org.	1	2	-1,20	1,744	,492	-4,63	2,23
		3	1,42	1,744	,416	-2,01	4,85
		4	-2,32	1,744	,184	-5,75	1,11
		5	1,00	1,744	,567	-2,43	4,43
		6	-1,06	1,744	,544	-4,49	2,37
		7	-3,68(*)	1,744	,035	-7,11	-,25
		8	,46	1,744	,792	-2,97	3,89
		9	-3,20	1,744	,067	-6,63	,23
		10	-6,74(*)	1,744	,000	-10,17	-3,31
	2	1	1,20	1,744	,492	-2,23	4,63
		3	2,62	1,744	,134	-,81	6,05
		4	-1,12	1,744	,521	-4,55	2,31
		5	2,20	1,744	,208	-1,23	5,63
		6	,14	1,744	,936	-3,29	3,57
		7	-2,48	1,744	,156	-5,91	,95
		8	1,66	1,744	,342	-1,77	5,09
		9	-2,00	1,744	,252	-5,43	1,43
		10	-5,54(*)	1,744	,002	-8,97	-2,11
	3	1	-1,42	1,744	,416	-4,85	2,01
		2	-2,62	1,744	,134	-6,05	,81
		4	-3,74(*)	1,744	,032	-7,17	-,31
		5	-,42	1,744	,810	-3,85	3,01
		6	-2,48	1,744	,156	-5,91	,95
		7	-5,10(*)	1,744	,004	-8,53	-1,67
		8	-,96	1,744	,582	-4,39	2,47
		9	-4,62(*)	1,744	,008	-8,05	-1,19
		10	-8,16(*)	1,744	,000	-11,59	-4,73
	4	1	2,32	1,744	,184	-1,11	5,75
		2	1,12	1,744	,521	-2,31	4,55
		3	3,74(*)	1,744	,032	,31	7,17
		5	3,32	1,744	,058	-,11	6,75
		6	1,26	1,744	,470	-2,17	4,69
		7	-1,36	1,744	,436	-4,79	2,07
		8	2,78	1,744	,112	-,65	6,21

Continued on the following page

			9	-,88	1,744	,614	-4,31	2,55
			10	-4,42(*)	1,744	,012	-7,85	-,99
		5	1	-1,00	1,744	,567	-4,43	2,43
			2	-2,20	1,744	,208	-5,63	1,23
			3	,42	1,744	,810	-3,01	3,85
			4	-3,32	1,744	,058	-6,75	,11
			6	-2,06	1,744	,238	-5,49	1,37
			7	-4,68(*)	1,744	,008	-8,11	-1,25
			8	-,54	1,744	,757	-3,97	2,89
			9	-4,20(*)	1,744	,016	-7,63	-,77
			10	-7,74(*)	1,744	,000	-11,17	-4,31
		6	1	1,06	1,744	,544	-2,37	4,49
			2	-,14	1,744	,936	-3,57	3,29
			3	2,48	1,744	,156	-,95	5,91
			4	-1,26	1,744	,470	-4,69	2,17
			5	2,06	1,744	,238	-1,37	5,49
			7	-2,62	1,744	,134	-6,05	,81
			8	1,52	1,744	,384	-1,91	4,95
			9	-2,14	1,744	,220	-5,57	1,29
			10	-5,68(*)	1,744	,001	-9,11	-2,25
		7	1	3,68(*)	1,744	,035	,25	7,11
			2	2,48	1,744	,156	-,95	5,91
			3	5,10(*)	1,744	,004	1,67	8,53
			4	1,36	1,744	,436	-2,07	4,79
			5	4,68(*)	1,744	,008	1,25	8,11
			6	2,62	1,744	,134	-,81	6,05
			8	4,14(*)	1,744	,018	,71	7,57
			9	,48	1,744	,783	-2,95	3,91
			10	-3,06	1,744	,080	-6,49	,37
		8	1	-,46	1,744	,792	-3,89	2,97
			2	-1,66	1,744	,342	-5,09	1,77
			3	,96	1,744	,582	-2,47	4,39
			4	-2,78	1,744	,112	-6,21	,65
			5	,54	1,744	,757	-2,89	3,97
			6	-1,52	1,744	,384	-4,95	1,91
			7	-4,14(*)	1,744	,018	-7,57	-,71
			9	-3,66(*)	1,744	,036	-7,09	-,23
			10	-7,20(*)	1,744	,000	-10,63	-3,77
		9	1	3,20	1,744	,067	-,23	6,63

Continued on the following page

			2	2,00	1,744	,252	-1,43	5,43
			3	4,62(*)	1,744	,008	1,19	8,05
			4	,88	1,744	,614	-2,55	4,31
			5	4,20(*)	1,744	,016	,77	7,63
			6	2,14	1,744	,220	-1,29	5,57
			7	-,48	1,744	,783	-3,91	2,95
			8	3,66(*)	1,744	,036	,23	7,09
			10	-3,54(*)	1,744	,043	-6,97	-,11
		10	1	6,74(*)	1,744	,000	3,31	10,17
			2	5,54(*)	1,744	,002	2,11	8,97
			3	8,16(*)	1,744	,000	4,73	11,59
			4	4,42(*)	1,744	,012	,99	7,85
			5	7,74(*)	1,744	,000	4,31	11,17
			6	5,68(*)	1,744	,001	2,25	9,11
			7	3,06	1,744	,080	-,37	6,49
			8	7,20(*)	1,744	,000	3,77	10,63
			9	3,54(*)	1,744	,043	,11	6,97
ACCURACY		1	2	-,02	,815	,980	-1,62	1,58
			3	-,22	,815	,787	-1,82	1,38
			4	-,48	,815	,556	-2,08	1,12
			5	,78	,815	,339	-,82	2,38
			6	-,62	,815	,447	-2,22	,98
			7	-,78	,815	,339	-2,38	,82
			8	,26	,815	,750	-1,34	1,86
			9	-,74	,815	,364	-2,34	,86
			10	-2,26(*)	,815	,006	-3,86	-,66
		2	1	,02	,815	,980	-1,58	1,62
			3	-,20	,815	,806	-1,80	1,40
			4	-,46	,815	,573	-2,06	1,14
			5	,80	,815	,327	-,80	2,40
			6	-,60	,815	,462	-2,20	1,00
			7	-,76	,815	,351	-2,36	,84
			8	,28	,815	,731	-1,32	1,88
			9	-,72	,815	,377	-2,32	,88
			10	-2,24(*)	,815	,006	-3,84	-,64
		3	1	,22	,815	,787	-1,38	1,82
			2	,20	,815	,806	-1,40	1,80
			4	-,26	,815	,750	-1,86	1,34
			5	1,00	,815	,220	-,60	2,60
				Continued on the following page				

			6	-,40	,815	,624	-2,00	1,20
			7	-,56	,815	,492	-2,16	1,04
			8	,48	,815	,556	-1,12	2,08
			9	-,52	,815	,524	-2,12	1,08
			10	-2,04(*)	,815	,013	-3,64	-,44
		4	1	,48	,815	,556	-1,12	2,08
			2	,46	,815	,573	-1,14	2,06
			3	,26	,815	,750	-1,34	1,86
			5	1,26	,815	,123	-,34	2,86
			6	-,14	,815	,864	-1,74	1,46
			7	-,30	,815	,713	-1,90	1,30
			8	,74	,815	,364	-,86	2,34
			9	-,26	,815	,750	-1,86	1,34
			10	-1,78(*)	,815	,029	-3,38	-,18
		5	1	-,78	,815	,339	-2,38	,82
			2	-,80	,815	,327	-2,40	,80
			3	-1,00	,815	,220	-2,60	,60
			4	-1,26	,815	,123	-2,86	,34
			6	-1,40	,815	,086	-3,00	,20
			7	-1,56	,815	,056	-3,16	,04
			8	-,52	,815	,524	-2,12	1,08
			9	-1,52	,815	,063	-3,12	,08
			10	-3,04(*)	,815	,000	-4,64	-1,44
		6	1	,62	,815	,447	-,98	2,22
			2	,60	,815	,462	-1,00	2,20
			3	,40	,815	,624	-1,20	2,00
			4	,14	,815	,864	-1,46	1,74
			5	1,40	,815	,086	-,20	3,00
			7	-,16	,815	,844	-1,76	1,44
			8	,88	,815	,281	-,72	2,48
			9	-,12	,815	,883	-1,72	1,48
			10	-1,64(*)	,815	,045	-3,24	-,04
		7	1	,78	,815	,339	-,82	2,38
			2	,76	,815	,351	-,84	2,36
			3	,56	,815	,492	-1,04	2,16
			4	,30	,815	,713	-1,30	1,90
			5	1,56	,815	,056	-,04	3,16
			6	,16	,815	,844	-1,44	1,76
			8	1,04	,815	,202	-,56	2,64

Continued on the following page

			9	,04	,815	,961	-1,56	1,64
			10	-1,48	,815	,070	-3,08	,12
		8	1	-,26	,815	,750	-1,86	1,34
			2	-,28	,815	,731	-1,88	1,32
			3	-,48	,815	,556	-2,08	1,12
			4	-,74	,815	,364	-2,34	,86
			5	,52	,815	,524	-1,08	2,12
			6	-,88	,815	,281	-2,48	,72
			7	-1,04	,815	,202	-2,64	,56
			9	-1,00	,815	,220	-2,60	,60
			10	-2,52(*)	,815	,002	-4,12	-,92
		9	1	,74	,815	,364	-,86	2,34
			2	,72	,815	,377	-,88	2,32
			3	,52	,815	,524	-1,08	2,12
			4	,26	,815	,750	-1,34	1,86
			5	1,52	,815	,063	-,08	3,12
			6	,12	,815	,883	-1,48	1,72
			7	-,04	,815	,961	-1,64	1,56
			8	1,00	,815	,220	-,60	2,60
			10	-1,52	,815	,063	-3,12	,08
		10	1	2,26(*)	,815	,006	,66	3,86
			2	2,24(*)	,815	,006	,64	3,84
			3	2,04(*)	,815	,013	,44	3,64
			4	1,78(*)	,815	,029	,18	3,38
			5	3,04(*)	,815	,000	1,44	4,64
			6	1,64(*)	,815	,045	,04	3,24
			7	1,48	,815	,070	-,12	3,08
			8	2,52(*)	,815	,002	,92	4,12
			9	1,52	,815	,063	-,08	3,12

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

APPENDIX P

Comparison of the grades given with the 1st instrument in terms of the reliability degrees at the basis of 7 points tolerance interval.

INSTRUMENT 1																														
Grader	1			2			3			4			5			6			7			8			9			10		
Paper	Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading		
	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp
1	55	60	1	35	55	0	55	65	0	35	55	0	68	45	0	45	55	0	45	60	0	65	55	0	25	45	0	60	73	0
2	65	55	0	60	70	0	55	60	1	60	65	1	40	65	0	65	55	0	45	65	0	88	65	0	75	68	1	78	91	0
3	25	25	1	25	35	0	45	35	0	35	55	0	65	35	0	85	45	0	55	45	0	25	35	0	25	45	0	70	50	0
4	50	50	1	70	78	0	65	70	1	60	78	0	68	60	0	70	73	1	50	70	0	73	65	0	86	50	0	75	83	0
5	45	50	1	25	35	0	45	30	0	30	25	1	50	35	0	65	85	0	35	45	0	55	25	0	45	25	0	65	55	0
6	50	50	1	60	40	0	60	35	0	60	70	0	53	55	1	75	45	0	65	60	1	50	70	0	45	60	0	83	75	0
7	45	45	1	35	25	0	45	30	0	35	25	0	55	25	0	55	35	0	45	25	0	25	30	1	75	25	0	40	50	0
8	70	70	1	50	70	0	65	50	0	75	70	1	35	60	0	65	58	1	45	70	0	70	55	0	78	70	0	83	75	0
9	40	45	1	60	50	0	60	50	0	50	60	0	60	50	0	50	58	0	55	68	0	38	70	0	58	68	0	63	86	0
10	45	45	1	25	35	0	55	25	0	25	55	0	35	25	0	45	25	0	35	25	0	60	25	0	25	35	0	40	45	1
11	45	45	1	35	45	0	45	50	1	35	45	0	45	25	0	55	45	0	75	45	0	45	25	0	25	35	0	75	60	0
12	45	45	1	35	55	0	45	30	0	25	55	0	25	35	0	55	35	0	65	35	0	45	25	0	45	35	0	60	45	0
13	70	70	1	60	73	0	70	70	1	75	70	1	78	55	0	78	70	0	50	86	0	88	70	0	86	83	1	80	91	0
14	45	45	1	25	35	0	45	50	1	25	60	0	35	25	0	35	45	0	55	30	0	45	50	1	30	25	1	55	65	0
15	45	70	0	60	55	1	50	55	1	40	55	0	50	35	0	70	55	0	45	60	0	65	60	1	35	50	0	65	70	1
16	75	82	1	70	78	0	70	75	1	55	70	0	83	55	0	58	70	0	80	91	0	86	75	0	88	83	1	91	91	1
17	45	65	0	60	50	0	60	65	1	60	60	1	55	45	0	75	68	1	65	65	1	60	70	0	73	86	0	91	83	0
18	75	78	1	70	38	1	75	75	1	70	78	0	83	60	0	70	85	0	78	91	0	78	75	1	94	86	0	83	91	0
19	75	75	1	65	75	0	65	60	1	75	83	0	94	75	0	93	75	0	71	91	0	75	70	1	91	96	1	91	83	0
20	70	70	1	75	83	0	83	60	0	70	78	0	86	75	0	85	85	1	76	96	0	98	75	0	96	88	0	75	91	0
21	75	83	0	83	91	0	45	75	0	65	78	0	83	70	0	85	75	0	78	83	1	65	83	0	98	86	0	91	83	0
22	75	83	0	75	65	0	65	50	0	78	78	1	78	65	0	75	55	0	70	83	0	85	65	0	88	68	0	73	91	0
23	70	75	1	83	80	1	95	75	0	78	86	0	83	70	0	83	70	0	75	91	0	81	70	0	88	91	1	50	91	0
24	75	83	0	60	83	0	70	83	0	75	96	0	55	60	1	88	85	1	78	60	0	85	75	0	98	91	1	60	83	0
25	75	65	0	50	70	0	60	65	1	60	68	0	65	60	1	83	90	1	91	86	1	40	70	0	86	75	0	98	96	1

Continued on page 202

26	83	81	1	91	88	1	75	96	0	78	73	1	75	70	1	98	70	0	90	91	1	90	73	0	78	83	1	91	93	1
27	75	75	1	60	78	0	65	75	0	75	68	1	86	60	0	68	85	0	86	83	1	70	70	1	86	75	0	75	83	0
28	45	55	0	60	60	1	55	65	0	65	73	0	68	50	0	78	78	1	80	65	0	55	60	1	48	55	1	88	75	0
29	45	40	1	50	76	0	55	65	0	75	63	0	68	55	0	78	58	0	96	65	0	50	60	0	58	65	1	83	65	0
30	65	65	1	75	70	1	65	55	0	65	83	0	65	60	1	83	75	0	85	83	1	78	70	0	58	75	0	83	75	0
31	65	65	1	50	65	0	55	60	1	65	78	0	75	65	0	86	70	0	78	83	1	65	50	0	75	75	1	75	83	0
32	60	60	1	65	70	1	70	75	1	75	88	0	94	75	0	96	93	1	91	91	1	78	70	0	96	78	0	96	83	0
33	45	45	1	60	65	1	55	60	1	65	60	1	73	60	0	88	50	0	70	65	1	55	55	1	65	68	1	65	83	0
34	91	91	1	93	94	1	91	96	1	94	96	1	98	94	1	98	100	1	100	88	0	98	65	0	100	80	0	90	98	0
35	91	91	1	91	94	1	75	88	0	86	91	1	80	94	0	93	95	1	75	98	0	96	90	1	98	96	1	83	96	0
36	83	91	0	75	75	1	65	83	0	86	94	0	86	91	1	93	65	0	91	91	1	78	70	0	91	96	1	91	83	0
37	78	75	1	86	83	1	91	88	1	75	83	0	94	83	0	98	85	0	95	94	1	96	93	1	98	88	0	94	98	1
38	75	75	1	86	86	1	83	96	0	75	91	0	94	88	1	96	88	0	91	98	1	94	83	0	100	80	0	94	98	1
39	70	75	1	76	75	1	65	91	0	94	94	1	94	76	0	93	86	1	70	94	0	68	83	0	98	90	0	94	100	1
40	65	55	0	78	86	0	90	75	0	91	91	1	68	90	0	98	90	0	78	94	0	78	85	1	75	88	0	94	98	1
41	78	75	1	75	78	1	75	91	0	83	83	1	94	96	1	74	98	0	80	70	0	75	80	1	86	86	1	83	96	0
42	86	91	1	78	86	0	75	88	0	94	83	0	96	94	1	70	91	0	98	75	0	94	86	0	94	94	1	83	96	0
43	86	91	1	83	86	1	83	88	1	87	63	0	98	80	0	78	96	0	96	85	0	75	86	0	100	96	1	83	96	0
44	94	94	1	86	91	1	91	96	1	86	94	0	98	90	0	96	88	0	100	91	0	95	90	1	78	96	0	83	96	0
45	86	78	0	63	94	0	91	83	0	91	94	1	96	88	0	96	98	1	98	94	1	68	90	0	88	88	1	83	88	1
46	91	94	1	94	93	1	65	83	0	83	83	1	96	91	1	96	91	1	90	91	1	83	86	1	94	88	1	75	91	0
47	78	78	1	86	91	1	83	75	0	83	83	1	94	88	1	86	94	0	76	90	0	70	73	1	98	86	0	83	88	1
48	78	86	0	78	68	0	95	55	0	73	75	1	98	75	0	98	86	0	90	83	1	73	70	1	78	73	1	73	80	1
49	86	78	0	94	97	1	75	83	0	65	83	0	94	83	0	98	86	0	96	75	0	75	70	1	100	88	0	78	91	0
50	94	94	1	94	98	1	83	71	0	83	94	0	98	94	1	100	100	1	98	94	1	80	88	0	100	86	0	94	96	1
Sum of 1			37			21			18			18			13			14			17			17			20			13
Reliability			**0,74**			**0,42**			**0,36**			**0,36**			**0,26**			**0,28**			**0,34**			**0,34**			**0,40**			**0,26**
																											Mean: 0,38			

Comp= Comparison of the grades with an interval of 7 points. 0 was given if the difference between two grades of the same grader was more than 7, 1 was given if the difference between two grades of the same grader was 7 or less than 7. The sum of the value 1 gave the reliability level when the totals were regarded as percentages since a total of 50 would give the 100%.

Comparison of the grades given with the 2nd instrument in terms of the reliability degrees at the basis of 7 points tolerance interval.

INSTRUMENT 2																													
Grader	1			2			3			4			5			6			7			8			9			10		
Paper	Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading			Grading		
	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp	1	2	Comp
1	40	41	1	44	42	1	52	50	1	43	46	1	37	44	1	50	52	1	41	41	1	52	52	1	41	40	1	48	48	1
2	59	57	1	58	57	1	59	57	1	57	59	1	50	52	1	60	63	1	54	57	1	55	62	1	50	53	1	50	50	1
3	24	24	1	24	19	1	27	29	1	35	27	0	28	29	1	32	26	1	26	27	1	32	31	1	25	21	1	32	26	1
4	66	61	1	57	59	1	56	58	1	65	57	0	60	59	1	55	68	0	66	69	1	61	62	1	56	58	1	60	68	0
5	39	40	1	31	35	1	27	33	1	34	28	1	26	29	1	24	31	1	27	33	1	37	33	1	29	31	1	26	25	1
6	52	53	1	48	44	1	37	44	1	44	42	1	50	45	1	41	43	1	44	50	1	33	40	1	33	39	1	48	44	1
7	20	21	1	20	18	1	19	21	1	22	20	1	24	25	1	21	21	1	22	22	1	27	26	1	22	20	1	24	18	1
8	48	45	1	45	43	1	39	43	1	40	39	1	36	40	1	40	38	1	43	48	1	46	41	1	31	37	1	44	47	1
9	39	38	1	41	36	1	41	45	1	51	44	1	31	39	0	37	32	1	40	43	1	44	40	1	36	43	1	35	31	1
10	21	24	1	20	22	1	26	27	1	33	29	1	31	28	1	22	24	1	24	25	1	30	31	1	24	22	1	20	22	1
11	27	27	1	34	30	1	34	34	1	29	32	1	34	29	1	36	38	1	34	37	1	38	40	1	38	32	1	40	36	1
12	19	19	1	21	17	1	26	28	1	26	22	1	24	23	1	24	22	1	24	27	1	28	25	1	17	21	1	21	23	1
13	65	61	1	58	52	1	55	54	1	65	60	1	57	55	1	65	67	1	64	63	1	67	68	1	57	63	1	62	67	1
14	29	27	1	25	24	1	25	24	1	37	34	1	26	28	1	30	23	1	37	32	1	36	35	1	25	28	1	28	23	1
15	50	51	1	58	56	1	37	50	0	51	57	1	60	59	1	63	66	1	67	61	1	57	52	1	56	61	1	53	58	1
16	61	59	1	70	67	1	76	72	1	73	68	1	68	62	1	75	70	1	80	78	1	76	73	1	63	69	1	75	71	1
17	61	61	1	44	50	1	67	67	1	57	60	1	48	56	0	48	53	1	58	58	1	55	61	1	48	54	1	66	59	1
18	77	77	1	58	54	1	60	59	1	52	57	1	51	56	1	66	63	1	73	71	1	76	69	1	52	52	1	65	65	1
19	66	68	1	62	64	1	69	65	1	74	70	1	62	67	1	70	72	1	66	66	1	63	68	1	59	64	1	67	72	1
20	80	77	1	70	75	1	80	81	1	67	70	1	71	71	1	86	80	1	77	81	1	81	77	1	73	75	1	78	83	1
21	77	77	1	70	78	0	83	86	1	76	79	1	75	74	1	79	74	1	77	73	1	80	76	1	68	75	1	70	83	0
22	58	57	1	48	48	1	70	72	1	62	68	1	59	54	1	50	59	0	64	65	1	62	60	1	57	66	0	55	65	0
23	58	58	1	64	60	1	68	67	1	62	58	1	54	59	1	70	68	1	69	73	1	75	70	1	57	60	1	70	67	1
24	68	70	1	73	66	1	62	68	1	69	74	1	74	73	1	83	75	0	71	69	1	76	67	0	64	70	1	65	75	0

Continued on page 204

25	51	51	1	66	59	1	50	57	1	58	55	1	57	68	0	65	70	1	61	60	1	59	62	1	63	66	1	71	70	1
26	63	65	1	72	70	1	71	67	1	65	68	1	63	66	1	77	74	1	72	73	1	79	70	0	62	71	0	75	72	1
27	65	67	1	70	74	1	67	69	1	73	71	1	75	70	1	78	74	1	70	71	1	73	77	1	74	75	1	71	78	1
28	57	55	1	54	60	1	63	60	1	67	60	1	58	53	1	67	62	1	52	56	1	55	62	1	60	66	1	52	58	1
29	59	59	1	51	58	1	61	61	1	56	61	1	60	61	1	61	60	1	61	63	1	61	59	1	61	60	1	59	60	1
30	64	65	1	66	69	1	70	72	1	69	65	1	73	78	1	82	80	1	74	76	1	74	75	1	77	70	1	72	72	1
31	51	54	1	53	50	1	49	47	1	48	50	1	49	48	1	50	56	1	57	54	1	63	59	1	54	58	1	50	55	1
32	56	55	1	58	57	1	60	59	1	66	58	0	54	58	1	63	63	1	72	68	1	73	66	1	58	63	1	65	70	1
33	36	43	1	36	41	1	34	38	1	55	51	1	39	44	1	40	48	0	45	44	1	48	51	1	39	40	1	42	45	1
34	89	91	1	85	86	1	91	92	1	86	85	1	86	84	1	91	95	1	87	86	1	90	85	1	93	93	1	89	91	1
35	88	91	1	92	86	1	79	81	1	86	84	1	81	84	1	90	91	1	79	86	1	90	89	1	92	96	1	90	91	1
36	75	75	1	74	76	1	70	70	1	72	74	1	81	73	0	85	79	1	83	86	1	75	80	1	85	75	0	70	78	0
37	74	76	1	73	78	1	77	77	1	62	74	0	64	71	1	76	77	1	83	80	1	74	78	1	76	77	1	78	80	1
38	81	84	1	83	81	1	94	93	1	74	82	0	89	87	1	90	86	1	84	81	1	79	86	1	89	87	1	82	84	1
39	72	78	1	71	72	1	72	69	1	80	81	1	75	75	1	88	79	0	77	80	1	86	83	1	86	79	1	86	74	0
40	82	82	1	74	80	1	84	84	1	84	88	1	75	78	1	91	87	1	75	81	1	79	84	1	88	85	1	74	82	0
41	78	79	1	77	75	1	80	79	1	79	72	1	76	79	1	85	87	1	87	79	0	79	76	1	88	83	1	87	83	1
42	83	87	1	83	86	1	80	79	1	84	90	1	87	80	1	92	80	0	89	84	1	80	82	1	93	83	0	87	85	1
43	79	82	1	80	79	1	87	87	1	84	82	1	79	75	1	93	88	1	83	84	1	79	83	1	88	87	1	86	80	1
44	83	84	1	87	79	0	85	81	1	79	78	1	84	81	1	91	87	1	84	80	1	83	85	1	97	89	0	89	82	1
45	79	81	1	82	81	1	87	83	1	78	79	1	81	76	1	94	85	0	87	84	1	79	84	1	96	90	1	91	87	1
46	74	74	1	75	76	1	80	79	1	81	79	1	81	76	1	87	82	1	77	70	1	78	78	1	84	87	1	84	81	1
47	71	71	1	82	76	1	79	75	1	75	76	1	75	75	1	88	79	0	81	81	1	76	79	1	88	76	0	86	82	1
48	77	71	1	84	80	1	73	76	1	75	74	1	73	74	1	85	85	1	81	77	1	77	80	1	81	84	1	77	79	1
49	75	78	1	77	70	1	88	80	0	88	75	0	75	84	0	82	85	1	77	75	1	76	78	1	95	86	0	80	74	1
50	92	95	1	84	80	1	91	94	1	88	83	1	83	80	1	89	86	1	87	82	1	80	83	1	89	86	1	91	92	1
Sum of 1			50			48			48			44			45			42			49			48			43			43	
Reliability			**1,00**			**0,96**			**0,96**			**0,88**			**0,90**			**0,84**			**0,98**			**0,96**			**0,86**			**0,86**	

Mean : 0,92

Comp= Comparison of the grades with an interval of 7 points. 0 was given if the difference between two grades of the same grader was more than 7, 1 was given if the difference between two grades of the same grader was 7 or less than 7. The sum of the value 1 gave the reliability level when the totals were regarded as percentages since a total of 50 would give the 100%.