İNGİLİZCE YABANCI DİL ÖĞRENCİLERİNİN YAZILI ANLATIM KAĞITLARINDAKİ DOĞRU GRAMER KULLANIMININ, DENEYİMLİ VE DENEYİMSİZ ÖĞRETMENLERİN DEĞERLENDİRMELERİNE OLAN ETKİSİ

The Impact of EFL Students' Accurate Use of Language

on

Experienced and Inexperienced Teachers'

Scoring the Written Compositions

Mehmet Duranlıoğlu

(Yüksek Lisans Tezi)

ESKİŞEHİR-2004

The Impact of EFL Students' Accurate Use of Language
on
Experienced and Inexperienced Teachers' Scoring the Written Compositions

Mehmet Duranlıoğlu

Master Thesis

Department of English Language Teaching

Advisor: Asst.. Prof. Dr. Hasan ÇEKİÇ

ESKİŞEHİR

Anadolu University, Institute of Educational Sciences

January-2004

YÜKSEK LİSANS TEZ ÖZÜ

İNGİLİZCE YABANCI DİL ÖĞRENCİLERİNİN YAZILI ANLATIM KAĞITLARINDAKİ DOĞRU GRAMER KULLANIMININ, DENEYİMLİ VE DENEYİMSİZ ÖĞRETMENLERİN DEĞERLENDİRMELERİNE OLAN ETKİSİ

Mehmet DURANLIOĞLU
İngilizce Öğretmenliği Ana Bilim Dalı
Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü, Ocak 2004
Danışman: Yard. Doç. Dr. Hasan ÇEKİÇ

Yabancı Dil Yazılı Anlatım Kağıtlarını Değerlendirme Kriteri kullanılarak, bu çalışma, öğrencilerin yazılı anlatımdaki doğru gramer kullanımları ve öğretmenlerin mesleki deneyimlerinin, yazılı anlatım kağıtlarının değerlendirilmesi üzerine bir etkisi olup olmadığını saptamayı hedeflemiştir. Bu amaçla, 10'u deneyimli ve 14'ü deneyimsiz olmak üzere toplam 24 yabancı dil öğretmeni bu çalışmaya katılmıştır.

Tüm öğretmenlerden, toplam 40 yazılı anlatım kağıdını farklı zamanlarda olmak üzere iki kez değerlendirmeleri istenmiştir. Bu toplam 40 kağıt 2 gruba ayrılmıştır: Birinci grup kağıtlar öğretmenlerin farklı zamanlarda aynı kağıda verdikleri notların tutarlı olup olmadığını tespit etmek, ikinci grup kağıtlar ise öğrencilerin doğru gramer kullanımı ve öğretmenlerin mesleki deneyimlerinin notlandırma üzerine etkisi olup olmadığını saptamak için kullanılmıştır. İkinci değerlendirme, birinci değerlendirmeden

1 ay sonra yapılmıştır. Birinci grup 20 kağıt, her iki değerlendirmede de, kağıtlar üzerinde herhangi bir değişiklik yapılmadan öğretmenlere verilmiş, ancak ikinci grup 20 kağıdın cümle düzeyindeki gramer hataları ikinci değerlendirmeden önce düzeltilerek verilmiştir.

3 öğretmen hariç, diğer tüm öğretmenlerin birinci ve ikinci değerlendirmede birinci grup 20 kağıda verdikleri notlar tutarlı bulunmuş ($r \geq 0{,}70$), ve söz konusu 3 öğretmen diğer istatistiksel analizlere dahil edilmemiştir. Bu sebeple kalan toplam 21 öğretmenin, her iki değerlendirmede, ikinci grup 20 kağıda verdikleri notlar istatistiksel olarak incelenmiştir.

Sonuç olarak, uygulanan ikili t-test analizlerine göre, öğrenciler grameri doğru kullandıkları zaman, öğretmenlerin değerlendirmelerini etkilediği ve çözümlemeli kritere göre, içerik, düzen, kelime kullanımı ve yazım kuralları gibi alt kategorilere daha fazla not vererek toplam notları artırdıkları görülmüştür. Bağımsız t-test sonuçları ise, öğretmenlerin mesleki deneyimlerinin, yazılı anlatım kağıtlarının değerlendirilmesi üzerine istatistiksel olarak anlamlı bir etkisinin olmadığını göstermiştir.

ABSTRACT

THE IMPACT OF EFL STUDENTS' ACCURATE USE OF LANGUAGE ON EXPERIENCED AND INEXPERIENCED TEACHERS' SCORING THE WRITTEN COMPOSITIONS

Mehmet DURANLIOĞLU
Department of English Language Teaching
Anadolu University Institute of Educational Sciences, January 2004
Advisor: Asst. Prof. Dr. Hasan ÇEKİÇ

This study aims at investigating whether such factors as grammatical accuracy of students' written texts and teaching experience of raters have an impact on the distribution of the scores that teachers assign through the use of ESL Composition Profile, an analytic instrument for marking student compositions. For that purpose, totally 24 language teachers, 10 of whom were experienced and 14 of whom were inexperienced, participated in the study.

All the teachers were asked to grade totally 40 essays twice at different times. These essays were divided into two sets: The first set of 20 essays was used to observe the internal consistency of the graders over time, and the second set of 20 essays served the purpose of observing whether students' accurate use of grammar and raters' experience in teaching writing influence the total scores assigned to these papers. The second grading was held one month after the first one. Before the second grading, the

sentence-level grammar errors of the second set of 20 essays were corrected, while the first set of 20 remained the same for the second grading. Both of the sets of totally 40 essays were given to the raters together for marking in both gradings.

The internal consistency results of the first set of 20 essays indicated that, except for 3 raters from the inexperienced group, all of the others were found to assign consistent scores over time ($r \geq 0,70$). Hence, the statistical analysis were applied for the remaining 21 teachers' scores that were assigned to the second set of 20 essays in the first and second gradings.

As a result of the statistical analyses, paired t-tests revealed that students' accurate use of grammar influenced the raters' sub-scores that they assigned to the sub-components of content, organization, vocabulary use and mechanics. Thus their total scores that they assigned to the papers were found to increse. On the other hand, no statistically significant effect was found on the total scores in terms of the participants' teaching experience when independent t-test was applied.

JÜRİ VE ENSTİTÜ ONAYI

Mehmet DURANLIOĞLU'nun, " The Impact of EFL Students' Accurate Use of Language on Experienced and Inexperienced Teachers' Scoring the Written Compositions " başlıklı tezi 14/01/2004 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 oy çokluğuyla kabul edilmiştir.

	Adı-Soyadı	İmza
Üye (Tez Danışmanı)	: Yrd.Doç.Dr. Hasan ÇEKİÇ	
Üye	: Doç.Dr. Handan YAVUZ	olumsuz.
Üye	: Yrd.Doç.Dr. Mine DİKDERE	
Üye	: Yrd.Doç.Dr. Aynur BOYER	
Üye	: Yrd.Doç.Dr. İlknur MAVİŞ	

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

ACKNOWLEDEMENTS

Firstly, I would like to thank to my supervisor and instructor Asst. Prof. Dr. Hasan Çekiç for his invaluable suggestions and contributions to my thesis. I certainly believe that without his encouragement, I would have never been able to complete this research.

Many thanks to Asst. Prof. Dr. Aysel Bahçe for her supportive advice in the preparation phase of the study and to my other teachers who shared their thoughts by commenting on my thesis proposal of the present study.

Also, I will always feel myself in debt to my colleagues for their invaluable support and efforts in their reading and assessing the papers for this study.

I also thank to my teachers, Assoc. Prof. Dr. Handan K. Yavuz, Asst. Prof. Dr. İlknur Maviş, Asst. Prof. Dr. Mine Dikdere and Asst. Prof. Dr. Aynur Boyer for their suggestions, advice and help.

Lastly, I can never forget my wife, my little children, my parents and my mother-in-low's patience, support and understanding during this study

TABLE OF CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

4. PRESENTATION AND ANALYSES OF DATA

CHAPTER 5

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

1.1. Introduction

In this chapter, a brief background information to the study and to the evaluation of students' writing abilities at Anadolu University, Preparatory School of Foreign Languages is given. Following this, the statement of problem and the purpose of the study are introduced. In the end, research questions are directed.

1.2. Background to the Study

1.2.1. Assessment in Language Teaching

When the process of teaching and/or learning a language is concerned, one can neither ignore the importance of assessment nor thus avoid assessing the learners' abilities. Therefore, testing and teaching can not be considered as separate terms; in contrast, they are, as Brossell (1996) sates, closely interrelated. Heaton (1975) explains the degree of this relation between testing and teaching saying that ''it is virtually impossible to work in either field without being constantly concerned with the other'' (p:1).

To make this relation clearer, we should know why assessment is necessary in language teaching/learning. In fact, the reason for the need for assessment of language abilities lies under its purpose. Johnson and Johnson (2002) regard assessment as a means of ''collecting information about the quality and quantity of a change in a student, group, class, school, teacher or administrator'' (p:2). In other words, for instance, assessing students' certain abilities, teachers can find out whether the students have developed their particular abilities; that is, teachers can determine how much the students have achieved the objectives of a certain course.

Therefore, without being aware of the outcomes of a particular instruction, it would not be reasonable to go on providing students with further instruction since we don't know yet whether the students have processed the previous instruction or not. Here, it would be better to explain the importance of testing in the teaching and learning process with a simile from Heaton (1975):

> ''just as it is necessary for the doctor first to diagnose his patient's illness, so it is equally necessary for the teacher to diagnose his student's weaknesses and difficulties. Unless the teacher is able to identify and analyse the errors a student makes in handling the target language, he will be in no position to render any assistance at all through appropriate anticipation, remedial work and additional practice'' (p:2).

In general, in language teaching, students are expected to gain the four basic skills at the end of a treatment. Among these four skills, two of them are receptive skills such as reading and listening, and the other two are productive skills such as speaking and writing. In terms of testing of reading and listening abilities, students can be given, for example, a multiple-choice test which makes the grading process quite easy since such tests enable teachers only to count the correct and incorrect responses. On the contrary, when the assessment of productive abilities of students are concerned, the grading process can become rather difficult. For instance, when students are asked to write an essay or a paragraph in a writing test, teachers will have to deal with the difficult task of grading each written text. Therefore, Camp (1996) believes that assessment of writing is not merely a test instrument or a method of scoring since the quality of assessment is very important in terms of determining whether the information it yields can be trusted.

1.2.2. Assessment of Writing Ability

In order for teachers to elicit some behaviour from learners and to assess their overall ability, Hughes (1989) suggests *indirect assessment* techniques such as multiple-choice tests or cloze tests. However, these tests are criticized to test only recognition knowledge. That is, these tests only enable us to determine whether the student is aware of, for instance, a particular grammatical form when he chooses the correct option in a multiple-choice test. However, we will have no idea about whether the student can put this knowledge into practice. Hence, we can solely test if the student can recognize that particular form or not.

Therefore, for some particular abilities, as in the case of assessment of writing ability, Kroll (1991) claims the best way for assessment of writing to be to use direct tests of writing. In direct tests of writing, students could be asked to compose a text, for instance, in an essay format or in a paragraph form. However, if we want to test students' writing ability in an indirect way, then we will not be able to ask students to compose a text. Instead, we will have to prepare a test that includes several items, for instance, some of which would ask students to put jumbled paragraphs into the correct order so as to investigate their organizational skills. Meanwhile, some of the items would ask students to correct grammar errors found in the given sentences to test their grammatical abilities. Hence, Hughes (1989) explains the underlying reason for prefering the use of direct tests of writing to indirect tests stating,

> ''even professional testing institutions are unable to construct indirect tests which measure writing ability accurately. And if in fact satisfactory accuracy were a real possibility, considerations of backwash and ease of construction would still argue for the direct testing of writing within teaching institutions'' (p:75).

If a direct test of writing ability is of concern, then students can be asked to produce a written text in a paragraph form or in an essay format, which is also called performance testing (Gronlund, 1988). One main advantage of a performance test is that the construction of a performance test, such as an essay test, is quite easy as compared to that of any other type of tests mentioned above. Another advantage of such a test is that an essay test, for instance, helps to see the whole picture of students' writing abilities since they are expected to express themselves through an accurately-written work with a sufficient amount of vocabulary and in an organized manner. In other words, when students are asked to write an essay, they have to put their knowledge into practice which indirect tests, such as recognition tests as mentioned above, fail to do so.

However, what makes an essay test difficult and somehow problematic is its scoring (Baker, 1989). As one major problem with the assessment of written texts, Gay (1985) warns raters about the scoring procedure of an essay test since he as well as many other researchers (Henning, 1986; Harrison, 1983; Kubiszyn and Borich, 1990) consider it to be potentially a very subjective process that involves low scorer reliability. Despite the probability of such a problem, Gay (1985) also reminds that ''the degree of subjectivity can be considerably minimized by careful planning and scoring'' (p:226). That is, a scoring procedure which is most appropriate, for instance, to the raters and/or to the purpose of testing should be followed when scoring the written texts as explained in the following part of this chapter.

1.2.3. Approaches to Scoring Essay Tests

As mentioned before, objective assessment of students' performance on a piece of written work can be considered to be a highly complex task and a time-consuming activity on the part of the teacher. So as to achieve this difficult task and to increase the level of objectivity, as the very first step, Brown (1996) suggests that teachers should decide on which approach is better for their assessment purposes, such as an analytic approach, ''in which the teachers rate various aspects of each student's language production seperately'' (p:61) or a holistic approach, ''in which the teachers use a single general scale to give a single global rating for each student's language production'' (p:61).

In addition, Elbow (1996) mentions a third approach, multiple-trait scoring, which allows holistic grading. Ferris and Hedgcock (1998) explains that the goal of this approach is ''to develop criteria for successful writing on a given topic and/or in a selected genre so that teachers and writers alike can focus on a narrow range of textual aspects or traits'' (p:242). That is, the raters focus on a certain feature of student texts, and they assign their scores holistically keeping that particular feature in mind as a basis for their judgement. It would be easier to better understand this approach if we give an example from Omaggio (1986):

> ‘‘if a student’s essay was designed to persuade others to adopt his point of view on an issue, the grade might be based on the number of reasons given in the support of his argument, the elaboration of those reasons, the authorities to which he appealed, and other features of the discourse related to the function of persuation’’ (p:268).

Though there are such different ways to make judgements about a paper, Ferris and Hedgcock (1998) state that these three approaches are only scoring options among which teachers can choose to apply. The researchers also note that these approaches should not be considered as preferred or recommended methods. Here, among these three approaches, only two of them, holistic and analytic approaches, will be mentioned in terms of their pros and cons because of two basic reasons: First, this study requires an analytic scoring procedure to be used (see chapter 3). Next, the previous research in the field mostly used holistic and analytic approaches for assessment purposes (see chapter 2).

1.2.3.1. Holistic Scoring

Holistic approach to scoring is often referred to as *global approach* (Gay, 1985) or as *impressionistic approach* to scoring (Hughes, 1989). This approach has always been exposed to criticism mostly because it results in more subjective and less reliable scores (Gay, 1985; Hamp-Lyons, 1995, 1996; Connor-Linton, 1995; Ruetten, 1984). The reason for the subjectivity is that the scoring procedure ‘‘involves the assignment of a single score to a piece of writing on the basis of an overall impression of it’’ (Hughes, 1989; p:86). That is, the scores assigned to each paper depend on what constitute the raters’ overall impression of a paper. Therefore, impressionistic approach to scoring makes it necessary for raters to look at the same points in a paper in the process of marking.

In order to ease the raters’ work and to increase their consistency in the distribution of their marks, a scoring guide may accompany holistic scoring (see appendix C). The descriptions in the holistic scoring guide ‘‘imply a pattern of development common to all language learners. They assume that a particular level of grammatical ability will always be associated with a particular level of lexical ability’’ (Hughes, 1989; p:91). That is, raters do not give separate marks to the grammatical ability of a student nor to the use of vocabulary, but they assign a single score that represents the descriptors in the holistic scoring guide.

On the other hand, in spite of such a potential subjectivity and thus a possible low reliability of the assigned scores, holistic scoring is favoured by teachers and institutions because of two major reasons. First of all, raters can assign their marks very fast since the approach requires the assignment of a single score. Hughes (1989) first gives an example for how rapid raters can be and then mentions another related advantage, stating that ‘‘Experienced scorers can judge a one-page piece of writing in just a couple of minutes or even less.... This means that it is possible for each piece of work to be scored more than once, which is fortunate, since it is also necessary!’’ (p:86). Another reason for the popularity of holistic scoring concerns especially the institutions in that its cost is less than other approaches as the rating process doesn’t take much time for graders.

Despite the availability of a scoring guide in holistic approach and despite its advantages such as its being faster and less costly, Hughes (1989) considers this approach to be highly questionable in terms of raters’ potential inconsistency in assigning their single scores. For that reason, the desire to have the raters look at the same features of a written text and thus to obtain more consistently-assigned scores leads us to focus on the analytic approach to assessment of writing.

1.2.3.2. Analytic Scoring

One way to decrease subjectivity is to provide raters with detailed criteria which allow them to focus their attention on some common standards. Therefore, analytical procedures appeared due to the need for more objective scores that impressionistic scoring approach mostly fails to produce. Analytic scoring is commonly defined by researchers (Hughes, 1989; Kroll, 1991) as the assignment of a separate score for each of a certain number of features found in a written text.

There are several advantages of such a procedure of scoring where teachers have the chance to reach a total score through some subscores and where students are able to see what constitutes this total score. Hughes (1989) mentions these advantages in terms of both students and teachers: ''First, it disposes of the problem of uneven development of subskills in individuals. Secondly, scorers are compelled to consider aspects of performance which they might otherwise ignore. And thirdly, the very fact that the scorer has to give a number of scores will tend to make the scoring more reliable'' (p:94). That is to say, teachers will not ignore students' ability to use language accurately, nor will they forget about the appropriate use of vocabulary while they are assigning their marks. In this way, students will also be able to see in what areas in writing skills they have to develop themselves.

Therefore, in analytic scoring, raters are supposed to refer to an analytic scoring guide while assigning their scores. A well-known analytic scoring guide, ESL Composition Profile, is developed by Jacobs et al (1981). The profile consists of five features such as content, organization, vocabulary, language use and mechanics, each of which has different weights of scores (for details, see appendix A). Using such a profile, raters assign separate sub-scores to each of these sub-components to reach the total score of a paper. Omaggio (1986) points out one instructional advantage of scoring such features separately; that is, ''more precise diagnostic feedback can be provided to the student'' (p:268). In addition, with respect to the reliability of the scores, Hughes (1989) mentions another advantage of assigning separate scores stating that ''the mere fact of having (in this case) five 'shots' at assessing the student's performance should lead to greater reliability'' (p:94).

However, although analytic scoring makes it compulsory to use an analytic scoring guide, Hughes (1989) is concerned about whether scorers can judge each of these aspects in the guide independently of the others (which is called *halo effect*). Hughes (1899) first explains the reason for such a possibility and finally suggests a solution:

> ''Concentration on the different aspects may divert attention from the overall effect of the piece of writing. Inasmuch as the whole is often greater than the sum of its parts, a composite score may be very

> reliable but not valid.... To guard against this, an additional, impressionistic score on each composition is sometimes required of scores, with significant discrepancies between this and the analytic total being investigated'' (p:94).

Perkins (1983) also holds a similar view with Hughes (1989) and draws attention to this potential problem stating that ''the features to be analyzed are isolated from context and are scored separately. Discourse analysis and good sense tell us that a written or spoken text is more than the sum of its parts'' (p:657).

In addition to this potential problem within analytic scoring approach, Hughes (1989) is of the opinion that ''the main disadvantage of the analytic method is the time that it takes. Even with practice, scoring will take longer than with the holistic method'' (p:94). That's why, analytic scoring is considered to be more costly, which leads some institutions to use holistic approach especially in the case of a large number of students to be assessed.

Holistic or analytic? Despite the advantages and disadvantages of both of the scoring approaches, it is true, as mentioned before, that ''we present these approaches as scoring options from which teachers can select, rather than as preferred or prescribed methods'' (Ferris and Hedgcock, 1998; p:232). Hughes (1989) gives an example for a possible rationale to choose between the two methods; ''the choice between holistic and analytic scoring depends in part on the purpose of the testing. If diagnostic information is required, then analytic scoring is essential.'' (p:97).

1.3. Background to the Testing System at Anadolu University, Preparatory School of Foreign Languages

Students at Anadolu University, Preparatory School of Foreign Languages are placed at different language levels, from beginner to advanced levels, through a placement test given at the beginning of an academic year. All through the academic year students take several courses, including a separate writing course, and they take different exams for each course, such as quizes and midterm exams. At the end of the academic year, all students are required to take the same achievement test which is made up of three parts. The first part comprises a multiple-choice test that covers grammar, listening and reading. In the second part, students are required to accomplish an oral task which is graded by two teachers. Finally, in the last part of the achievement test, students are asked to write a five-paragraph essay. All the essays written by students at different levels are graded by a committee of about 40 teachers in such a way that one paper is to be scored by two graders. The two marks assigned to a paper are averaged unless there is more than a 10-point difference between the two scorers. In the case of an 11-point difference or more, a third grader is consulted.

In order to determine whether a student passes the preparatory class or fails to achieve the passing grade, the average of all the marks that the students get from all the courses throughout the academic year and the marks received at the achievement test are taken into account in order to calculate the averaged-total score of a student. This total average has to be 70 or above, as determined by the administration, so that the students can be considered as successful enough to meet the overall objectives of the program. What is more and the most important of all, if the achievement test score of a student is below 70, then he or she is considered to fail the program whatever the average grade obtained throughout the academic year is.

1.4. Problem And Purpose

Students consider this achievement test to be the most important and difficult part of the preparatory class. To say the truth, they have the right to be afraid of this exam since it is not officially important how high their in-year avarage grade is if their end-of-year grade is not 70 or above. Therefore, this fear and reality place more importance on teachers' evaluation of students' productive performance in the achievement test. Hence, teachers' judgements of the papers written both throughout the academic year and especially in the achievement test are of great importance in terms of the students' failure or success.

This fact led us to focus on how trustworthy teachers' scores are that they assign to the students' written texts in the achievement test. The reason for focusing on teachers in this study is the fact that the preparatory school at Anadolu University employs teachers with various teaching backgrounds, such as native English teachers who are in fact very few in number and therefore were not included in this study, non-native inexperienced teachers most of whose language teaching experience varies between one or two years or a little above, and those non-native experienced teachers who have taught English language for approximately ten years or so.

As stated earlier in the previous part, despite using even the same analytic scoring criteria, different teachers' gradings for the same student's composition can result in inconsistent scores which, in turn, lead to high subjectivity. Such a case is likely to be encountered due to the readers' characteristics. There is a considerable number of studies carried out to investigate whether differences among raters' background, such as nationality and/or teaching experience, have effects on composition marking. The results of these studies involving the effects of teachers' teaching experience do not seem to be consistent (see chapter 2). Furthermore, with respect to experience in teaching a language, Hamp-Lyons and Kroll (1996) hold the belief that the assessment of writing requires teachers to be skilled enough to cope with the complex process of grading papers. Therefore, here in our own context, it would be worth examining scientifically whether experienced and inexperienced teachers at the preparatory school of Anadolu University, all being non-native, assign their scores consistently to the papers written at an end-of-year exam, which, to a large extend, determines the students' fate in the school.

In addition, there is another factor, which hasn't been addressed adequately in the field: How effective are certain qualities of students' written texts on raters' scores? That is, for instance, does students' accurate use of grammar affect raters' scores that they assign to papers? To understand the need for and the importance of exploring such a factor, it would be better to recall what testing of writing skills involves.

In general sense, testing allows us to determine, at a time, whether our students have achieved our pre-determined objectives of a particular course or not. This is also true for the testing of our students' writing skills in the way as follows; students who are to pass the preparatory class at Anadolu University at the end of an academic year are expected to express themselves effectively and efficiently in their written works so that they can successfully convey their message to others using the language they have been learning. That is, for the students to be successful in written communication, they are expected to:

- put forward ideas rich in content and relevant to the present context
- have a wide range of vocabulary so that they can present their ideas effectively
- produce structurally correct sentences so as to have meaningful statements
- obey the rules of writing such as punctuation and spelling
- present their ideas in an organized manner (i.e. students should be able to choose an appropriate genre and a suitable rhetorical pattern, include a clear main idea with a sufficient amount of supportive evidence or examples and pay attention to the use of transitions that help coherence and cohesion in the text).

In other words, a well written text is expected to involve a rich content with well-organized ideas, a wide range of vocabulary and grammatically correct sentences. Therefore, a badly-organized text, for instance, will certainly cause readers or raters to have problems in following the ideas. Similarly, wrong word-choice in a text may also confuse readers or might even lead them to misunderstand the ideas presented especially if the context doesn't help. Once more, grammar errors found in a text may irritate the reader as well.

Eventually, when we assess a student's paper, using either a holistic scale or even an analytic one, we should judge the paper considering what qualities a good paper should have. That is, a rater should keep it in mind that a good paper has to include such certain features as mentioned above which are important for a successful written communication.

However, in almost all preparatory schools in Turkey, as in that of Anadolu University, students at different proficiency levels (from beginners to advanced) take a writing test as part of an end-of-year achievement exam. In this writing test, students are asked to write an essay. Due to the examinees at various levels, it is always highly likely that we end up with papers that have varying qualities in terms of the five features mentioned above (content, organization, vocabulary use, language use and mechanics). For instance, one can expect an upper-level student to be quite successful in each of these five features. However, it is also possible that s/he may fail, say, to organize his/her ideas even though he has mastered the English language grammar. Similarly, a student at a lower level might present his/her ideas rich in content in a well-organized manner yet may still have problems in the word-choice and/or grammar errors in his/her sentences. These are all possible illustrations that we may encounter when we assess papers. Thus when using an analytic scoring scale, as in the case of this study, one should never forget that each of the five components (content, organization, vocabulary use, language use and mechanics) refers to a different feature of a text.

Hence, in this study, we also wanted to find out if the scores of teachers, either experienced or inexperienced in teaching writing, change when they grade the same paper twice: In the first grading, they are requested to mark the original papers and in the second grading they are asked to mark the same paper, yet the sentence-level grammar errors of which are corrected. This would help us to see if there is an increase in the total scores between the two gradings or not. If so, it would also enable us to see in which of the other four components a change occurs.

Therefore, the purpose of this study was to find an answer to the following question:

Do factors such as years of experience in teaching writing and the quality of students' language use in their texts have an influence on the teachers' composition grading through the use of analytic writing criteria?

Once the importance of the writing skill in any language teaching syllabus is concerned, it is clear that a careful, free-from subjectivity type of an assessment of this skill is necessary in terms of fairness to the students.

For the assessment of final exams over the last few years in the school, a scoring rubric has been in use (see appendix F). This rubric can be said to look analytic since it requires teachers to focus on certain aspects of a written text, such as content, organization and language use. However, the descriptors are so general and wide that the scoring rubric may result in inconsistent scores. This scoring rubric was previously developed by Oruç (1999), and in a later study by Polat (2003) it was found to cause teachers to end up with inconsistent scores in terms of both inter- and intra-rater reliability coefficients. Therefore, with the use of a well-known analytic scoring guide, ESL Composition Profile developed by Jacobs et al (1981), where the descriptors of each component are clearer, the results of this study will also help comment on the intra-rater reliability coefficients of the scores assigned by teachers themselves and on the inter-rater reliability coefficients of the scores assigned by experienced and inexperienced teachers.

Hence, this study not only will provide insights into the writing assessment at the preparatory school in particular, but also, in general, will enable us to compare with the results of other similar studies carried out in the field, in terms of the effects of such factors mentioned above.

1.5. Research Questions

The present study is designed in order to investigate whether students' accurate use of grammar in their written essays will shape the experienced and inexperienced teachers' re-assessing the overall writing quality of the original papers when their sentence-level grammar errors are corrected. Furthermore, it was also aimed to find out whether the scores assigned by two groups of teachers were consistent in-between or not. For that purpose, the following research questions were asked:

1. When compared with the total scores assigned by the inexperienced teachers in the first grading, do the total scores assigned by the same group of teachers increase in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?

    1.1.If so, in which of the four sub-components (content, organization, vocabulary, and mechanics) does a significant change occur between the first and second gradings?

2. When compared to the total scores assigned by the experienced teachers in the first grading, do the total scores assigned by the same group of teachers increase in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?

2.1. If so, in which of the four sub-components (content, organization, vocabulary, and mechanics) does a significant change occur between the first and second gradings?

3. Is there a significant difference between the two groups of teachers' total scores assigned to the corrected-version essays in the second grading?

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

This chapter begins with the presentation of a measurement theory with respect to assessment of writing ability. In the light of the theory, how reliable a rater can be in assigning their scores to the written papers is discussed. Following this, some factors that are likely to affect raters' judgements in the assessment process and thus the reliability of the assigned scores are presented. In the end, several studies that investigated the possible effects of sentence-level grammar errors on graders' scores are summarized in detail.

2.2. Classical True Score Measurement Theory

Whichever approach to scoring, either holistic or analytic, is used in the grading process, the mere concern of teachers should be to assign the scores that the students really deserve. However, it is not an easy job for the teachers to decide what mark a paper really has to receive. Sometimes, many teachers suffer from spending hours while marking their students' written texts just to be fair in their scores. However invaluable efforts teachers show in the marking process, the scores assigned to student papers are claimed, by many researchers, to consist of measurement errors (Hughes, 1989; Bachman, 1990; Thorndike et al, 1991; Brown, 1988, 1996; Nitko, 1996; White et al, 1996). Due to these measurement errors, students may not receive the actual scores that they in fact deserve. Therefore, it would be better here to define what constitutes a student's score: A student's score - also called *obtained score* (Nitko, 1996) or *observed score* (Brown, 1996) – is made up of two parts: a *true score* and an *error score*. Nitko's (1996) definitions of these two parts are clear enough to understand what a true score and an error score are:

> ''The sum of these two scores (true score and error score) equals the obtained score. Whenever we assess a student, we want to know the student's true score. However, we are always 'stuck' with the obtained score.... If you could quantify the amount of error in a student's obtained score, you would have the **error score**. (Often the error score is referred to as *error of measurement*). The **true score** is the remaining portion of the observed score and contains no measurement error'' (p:63).

In terms of classical true score measurement theory, Bachman (1990) further discusses two assumptions about the relationship between the observed scores and some factors that cause error scores:

> ''The first assumption of this model states that an observed score on a test comprises two factors or components: a *true score* that is due to an individual's level of ability and an *error score,* that is due to factors other than the ability being tested.... A second set of assumptions has to do with the relationship between true and error scores. Essentially, these assumptions state that error scores are unsystematic, or random, and are uncorrelated with true scores. Without these assumptions it would not be possible to distinguish true scores from error scores. These assumptions constitute the CTS model's definition of measurement error as that variation in a set of test scores that is unsystematic or random'' (p:167).

Brown (1988) considers *true score* in terms of reliability. He strongly believes that a true score is impossible to achieve. Therefore, he suggests estimating how close the students' scores are to the ideal true scores. In this way, the *reliability coefficient,* which is one way of looking at the consistency of test scores, can help researchers ''estimate the percentage of variation in the observed scores (the scores that are actually obtained on a test) that can be attributed to true score variation'' (p:99).

The fact that raters are not likely to be consistent in their scoring the papers and that they fail to reach true scores lead to a fundamental concern of reliability.

2.3. Reliability in Language Testing

In the field of language testing, in general, the term ``*reliability* refers to the consistency of assessment scores'' (Nitko, 1996; p:62). To understand the term better, a clearer explanation is provided by Thorndike et al (1991), ``when we ask about a test's reliability, we are asking not what it measures, but instead how accurately it measures whatever it does measure. What is the precision of the resulting score? How accurately will the score be reproduced if we measure the individual again?'' (p:91). The term reliability, therefore, involves the students whose performances are assessed, the testing material which is used for assessment, and the raters who aim and try to assign the fair scores that students really deserve.

2.4. Rater Reliability in Composition Assessment

However, when the issue of reliability is taken into consideration in terms of a direct test of writing, say testing students' writing abilities by asking them to write essays, reliability is not thought to be concerned with the administration of a test or with the writers' performance on the task. Instead, reliability calls for the consistency of the raters' judgements on their scores as explained by Shale (1996), ``with essay tests, the presence of a rater (or raters) has clouded the issue, and there is often confusion about what attribute *reliability* refers to .... The term refers to a measure (or, alternatively, a measurement) – a measure being a procedure for producing a score for each examinee'' (p:79). In addition, he believes that reaching such a definition of reliability in terms of essay tests has some advantages since it makes it possible to examine the rater's individual scores or the consistency across raters.

Furthermore, Hughes (1989), who also calls the term *reliability* in essay tests as *consistency* of a rater's scores, claims that perfect consistency of a single rater or between/among raters can not be expected in the case of a performance test, such as compositions or interviews. ``Such subjective tests will not have reliability coefficients of 1! ... While the perfect reliability of objective tests is not obtainable in subjective tests, there are ways of making it sufficiently high for test results to be valuable. It is possible, for instance, to obtain scorer reliability coefficient of over 0,9 for the scoring of compositions'' (p:36).

Therefore, in such performance tests, which result in subjectively assigned scores, it is inevitable to have measurement errors in scores, and Bachman (1990) holds the view that the source of possible errors in measurement results from the inconsistent ratings. Hence, he makes the point clearer stating,

> ''in the case of a single rater, we need to be concerned about the consistency within that individual's ratings, or with intra-rater reliability. When there are several different raters, we want to examine the consistency across raters, or inter-rater reliability. In both cases, the primary causes of inconsistency will be either the application of different rating criteria to different samples or the inconsistent application of the rating criteria to different samples'' (p:178).

2.4.1. Inter-Rater Reliability

In order for students to receive scores that are closer to their true scores, Hamp-Lyons (1990) suggests multiple scoring since multiple judgements result in a final score which the researcher considers to be closer to a 'true' score than any single judgement. However, inviting more than one rater to assign individual scores to each paper is not enough since there is a high possibility that raters can be looking at different things. Heaten (1975), referring to previous research, expresses the fact that raters are extremely unreliable due to ''their failure to agree with colleagues on the relative merits of a student's composition'' (p:134).

Hamp-Lyons (1990) explains a sample case where such a disagreement between raters on the same paper may occur by giving a striking example:

> 'Reader A may assign a score of 2 on a six-point scale, for example, while reader F may assign a score of 5. If the two scores are averaged, a score of 3,5 will be reported. Yet we can quickly see that 3,5 bears no resemblance to the actual scores assigned. What often happens in these cases is that a third reader brought in. The three scores can be handled in different ways: all three may be averaged, or only the two closest scores may be averaged. Let us say that in the foregoing example the third reader, Reader P gives a score of 3: the reported score may be 2,5 (average of two closest scores) or 3 (average of all three scores, rounded to nearest whole number). In either case, how do we know the result is in fact a 'true' score? In both cases, Reader F's

> score is effectively discounted; yet reader F is a trained reader whose scores on other essays are treated as valid. Indeed, Reader F may be the third reader of some other essays over which two readers have disagreed'' (p:80).

The researcher considers such a case as a problem that still hasn't been solved by researchers and goes on explaining the hidden reason: ''We do not share a construct of writing quality. It seems that writing quality is not a simple construct, and until we arrive at scoring procedures that respect that fact, we will continue to have both validity and reliability problems'' (p:80).

2.4.2. Intra-Rater Reliability

Although raters' internal consistency is not mentioned much in the literature of writing assessment, it is important to see whether a rater assigns similar scores if he marks the same paper over time (Heaton, 1975). Brown (1988) relates *intra-rater reliaiblity* closely to *test-retest* type and defines it as follows, ''two sets of scores are produced by the same rater on two separate occasions for the same group of students'' (p:100).

Whether a rater uses holistic or analytic approaches to assessment of writing ability, he is expected to apply the same set of criteria and thereby to be consistent in his rating. If this is the case, then ''this will yield a reliable set of ratings. That is, assuming that the language samples themselves are error free, individuals' true scores will be determined to a large extent by the set of criteria by which their performance is judged'' (p:179).

However, this type of reliability is significant mostly for research purposes, and it is usually ignored in a typical writing assessment procedure in a language program. This is basically due to its lack of practicality; that is, no administration requires their raters to re-judge the papers (after a certain period of time) that they scored previously.

2.5. Factors That Affect True-Scores And Thus the Reliability of Obtained Scores

The subjective nature of writing assessment has led much of the research in the field to the process of assessing learners' writing ability. This tendency has been due to a wide range of factors that are present in the process of writing assessment. Therefore, in many studies, these factors have been investigated whether they have any effects on the effects on the grading process and hence on the objective marks.

These studies have mostly been concerned with the demographic and background characteristics of raters (such as their age, experience, race or gender), with the type of criteria itself used for assessment purposes (either as holistic or analytic), and with the poor or high quality of the learners' written work (in terms of, for instance, content, grammar, use of vocabulary, or the organization of ideas). In the following parts, a considerable amount of studies that focus on these issues will be summarized in detail.

2.5.1. Factors Related to Raters

Among the other factors such as the scoring guide used (holistic or analytic) and the quality of the text being graded (poor or good), Brown (1996) considers teachers to be the most responsible for the unfairly-assigned marks,

> ''Teachers would generally like to ensure that their personal feelings do not interfere with fair assessment of the students or bias the assignment of scores. The aim in maximizing objectivity is to give each student an equal chance to do well .... The problem is not with the scale itself but rather with the person who would inevitably assign the scores on such a test. Can any person ever be completely objective when assigning such ratings? Of course not.... Such tests ultimately require someone to use some scale to rate the written or spoken language that the students produce. The results must eventually be rated by some scorer, and there is always a threat to objectivity when such tests are used. The problem is not whether the test is objective but rather the degree of subjectivity that the teachers are willing to accept'' (p:32).

Brown (1996) mentions his experience in one composition scoring situation, which is striking and hence is worth mentioning: ''I found that ten language teachers made numerous mistakes in adding five two-digit subscores to find each student's total score. These mistakes affected 20% of the compositions, and no teacher (myself included) was immune'' (p:35). On the other hand, even though scoring mistakes are undesirable in terms of fairness to the students, the researcher believes that ``... any teacher who has served as a scorer in a pressure-filled testing situation has made such scoring mistakes'' (p:35).

2.5.2. Factors Related to Approaches to Scoring

An approach that is most suitable for testing purposes and for raters as well should be applied while assessing students' papers. Holistic approach, for instance, requires raters to be experienced enough to be consistent across papers while assigning their scores. On the other hand, in the case of an analytic approach to scoring, categories of language to be judged should be determined prior to marking process if raters don't have an appropriate scoring scale ready-in-hand. Brown (1996) explains the need to define language categories stating that ``because such decisions are often very different from course to course and from program to program, decisions about which categories of language to rate should most often rest with the teachers who are involved in the teaching process'' (p:61).

Following the determination of the categories of language to rate, Brown (1996) considers it necessary, as the next step, to provide the clear definitions of the points on the scales for each category since he believes that ``written descriptions of the kinds of language that would be expected at each score level will help ... to ensure that the judgements of the scorers are relatively consistent within and across categories and that the scores will be relatively easy to assign and interpret'' (p:62).

However, no matter how explicitly the descriptions are written in the scoring guide, sometimes, raters may not use the given criteria effectively (Lumley, 2002). This may be due to the descriptions that do not satisfy them while grading certain features of particular papers. Eventually, the raters may end up with an indeterminate grading process though they are supposed to follow the scoring guide provided. In such a case, it is likely to end up with a low rater reliability and thus unreliable scores.

2.5.3. Factors Related to the Student Texts

In an essay test, students may be given only one topic to write about, or they may be asked to choose one among some topics. If the test provides students with the opportunity to choose one from the given topics, this will result in many texts written on different topics. Heaton (1975) considers such a case to involve some hidden dangers in terms of the scoring process since one paper on a particular topic will have a different content from the content of another paper on a different topic. However, ''if the composition test is intended primarily for assessment purposes, it is advisable not to allow for any choice of composition items to be answered. Examination scripts written on the same topic give the marker a common basis for comparison and evaluation'' (p:128).

Another factor is related to the quality of some aspects of students' texts. Some raters may regard a student's grammatical competence as their primary concern (Sweedler-Brown, 1993) while some others may be looking at the organization of the presented ideas while distributing their scores even though they may be using an analytic scoring guide (Santos, 1988). Another aspect could be the quality and variety of the vocabulary use of the students which might influence the graders' impression about a paper (Engber, 1995). What's more, teachers may react differently to the cultural rhetorical patterns provided in students' papers (Kobayashi and Rinnert, 1996). One more aspect of a student's paper can be the handwriting and general appearance of the paper. Some teachers may pay more attention to neatness more than the others while they are grading students' papers (Henning, 1986; Eames and Loewenthal, 2001). This is probably because these raters themselves produce neat papers while they are writing their own texts.

All these factors might play a role to a certain extend in the distribution of scores, and this may change from teacher to teacher. Certainly, this is not desirable on the part of the students when we consider the most important outcome of any testing procedure, which is the backwash effect of a test.

2.6. Backwash Effect of a Writing Test

Backwash effect can be considered as the effect of testing on teaching and learning process (Madsen, 1983). Backwash effect (sometimes called washback effect) occurs in two ways; beneficial (or positive) or harmful (or negative). It is important that the course objectives and the purpose of the test have to be interrelated. If so, then the backwash effect of the testing material can be regarded as beneficial. That is, the content of the material used for testing matches the course objectives. If the opposite is the case, then the effect is considered to be harmful.

This is also true for the testing of writing skills. To illustrate, if students believe that they receive higher marks when they write grammatically error-free papers, then, -the next time they are asked to compose a text- they will not pay so much attention to the other textual aspects such as the organization or the quality of their ideas. Therefore, in terms of the backwash effect of a writing test, it is important to avoid such a case so as to achieve a beneficial backwash effect. Similarly, if students become aware of the fact that they receive the marks that they really deserve, then ''composition can be used to provide not only a high motivation for writing but also an excellent backwash effect on teaching'' (Heaton, 1975; p:134).

In addition, the use of analytic approach to scoring will also improve the backwash effect of a writing test since students will also be able to recognize in what areas they are successful and in what areas they need encouragement thanks to the diagnostic information that analytic scoring yields.

As can be seen, all the terms mentioned so far including measurement errors, true scores, reliability and backwash effect, are so closely interrelated that if we have errors in our measures then we can neither obtain the true scores nor thus claim that our scores are reliable, which in turn calls for the backwash effect of that writing assessment. In the next part, some studies carried out to investigate the possible effect of grammar errors on the judgements of raters with different backgrounds are summarized in detail.

2.7. Research on the Factors That Affect Scoring of Written Texts

Teaching writing skill and the assessment of it have always been a concern for researchers and an issue to be discussed by the foreign and second language teachers. What makes writing skill so concerned, as mentioned before, is its being a productive skill. Since it has strongly been suggested that there should be a direct test of writing skill (Hughes, 1989), teachers always face a complex issue of the assessment of writing. Despite this difficulty, the researcher, on the other hand, notes that a more beneficial backwash effect of the testing material will be achieved when students take direct tests of writing.

If a direct test of writing is concerned, then it means that the students will be asked to produce a piece of writing, which will be graded either using a holistic method of scoring or an analytic one. As for both of the methods, the former implicitly requires the graders to keep certain features of papers in mind, such as grammatical correctness, while the latter method explicitly requires formal accuracy to be taken into consideration while assessing a paper. Therefore, if teachers ignore grammar errors in an EFL paper, then its backwash effect is likely to be harmful, which may in turn cause the students not to pay enough attention to presenting grammatically error-free products the next time. In contrast, if students recognize that they receive higher marks when their papers do not have errors in terms of grammar, then they will not be careful about, say, their presentation or organization of their ideas or their choice and use of appropriate vocabulary.

Therefore, in the literature, most of the studies tried to find answers to whether, teachers are reliable in their scoring during the grading process of students' written papers. That is, they tried to investigate if some factors such as certain features of a written text and characteristics of raters affect judgements or not. For instance, several studies tried to seek answers to whether raters' overall scores increase when they meet papers which show a high grammatical competence and whether they ignore the other features of the papers, such as content, organization and vocabulary.

As it is true that teachers, to a certain extend, should assign a score to the writing quality of a paper focusing not only on the content and the rhetorical patterns but also on an accurate use of formal structures, (then) graders are to scrutinize the papers so as to score them effectively and reliably. There are several factors that affect graders' scores; some of the factors, as mentioned before, are related to the graders themselves, and some are concerned with the papers being graded. The grader-related factors mainly include the teaching experience, the gender, the age, the race and the current psychological state at the time of the grading. Among the paper-related factors might be the (legible or illegible) handwriting of the student, the neatness of the paper, and the (high or poor) quality of the grammar or vocabulary, the range of ideas and the (effective or ineffective) organization of the ideas.

There are several studies that search for whether teachers with varying backgrounds are reliable in their scoring in terms of the possible effects of the quality of the students' competence in formal structures. In one study, Sweedler-Brown (1993) investigated whether experienced English writing instructors who are not yet trained to teach English as a second language are more influenced by grammatical and syntactical features of English or proficiency in the broader rhetorical features of writing when they holistically grade ESL essays. She conducted her study in a developmental writing program which offers instruction to about 700 students each semester. Over 60% of the students in the program at the time of the study were ESL students, and the rest were native speakers of English.

The researcher asked 6 writing instructors first to mark 6 essays written by ESL students. All the teachers were experienced in teaching L1 writing to native speaker students. Therefore, the researcher identified the sentence-level grammar errors in 6 papers, which are characterized with ESL students. Later, the researcher corrected these ESL errors so as to make the papers look as if they were written by native speaker students. Following the corrections of sentence-level grammar errors, the researcher re-wrote them with a similar hand-writing found in the original papers. The 6 teachers were then asked to re-mark the corrected papers.

The results of this study indicated that even experienced teachers but not trained to teach ESL paid more attention to the students' grammatical accuracy. They assigned higher holistic scores to the papers, whose grammar errors were corrected, when compared to the scores assigned to the original ones. Therefore, the researcher concludes that the teachers caused many ESL students to fail the writing program although these ESL students do not differ from their native speaker peers in terms of the quality of content and organization found in the papers of both groups of students.

In another study, Hamburg (1984) found out that errors were among the major determiners of the total scores of essays. In his study, investigating the process of holistic grading of compositions written by ESL students, the researcher tried to find out what features of paper teachers take into account while assigning their scores. For that purpose, the researcher used student papers which were written and holistically graded on a ten-point scale by the reading staff the previous year. Among a very large number of papers, the researcher randomly selected totally 30 papers that were labeled as 5, 6 and 7 out of 10 according to the holistic scale. For each group of scores (5, 6 and 7) there were 10 student papers.

Hamborg (1984), refering to Nas (1975), classifies errors found in a student paper as to be at three different levels with respect to their effect on the comprehensibility of a paper. In addition, for each of the three levels, he identifies three types of errors (spelling, lexical and grammatical errors) as follows:

First-Degree Errors

Spelling: deviation from correct spelling is minor, reader has no trouble recognizing the word.

Lexical: deviation from meaning is so minor that reader has no trouble substituting correct word.

Grammatical: 1) occurring in a form that is an exception to a grammatical rule; 2) form or structure would be correct in partly different context, no problem in understanding; 3) form used is correct only in immediate context; 4) error can be explained as the use of the wrong register.

Second-Degree Errors

Spelling: serious deviation from correct spelling; word interpretable in context.

Lexical: *so* serious that item is only interpretable with the help of context.

Grammatical: 1) results in alien word combination or word order, but sentence still interpretable; 2) would be fatal to communication, except that rest of sentence is interpretable even without wrong words; 3) results in a form that can only be interpreted in context.

Third-Degree Errors

Spelling: makes it impossible to be certain about the word that is meant.

Lexical: makes it impossible to be certain about the meaning, except with the help of context.

Grammatical: makes it impossible to be certain about the meaning of the sentence, even with the help of context.

Hamborg (1984) explains the importance of such a classification of errors in terms of holistic assessment of papers saying:

> ''This classification of errors emphasizes the readability characteristics of a composition; a reader would not need to do much work to understand a composition with only a few Firs*t*-Degree errors, while a composition containing many Third-Degree errors would be practically impossible to comprehend. There is, however, a subjective element to this error-classification scheme since it is the grader who makes the subjective decision about whether an error is Firs*t*-Degree, Second-Degree, or Third-Degree'' (p:94).

Taking these degrees of errors into consideration, Hamborg (1984) examines the 10 papers from each group of scores separately and identifies such errors in all papers as being first, second or third-degree errors. The statistical analysis in his study reveal that there were fewer firs*t*-degree errors in the papers that received a holistic score 5 and 6 with respect to the papers whose scores was 7 out of 10. Hence, in this study, the total scores, though they were holistically assigned, were found to be consistent with the degrees of errors found in three groups of essays whose total scores were 5, 6 and 7, which were the middle scores of a ten-point scoring scale.

Vann, Meyer and Lorenz (1984) found that the errors at sentence-level are judged by different standards. In their study, the researchers first elicit 12 different types of sentence-level errors commonly found in ESL papers. Then, they select 24 sentences from different papers written by ESL students. Eventually, the researchers obtain 2 sentences for each error type. All these 24 sentences were sent to totally 164 native speaker respondents from different faculties at Iowa State University, who were asked to rank the sentences that contain different types of second language errors. The ranking was made according to a 5-point acceptability scale with '1' being intolerable in all academic situations and '5' being tolerable in all academic situations.

The analysis of the responses to the error types shows that respondents did not all agree on the same certain type of errors. In addition, the gravity of errors was different from one respondent to another. That is, some errors were considered by some respondents to be less acceptable, while the same type of error was tolerable for some other respondents. Therefore, the study reaches the conclusion that the quality of compositions written by non-native speaker students is judged in terms of some factors such as comprehensibility and correctness.

However, the researchers mention one important limitation to their study. All the sentences were separate statements, and there was, hence, no content and organization provided. For that reason, the researchers suggest a further study that focuses on sentence-level errors yet in a context.

In a later study, with a research design similar to a previous study carried out by Vann, Meyer and Lorenz (1984), Janopoulos (1992) searched for the possible tolerance of NS and NNS writing errors. Using the same list of 12 sentence-level error types found in the previous study mentioned above, the researcher obtained 24 separate sentences which were written by non-native speaker of students and which contain the error types.

The sentences were sent to totally 177 instructors from different faculties, and they were asked to rank the sentences according to the seriousness of the error types on a 6-point scale. Half the respondents were informed that they were going to rank errors committed by non-native speaker students, and the other half were told that the errors were in native-speaker origin. Eventually, the results of this study revealed that there generally occurred more tolerance of non-native speaker errors than of errors that were perceived as being made by native speaker errors.

In a more comprehensive study on errors, Santos (1988) investigated how instructors from two different faculties react to essays written by non-native speaker students. The researcher, by commenting on the results of a questionnaire given to the graders, also searched for whether reader characteristics such as age, gender, and native language affect the scoring of essays. The researcher concluded that the rhetorical features of writing, such as content and organization, are among the major factors that influence the graders' overall scores.

From two different faculties, a total of 178 professors participated in Santos's (1988) study. They were given two compositions which were carefully selected on the basis of certain criteria. For instance, both of the essays had a similar total number of words. The two essays suited the standard five-paragraph-system of essay organization; namely, introduction, body (three paragraphs), and conclusion. In addition, the essays were representatives of a variety of errors made by non-native speaker students.

The graders were asked to rate the two essays on a 10-point scale in terms of content and language use. The results indicated significant difference between the components of essays. The professors were found to be more severe in judging the content and more tolerant towards the language use of students. That is, the graders were able to distinguish content from language use. Further analysis on language use errors made by students also revealed that the graders found the sentences containing these errors to be highly comprehensible and reasonably unirritating yet linguistically and academically unacceptable.

When the responses given to the questionnaire were taken into account, the ages and the native languages of the professors were found to be significant. That is, the older professors found errors to be less irritating than their younger colleagues did. Furthermore, the native-speaker professors were more tolerant towards errors, and non-native speakers judged the errors more severely.

A more recent study was carried out by Porte and Inglesa (1999) who examined the error-gravity perceptions of NS and NNS graders at Granada University. When the two groups' given scores were statistically compared, no significant difference was observed, while some differences did exist in terms of the perceived gravity of specific errors. However, the researchers concluded that errors were not being considered as serious as the previous studies found. In other words, the participants generally agreed in their judgements regardless of their native languages.

In another study, Shohamy et. al. (1992) investigated rater reliability in terms of whether experience in teaching and the training of raters make a difference on the scores that teachers assign to papers. For that purpose, four groups of raters participated in their study. There were five raters in each group who asked to grade totally fifty randomly-selected written texts using three different scoring scales which were developed for the study: A holistic scale, a communicative scale and an accuracy scale. Two groups of teachers, one of which consisted of experienced and the other inexperienced raters, received training, whereas the other two groups did not attend the training sessions. All the student texts were scored by all raters using the three scoring scales within two weeks, and the data collected were computed through three methods of analysis: (1) The Ebel intraclass correlation for inter-rater reliability, (2) Repeated Measures analysis of variance (ANOVA) for the effects of background and training on inter-rater reliability, and (3) the Spearman-Brown correlation formula for intra-rater relibility.

The findings of the study revealed that all the four groups of raters achieved high inter-rater reliability coefficients with a range of .80 and .93 regardless of their teaching backgrounds. Although it was also found out that the reliability coefficients for the experienced groups were higher than that of the inexperienced groups, the difference between the groups was not statistically significant. Hence, no effect was observed in relation to the raters' background. However, with respect to the other concern of the study, that is the question of whether training affects raters' judgements, the results of ANOVA indicated that training significantly influenced ratings. Finally, high intra-rater reliability coefficients ranging from .76 to .96 for the trained experienced teachers, suggested that the ratings for that group of teachers were consistent over time.

One more remarkable study in the field was carried by Unat (1999) who tried to find out whether two factors such as the raters' experience in teaching and their nationality play a role on the scores assigned to students' essays. In order to conduct her study, 30 teachers were asked to rate 3 essays using an analytic scoring guide. These essays differed in quality and were labeled as good, average and poor. In addition, the participating teachers were grouped with respect to their nationality and teaching experience as follows:

1- 10 native experienced teachers
2- 10 non-native experienced teachers
3- 10 non-native inexperienced teachers

The results of the independent *t*-test applied to find out the inter-rater reliability of the raters across groups reveled that all the teachers regardless of their teaching experience and nationality achieved highly consistent results on the total mean scores of the three essays. However, when the mean scores assigned to the sub-components were analyzed, non-native inexperienced teachers were found to judge the grammar component of the poor essay more harshly than the other teachers.

All these studies reveal the very fact that, in the assessment of writing, teachers might show differences in their judgments. As mentioned previously, there can be several underlying factors that account for these differences; the graders' age, gender, the interpretation of the topics, background to the topic, language teaching experience and experience in teaching writing are among the most cited, but language errors found in papers might be considered as the most important.

In the light of all these studies, it would be worth reminding here that writing assessment is a skilled, complex and time consuming activity. Therefore, Hamp-Lyons and Kroll (1996) suggest that graders be experienced as well as trained enough to reward effective academic writing. In addition, they claim that while using a scoring scale, the writing performance expectations stated in the scale will shape the readers' judgments. That is, the more experienced the teachers are and the more they use the scoring scales, the more effective they will be in using these scales. Hamp-Lyons and Kroll (1996) also believe that ''as these scales become known to writers ..., conventional expectations about the features of a good writing are gradually being developed'' (p:63).

CHAPTER 3

METHODOLOGY

3.1. Introduction

In this chapter, first, the subjects that participated in the study are described. Following this, data collection procedures, which consist of the selection of student essays as well as the grading process of these essays are presented. Next, the internal consistency of graders across two different occasions are discussed. At the end of this chapter comes the procedure to be followed for statistical analyses.

Previous research findings mentioned in Chapter 2 reveal that certain features of students' texts have effects on scoring. For instance, teachers may assign a higher total mark to a student's text if it doesn't involve grammar errors. This is even proved to be true for teachers experienced in teaching writing. On the other hand, sometimes, teachers tend to tolerate students' errors in language use, and they thus assign a higher total mark even though the text might have deserved a lower one.

Therefore, in this study, the primary purpose was to find out whether experienced and inexperienced teachers assign higher marks when they assess papers which have high quality of language use. If so, our secondary purpose was to reveal in which of the other sub-components (content, organization, vocabulary use and mechanics) a change occurs. In addition, our next concern was to determine whether teachers in both groups have any differences in their judgements when they use an analytic scoring guide.

3.2. Subjects

This study was carried out at Anadolu University, Preparatory School of Foreign Languages, and the subjects participating in the study were 28 English Language teachers, who were all non-native speakers of English with different teaching backgrounds.

These teachers were selected on the basis of their availability at the time of data collection and their willingness to act as the subjects of this study. Hence, they were asked whether they would like to take part in the study. Eventually, 28 teachers kindly volunteered to participate.

The criterion for the selection of the teachers was experience in teaching writing since this study also focused on the possible effect of this factor on scoring the students' essays. Teachers who had taught writing at least three years and above were considered to be experienced (Johnson et. al., 2000). At the time of the study, the number of experienced teachers was limited to ten. As for the novice group, eighteen teachers, whose language teaching experience varied between seven months and two years, agreed to rate the essays. The inexperienced teachers had taught writing for between only one term or three terms. On the other hand, experienced teachers had taught English language for 5 to 13 years. Hence the subjects of this study were as follows:

1. 10 experienced non-native teachers
2. 18 inexperienced non-native teachers

3.3. Materials

The materials used in this study consisted of 40 essays written by students under exam conditions. These 40 papers were chosen following a certain procedure as described below. Following the description of the selection procedure, ESL Composition Profile, the analytic writing criteria used in this study, is presented. The essays were graded using this analytic profile since this approach to writing assessment provides diagnostic information which this study needs so as to achieve its goals.

3.3.1. Selection of the Students' Essays

The study included a total of 40 student essays to be all graded by both groups of teachers. The essays were written in the final proficiency exam by lower intermediate students at the end of the previous academic year. The reason for choosing this proficiency level was the fact that the essays written by lower-intermediate students were more likely to include language use errors at sentence-level which do not affect the comprehensibility of a paper. A total of 417 students at this proficiency level had taken the exam and had been asked to write a five-paragraph essay by choosing one topic among five different essay topics. All the five topics required students to write a cause-and-effect essay.

In order to avoid the possible effect of different topics on the teachers' grading process (Ruth and Murphy, 1988), it was considered necessary to choose the essays written on the same topic. Therefore, by random selection, the topic was determined to be *'effects of working too much'*, which 97 students had chosen to write essays about. Among these essays, 6 were ignored due to being too short to meet the standard of five-paragraph essay organization; namely, introduction, body (three paragraphs), and conclusion. Furthermore, 5 essays were also ignored because they were too illegible. Eventually, there were 86 essays left from which a total of 40 essays were chosen at random to be used in this study. These 40 essays were then divided in two sets. Therefore, there were 20 essays in each set. One set of 20 essays was used to check the intra-rater reliability of all the teachers, and the other set of 20 essays was used to find answers to the research questions directed in this study, that is, to investigate whether accuracy in grammar affects the teachers' scores or not.

The second set of essays, which was used for the observation of the effect of grammar errors, was typed in computer, and the range of the number of words used in these 20 essays was found to be between 252 and 345 with a mean of 303, which was also between the required numbers by the final proficiency exam.

In addition, before this study was carried out, experts both in the field of writing skills and in other fields of English language teaching were consulted by the resaercher, and they considered the total number of these essays to be enough for this study. In addition, when compared to previous studies in the field which used a total of 3 to 6 essays for similar purposes, the number of essays in this study is considerably high. It is also a manageable number for the teachers who were going to rate the essays due to the heavy work load of them at the school at the time of the study. However, what is more important for this study was the large number of graders to participate in this study, which is 28.

Since the teachers were to mark essays twice and some corrections were to be made between the two gradings, all the essays, as mentioned before, were typed on computer beforehand. The purpose of doing so was to eliminate any possible effect of neatness, such as the handwriting of the researcher who, otherwise, had to write the corrected versions for the second grading.

As stated earlier, the 40 essays were randomly divided into two halves, having 20 essays in the first set and 20 essays in the second set. The first set of 20 essays served as a tool to see if teachers in both groups were consistent in their markings over time in terms of their intra-rater reliability. As for the second set, these 20 essays (see appendix D for a sample original version essay) were used, as explained above, for the purpose of the observation of whether grammatical accuracy had any effects on the gradings of both groups of teachers. The teachers were asked to grade both sets of essays twice. However, prior to the second grading, which was held a month after the first grading, the sentence-level errors found in the second set of essays were corrected (see appendix E for a sample essay whose sentence-level errors were corrected). Hence, in the second grading, the teachers marked the corrected versions of the second set of 20 essays, whereas they re-marked the same first set of 20 essays which were previously used in the first grading.

As can be understood, the teachers were given the two sets of essays together in the first grading as well as in the second grading. Therefore, since there was a certain period of time, which was a month, between the two gradings, it was considered necessary first to statistically determine the internal consistency of all the raters across two different occasions. In this way, we would be able to see whether the teachers participating in this study were capable of re-judging the papers with similar standards across the two gradings. The statistical results of the internal consistency coefficients for each teacher are presented at the end of this chapter.

3.3.2. ESL Composition Profile

As mentioned in the previous section, the teachers received a total of 40 essays together, and they were asked to mark them by using analytic writing criteria. As a scoring guide, ESL Composition Profile, developed by Jacobs et al (1981), was used. This scoring guide, according to Bahçe (1992), makes the scoring more reliable since the guide makes it possible to consider the same aspects during the assessment of compositions, which are content, organization, vocabulary use, language use and mechanics. Jacobs et al. (in Bahçe, 1981; p:39) consider this guide to be different from holistic approach saying:

> ''This is an important difference, since readers sometimes tend to value one aspect of a composition when using a purely impressionistic approach, yet it is only through a writer's successful production, integration and synchronization of all these component parts of a composition an effective whole is created. The profile asks readers to peer at the composition through as many windows as possible in arriving at their judgements of quality'' (p:31).

Therefore, the profile is made up of sub-components, all of which teachers are supposed to refer to while scoring the essays. The weighting of each component depends on the degree of its importance for written communication. That is, the ideas presented in a student paper are considered to be more important than the student's grammatical ability, whereas the importance of the quality of vocabulary use in that paper is thought to be equal to the importance of how that student organizes his ideas in the paper. Table-1 below illustrates the weighting of each of these sub-components out of a total hundred:

Table-1. The components of ESL Composition Profile

Content	30
Organization	20
Vocabulary	20
Language Use	25
Mechanics	5
TOTAL	100

Furthermore, each sub-component consists of four mastery levels with clear descriptors: excellent to very good, good to average, fair to poor and very poor (see appendix A). Jacobs et al. (1981) are of the opinion that ''this profile's mastery levels and associated shorthand criteria thus provide a well-defined standard and an interpretive framework for all readers as they read composition and judge its communicative effectiveness'' (in Bahçe, 1992; p:40).

3.4. Data Collection Procedures

The teachers in both groups were familiar with the criteria as, at the time of the study, they were in the committee of assessment of writing skills of students who were attending Anadolu University, Open Education Faculty, English Language Teaching Program. In this ELT program, for the assessment of writing skills, an adapted version of the profile was used. However, the descriptors were similar to the original's.

Despite the familiarity with the profile, the teachers participating in this study were also provided with a user-guide proposed by Jacobs et al (1981). This user-guide aims at making readers aware of how to use ESL Composition Profile by giving full descriptions of each mastery level for each sub-component (see appendix B). Therefore, all the teachers received a pack which included the ESL Composition Profile and the user-guide along with the 40 essays.

3.4.1. Grading of the Essays by Using ESL Composition Profile

In order to achieve the goals of this study, the teachers were asked to grade two sets of essays twice. In the first grading, the teachers marked both of the two sets. For the second grading, the sentence-level grammar errors of 20 papers in one set were corrected while the other set of 20 papers remained the same as in the first grading. The second grading was held one month after the first grading. The procedure is illustrated in table-2 as follows:

Table-2. The Scoring Procedure

First Grading		Second Grading
Set A1 * (20 essays)	$\xrightarrow{\textit{the same}}$	Set A1 ** (20 essays)
Set B1 *** (20 essays)	$\xrightarrow{\textit{corrected}}$	Set B2 **** (20 essays)

* One set of 20 original essays (Set-A1) used in the first grading

** The same set of 20 original essays (Set-A1, previously used in the first grading) was used in the second grading as well

*** Another set of 20 original essays (Set-B1) used in the first grading

**** The sentence-level errors of 20 papers (Set-B1, previously used in the first grading) were corrected before the second grading

3.4.1.1. First grading

For the first grading, all the teachers were asked to grade the 40 essays. Therefore, the teachers were provided with a pack that comprised 40 essays, ESL Composition Profile (see appendix A) and the user-guide of this profile (see appendix B). Since the grading process was not carried out under exam conditions which require teachers to do marking in pre-specified rooms in a certain period of time, the teachers were, for the sake of convenience, given a week so that they could rate the essays according to the analytic profile in their free time and then return the essays.

In addition, the teachers did not know anything about the purpose of the research. Hence, they were not told that half of the essays were going to be used to determine their intra-rater reliability and the other half were going to be used for the purpose of finding out answers to the research questions asked in the present study. Thus, they received both sets of essays (a total of 40 essays) together. All the essays were given in a random order.

3.4.1.2. Second grading

Before the second grading, as mentioned above, the sentence-level grammar errors of the 20 papers in one set of essays were corrected. For that purpose, a native speaker of English who had experience in teaching writing for above 5 years was paid to correct the grammar errors at sentence-level found in the 20 essays. This native speaker was also an instructor at the preparatory school. In the correction process, he was asked to refer to the descriptors in the language use component of the ESL Composition Profile so that any correction of other errors, for instance the errors that belong to the use of vocabulary, could be avoided.

Here, it is necessary to mention an important characteristic of sentence-level errors referring to Leki (1992), ''problems at the discourse level are often fairly subtle, leaving the reader with the feeling that something is not quite right with a text but with no clear picture of where the problem lies. At the sentence level, however, errors are relatively obvious'' (p:105). Therefore, it was not expected to have difficulty in correcting the errors. However, the native instructor reported something different that it was sometimes really difficult to distinguish between grammar errors and vocabulary errors. There were few occasions where he claimed some errors to belong to vocabulary component though they seemed to be language use errors. Eventually, two other instructors who were also native speakers of English were consulted, and the errors in question were agreed upon to be grammar errors when the descriptors for these two components provided in the ESL Composition Profile were taken as the criterion.

The other set of 20 essays remained the same. In other words, the same set used in the first grading was used again for the second grading without any change. An important point to mention once more here for the second grading is that the teachers were not informed that they were going to rate the same set of 20 essays, nor did they know any of the changes that were made in the other set of 20 corrected-papers. They merely received the second pack of 40 essays and were asked to mark all these essays. However, some of the teachers reported that they found these essays similar to those that they had marked a month before. Despite this, they were told nothing about the issue nor about the purpose of the study. Therefore, they were only informed that it was only for research purposes. Another important point for the second grading was that all the 40 papers in two sets were put into the packs in a random order so that the sequence of essays would not have an impact on markings.

Unfortunately, the study eventually had to be carried on with 24 teachers since 4 of the inexperienced teachers were unable to return the essays for some personal reasons at the end of either the first grading or the second grading. Therefore, the scores of the remaining 24 teachers were taken into account for the statistical analyses.

3.5. Determining the Internal Consistency of Raters' Scores across Two Gradings

In order to be able to compare the scores of 20 essays marked in the first grading with the scores of the same 20 essays re-marked in the second grading, the Spearman-Brown Corelation based on the Pearson Moment Product Correlation was run. This statistical tool would provide us with correlation coefficients of the internal consistency of the scores assigned by 24 teachers.

For this purpose, as mentioned above, the scores obtained in the first and second gradings were compared. These scores were assigned to one set of 20 essays which remained the same in the first and second gradings. To observe the consistency of scores of the essays marked by 14 inexperienced teachers, Spearman-Brown Correlation was applied for comparison, and the correlation coefficients were found for each of the inexperienced teachers as shown in table-3 below.

Table-3. Correlation coefficients of internal consistency of scores assigned by 14 inexperienced teachers

Teachers	r	Teachers	r
1	0,63 *	8	0,83
2	0,90	9	0,90
3	0,95	10	0,78
4	0,80	11	0,63 *
5	0,86	12	0,65 *
6	0,94	13	0,84
7	0,77	14	0,83

* Coeffients found to be below 0,70

The statistical results indicated that the range of reliability correlation coefficients for the inexperienced teachers varied between 0,63 and 0,95. These results revealed that 11 teachers out of 14 inexperienced teachers assigned consistent scores with a varying correlation coefficients between 0,77 and 0,95 with a mean of 0,85. However, the scores assigned by three teachers in the inexperienced group had the correlation coefficients of 0,63, 0,63 and 0,65. The correlation coefficients of these 3 inexperienced teachers were below 0,70, which is not desirable in terms of the consistency of the scores (Baker, 1989). Therefore, from the inexperienced group, these 3 teachers with low correlation coefficents were not included in the data analysis procedure described in the following chapter.

The same statistics were also applied for the experienced teachers. In table-4 below, the results of the correlation coefficients are presented:

Table-4. Correlation coefficients of internal consistency of scores assigned by 10 experienced teachers

Teachers	r	Teachers	r
1	0,75	6	0,94
2	0,80	7	0,91
3	0,81	8	0,91
4	0,90	9	0,86
5	0,82	10	0,73

The results indicated that the mean of correlation coefficients of the scores assigned in the first and second gradings by experienced teachers was 0,84 with a range of 0,73 and 0,94. That is, the coefficients for the experienced teachers were all above 0,70. Therefore, all the experienced teachers were included in the statistical analyses presented in the next chapter.

To sum up, in terms of at least for the marking of a set of 20 essays used as a part of this study, except for three teachers in the inexperienced group, all the rest of the inexperienced and experienced teachers can be considered to assign consistent scores when they were asked to re-mark the same essays after a certain period of time. This result was important because the other set of 20 essays (which was used to provide answers to the research questions asked in this study) were given to the teachers for the first and second gradings at the same time with the first set of 20 essays (through the scores of which the internal consistency of raters are calculated and presented above). Hence, 11 inexperienced and 10 experienced teachers, who were found to be consistent in assigning similar scores on two different occasions, were involved in the statistical data analyses process.

3.6. Statistical Analysis

For the statistical analysis of the scores obtained through the grading of the other set of 20 essays, two different statistics were applied: paried *t*-test and independent *t*-test First, in order to be able to see whether the teachers' scores were influenced after the sentence-level grammar errors were corrected, paired *t*-test was run for each group of teachers. To do this, first, two sets of scores assigned by 11 inexperienced teachers were taken into account: *the total scores* of the 20 essays marked in the first grading were compared with the *expected total scores* of the corrected set of 20 essays re-marked in the second grading. Following this, if a change was found in the total scores, we would go on to analyze the sub-scores assigned to each of the sub-components (content, organization, vocabulary use and mechanics) in the first and second gradings.

The same procedure was also followed for the comparison of the total scores as well as of the sub-scores assigned by 10 experienced teachers again running paired *t*-test. Table-5 below illustrates the comparisons of the total scores and sub-scores of each individual group of teachers for both gradings.

Table-5. The comparisons of the total scores and the sub-scores of both groups of teachers for the two gradings

Inexperienced teachers		
First Grading		Second Grading
TOTAL	v.s.	TOTAL
In the case of a significant difference between the total scores, then the sub-components would be compared		
Content	v.s.	Content
Organization	v.s.	Organization
Vocabulary	v.s.	Vocabulary
Mechanics	v.s.	Mechanics

Experienced teachers		
First Grading		Second Grading
TOTAL	v.s.	TOTAL
In the case of a significant difference between the total scores, then the sub-components would be compared		
Content	v.s.	Content
Organization	v.s.	Organization
Vocabulary	v.s.	Vocabulary
Mechanics	v.s.	Mechanics

Next, in order to be able to compare the scores across both groups of teachers, independent *t*-test was run. For this purpose, two sets of scores were used (see table-6 below): *the scores* which were assigned to the 20 corrected-version essays by the inexperienced teachers in the second grading were compared with *the scores* which were assigned to the same 20 corrected-version essays by the experienced teachers in the second grading. With the help of independent *t*-test, we would be able to see if there was any significant difference across the scores in terms of the graders' teaching backgrounds, that is experience in teaching writing. In other words, we would be able to see whether experienced and inexperienced teachers' scores differed across groups after the correction of sentence-level grammar errors.

Table-6. The comparison of the expected total scores across both groups of teachers for the second grading

Second Grading		
Inexperienced Teachers		Experienced Teachers
TOTAL expected scores	v.s.	TOTAL expected scores

In this study, for all of the statistical analyses, the level of significance was taken as *p*<0,05.

CHAPTER 4

PRESENTATION AND ANALYSIS OF DATA

4.1. Introduction

This study aims to identify the effect of students' accurate use of grammar in their written papers on the scores that experienced and inexperienced teachers assign. Therefore, it is aimed to examine whether correcting the students' language use errors will change the teacher's scores that they re-assign to the papers. The study also tries to find an answer to the question of whether there is a difference between the scores assigned by experienced teachers and the scores assigned by inexperienced teachers. Therefore, in this chapter answers will be provided for the following research questions previously asked in this study:

1. When compared with the total scores assigned by the inexperienced teachers in the first grading, do the total scores assigned by the same group of teachers increase in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?
    1.1.If so, in which of the four sub-components (content, organization, vocabulary, and mechanics) does a significant change occur between the first and second gradings?
2. When compared to the total scores assigned by the experienced teachers in the first grading, do the total scores assigned by the same group of teachers increase in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?

have been given due to the correction of language use errors. For this purpose, first, the teachers' sub-scores assigned to language use component in the first grading were subtracted from the sub-scores assigned to the same component in the second grading. Next, the obtained values were also subtracted from the total scores that the teachers assigned in the second grading. Consequently, we would have the *expected* total scores from the second grading as shown in table-7 below which illustrates a sample calculation of an expected total score for an inexperienced teacher's scores:

Table-7. A sample calculation of an expected total score

	GRADINGS				Calculation of an Expected Score in the Second Grading	
	Language Use Component (out of 25)		TOTAL (out of 100)		Increase in Language use in the second grading	**The Expected Total score in the second grading**
Teachers	First Grading	Second Grading	First Grading	Second Grading		
Inexperienced Teacher Number-11 Paper Number-7	11	19	**63**	82	(19-11)= 8	(82-8=) **74**
	Mean		Mean		Mean	
Inexperienced Teachers	14,3272	19,859	**69,1272**	78,1863	(14,3272-19,859)= 5,5318	(78,1863-5,5318)= **72,6545**
Experienced Teachers	13,98	19,085	**67,885**	76,245	(19,085-13,98)= 5,105	(76,245-5,105)= **71,14**

We called these calculated scores as *expected total scores* since if a teacher's scores are not influenced by the correction of language use errors, that teacher is supposed to re-assign an expected total score similar to the total score assigned by him or her in the first grading. In table-7 above, sample scores of an inexperienced teacher who participated in this study is shown. The teacher seems to assign a higher total score (82) when he marked a paper whose sentence-level grammar errors were corrected for the second grading. On the other hand, that teacher was previously observed to assign a total score of 63 to the original version in the first grading. As for the language use sub-component, the teacher assigned a sub-score of 11 in the first grading and 19 in the second grading. This 8-point increase in the language use component in the second grading is already expected to occur since language use errors were corrected for the second grading. However, when we subtract this 8-point from the total score of the second grading, we obtain a score of 74 (82-8=74), which is called an *expected total score* for the second grading. When this expected total score of 74 obtained from the second grading is compared with the total score of 63 assigned by the same teacher in the first grading, the teacher can now be thought to assign a higher mark in the second grading.

Following the same calculations above, the expected total scores were found for each teacher and for each of the 20 papers for the second grading. Table-7 above also shows the mean of the sub-scores of language use component and the mean of the total scores assigned in the first and second gradings by experienced and inexperienced teachers. Table-7 also presents the mean of the expected total scores for the second grading for both groups of teachers (see appendix G for all the sub-scores and total scores for both gradings and the expected total scores for the second grading found for each teacher). Eventually, the total scores assigned in the first grading and the *expected total scores* obtained from the second grading were statistically compared for each group of teachers as illustrated in the next part.

4.2.1. Analyses of the Scores Assigned by Inexperienced Teachers to the Original essays in the First Grading and to the Corrected Versions in the Second Grading

Paired *t*-test was applied to see whether there was a significant difference between the total scores assigned in the first grading and the *expected* total scores obtained from the second grading for the inexperienced group of teachers.

Table-8. The mean of the total scores of inexperienced teachers in the first grading and the mean of the expected total scores of the same group of teachers for the second grading

Inexperienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (total scores)	69,1272	20	3,0110	,6733
Second Grading (expected total scores)	72,6545	20	3,0047	,6719

As it is seen in table-8 above, the mean score of the total scores assigned by the inexperienced teachers in the first grading was 69,12. However, the mean score of the expected total scores obtained from the second grading for the same group of teachers was 72,65.

The paired *t*-test results are shown below in table-9, which indicated a significant difference between the two sets of scores at a significance level of 0,05.

Table-9. *t*-test results of the comparison of the total scores assigned in the first grading with the expected total scores obtained from the second grading for inexperienced teachers

Inexperienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading	-3,5273	,6717	,1502	-3,8416	-3,2129	-23,485	19	,000*

* *p* value is significant at ,05 level

Since a significant difference for the inexperienced teachers was found between the total scores assigned in the first grading and the expected total scores obtained from the second grading, it was thus necessary to compare the sub-scores assigned to each sub-component in the first and second gradings so that we could find out in which of the four sub-components (content, organization, vocabulary use and mechanics) a change had occurred.

For this purpose, all the sub-scores assigned to the four sub-components were analyzed by applying paired *t*-test.

4.2.1.1. Analysis of the Sub-scores Assigned by Inexperienced Teachers to the Content Components of Papers in the First and Second Gradings

The mean of the sub-scores given to the content component of the papers in the first grading was found to be 22,34 and 22,98 for the same component for the second grading, as seen in table-10 below:

Table-10. The mean of the sub-scores assigned to the sub-component of content by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (content)	22,3410	20	,9897	,2213
Second Grading (content)	22,9863	20	1,2385	,2769

In table-11 below, the *t*-test results are shown. if the *p* value is taken as 0,05, the difference between the two mean scores was statistically significant when paired *t*-test was applied.

Table-11. *t*-test results of the comparison of the mean scores assigned to content sub-component by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (content)	-,6455	1,2260	,2742	-1,2193	-7,1650E-02	-2,354	19	,029* -

* *p* value is significant at ,05 level

4.2.1.2. Analysis of the Sub-scores Assigned by Inexperienced Teachers to the Organization Components of Papers in the First and Second Gradings

In terms of the organization component, the mean of the sub-scores assigned in the first grading was 13,85. However, as for the second grading, it was 15,97 for the same component as can be seen in table-12 below:

Table-12. The mean of the sub-scores assigned to the sub-component of organization by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (organization)	13,8590	20	1,0168	,2274
Second Grading (organization)	15,9772	20	1,0415	,2329

Paired *t*-test revealed that the difference between these two mean sub-scores was statistically significant at the significance level of 0,05. Table-13 below shows the *t*-test results.

Table-13. *t*-test results of the comparison of the mean scores assigned to organization sub-component by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (organization)	-2,1182	1,3581	,3037	-2,7538	-1,4826	-6,975	19	,000*

* *p* value is significant at ,05 level

4.2.1.3. Analysis of the Sub-scores Assigned by Inexperienced Teachers to the Vocabulary Use Components of Papers in the First and Second Gradings

As for the vocabulary use sub-component, the mean of the sub-scores given in the first grading was calculated as to be 14,70 and 15,31 for the same component for the second grading, as shown in table-14 below:

Table-14. The mean of the sub-scores assigned to the sub-component of vocabulary use by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (vocabulary use)	14,7000	20	,7698	,1721
Second Grading (vocabulary use)	15,3136	20	,7632	,1706

According to the results of paired *t*-test presented in table-15 below, there was a significant difference between the two mean sub-scores assigned to the vocabulary use sub-component. The *p* value was found to be lower than the significance level of *p*<0,05.

Table-15. *t*-test results of the comparison of the mean scores assigned to vocabulary use sub-component by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (vocabulary use)	-,6136	,6948	,1554	-,9388	-,2885	-3,950	19	,001*

* *p* value is significant at ,05 level

4.2.1.4. Analysis of the Sub-scores Assigned by Inexperienced Teachers to the Mechanics Components of Papers in the First and Second Gradings

In terms of the sub-scores assigned to the mechanics component of the papers, the mean was found to be 3,90 in the first grading and 4,05 for the same component for the second grading, as seen in table-16:

Table-16. The mean of the sub-scores assigned to the sub-component of mechanics by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (mechanics)	3,9000	20	,3078	6,882E-02
Second Grading (mechanics)	4,0502	20	,3209	7,176E-02

When these two mean sub-scores were compared running paired *t*-test, the results, as shown in table-17 below, indicated a significant difference between the two mean sub-scores at the significance level of 0,05.

Table-17. *t*-test results of the comparison of the mean scores assigned to mechanics sub-component by inexperienced teachers in the first and second gradings

Inexperienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (mechanics)	-,1545	,3109	6,952E-02	-,3000	-9,0459E-03	-2,223	19	,039*

* *p* value is significant at ,05 level

4.2.2. Analyses of the Scores Assigned by Experienced Teachers to the Original Essays in the First Grading and to the Corrected Versions in the Second Grading

The same statistics used above for the comparison of the scores of the inexperienced group was repeated for the comparison of the scores of the experienced teachers as well. Therefore, paired *t*-test was run again so as to find out if the total scores assigned in the first grading and the *expected* total scores obtained from the second grading differed for the experienced group significantly.

Table-18. The mean of the total scores of experienced teachers in the first grading and the mean of the expected total scores of the same group of teachers for the second grading

Experienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (total scores)	67,885	20	2,9971	,6702
Second Grading (expected total scores)	71,14	20	3,1733	,7096

Table-18 above shows that the mean score of the experienced teachers for the first grading was 67,88. On the other hand, the mean of the expected total scores obtained from the second grading was 71,14. As for the *t*-test results, at the significance level of $p<0,05$, table-19 below shows that there is a significant difference between these two sets of mean scores.

Table-19. *t*-test results of the comparison of the total scores assigned in the first grading with the expected total scores obtained from the second grading for experienced teachers

Experienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First & Second Grading	-3,2550	,6708	,1500	-3,5689	-2,9411	-21,701	19	,000*

* p value is significant at ,05 level

Since a significant difference for the group of experienced teachers was found between the total scores assigned in the first grading and the expected total scores obtained from the second grading, we compared and statistically analyzed the sub-scores assigned to each sub-component in the first and second gradings by applying paired *t*-test. In this way, we would to be able to reveal in which of the four sub-components (content, organization, vocabulary use and mechanics) there had been a change.

4.2.2.1. Analysis of the Sub-scores Assigned by Experienced Teachers to the Content Components of Papers in the First and Second Gradings

When the sub-scores assigned to the sub-component of content of the papers in the first and second gradings were taken into account, the mean of the sub-scores for the first grading and for the second grading was found to be respectively 22,33 and 23,28 as can be seen in table-20 below:

Table-20. The mean of the sub-scores assigned to the sub-component of content by experienced teachers in the first and second gradings

Experienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (content)	22,3300	20	1,0142	,2268
Second Grading (content)	23,2850	20	,9034	,2020

When paired *t*-test was run on the sub-scores assigned to the content component, the results, illustrated in table-21 below, indicated a significant difference between these two mean sub-scores at the significance level of 0,05.

Table-21. *t*-test results of the comparison of the mean scores assigned to content sub-component by experienced teachers in the first and second gradings

Experienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (Content)	-,9550	1,0918	,2441	-1,4660	-,4440	-3,912	19	,001*

* p value is significant at ,05 level

4.2.2.2. Analysis of the Sub-scores Assigned by Experienced Teachers to the Organization Components of Papers in the First and Second Gradings

With respect to the sub-scores assigned to the organization component of the papers, the mean of these sub-scores in the first grading was found to be 13,52. On the other hand, the mean for the same component was calculated as to be 15,21 for the second grading, as seen in table-22 below:

Table-22. The mean of the sub-scores assigned to the sub-component of organization by experienced teachers in the first and second gradings

Experienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (organization)	13,5200	20	1,1660	,2607
Second Grading (organization)	15,2150	20	1,1731	,2623

In table-23 below, paired *t*-test results are shown. The difference between these two mean sub-scores assigned in the first and second gradings was statistically significant at the significance level of 0,05.

Table-23. *t*-test results of the comparison of the mean scores assigned to organization sub-component by experienced teachers in the first and second gradings

Experienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (organization)	-1,6950	1,0748	,2403	-2,1980	-1,1920	-7,053	19	,000*

* *p* value is significant at ,05 level

4.2.2.3. Analysis of the Sub-scores Assigned by Experienced Teachers to the Vocabulary Use Components of Papers in the First and Second Gradings

As for the vocabulary use component, the mean of the sub-scores given in the first grading was 14,15, whereas it was 14,68 for the same component for the second grading, as seen in table-24 below:

Table-24. The mean of the sub-scores assigned to the sub-component of vocabulary use by experienced teachers in the first and second gradings

Experienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (vocabulary use)	14,1500	20	1,1390	,2547
Second Grading (vocabulary use)	14,6800	20	,9812	,2194

Though the same component seems to have received higher marks in the second grading when the mean sub-scores are taken into consideration, the difference between these two mean sub-scores was not statistically significant as shown in table-25 below. The *p* value was found to be higher than the significance level of 0,05.

Table-25. *t*-test results of the comparison of the mean scores assigned to vocabulary use sub-component by experienced teachers in the first and second gradings

Experienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (vocabulary use)	-,5300	1,1585	,2591	-1,0722	1,221E-02	-2,046	19	,055

4.2.2.4. Analysis of the Sub-scores Assigned by Experienced Teachers to the Mechanics Components of Papers in the First and Second Gradings

As shown in table-26 below, the mean of the sub-scores given to the mechanics component of the papers in the first grading was found to be 3,90. The mean of the sub-scores assigned to the same component was 3,98 for the second grading.:

Table-26. The mean of the sub-scores assigned to the sub-component of mechanics by experienced teachers in the first and second gradings

Experienced Teachers	Mean	N	Std. Deviation	Std. Error Mean
First Grading (mechanics)	3,9050	20	,3120	6,976E-02
Second Grading (mechanics)	3,9800	20	,2560	5,725E-02

When compared with the mean sub-scores assigned to the mechanics component in the first grading, the experienced teachers valued the mechanics component a little more with a mean difference of 0,08 in the second grading. However, the difference found between these two mean sub-scores was not statistically significant at the significance level of 0,05 when paired *t*-test was applied. Table-27 below shows the *t*-test results.

Table-27. *t*-test results of the comparison of the mean scores assigned to mechanics sub-component by experienced teachers in the first and second gradings

Experienced Teachers	Paired Differences					*t*	*df*	*p*
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
First and Second Grading (mechanics)	-8,0000E-02	,2783	6,224E-02	-,2103	5,027E-02	-1,285	19	,214

4.3. Comparison of the Expected Total Scores of Inexperienced Teachers with the Expected Total Scores of Experienced Teachers in the Second Grading

In order to be able to see whether experienced and inexperienced teachers differ in their scores, independent *t*-test was run since both groups of teachers represent different samples. To do this, the two groups of teachers were compared in terms of whether they assigned different scores to the corrected-version essays in the second grading.

In order to be able to observe if there was a significant difference between the experienced and inexperienced teachers, two sets of scores were compared: the *expected total scores* previously calculated through the total scores that inexperienced teachers assigned to the essays whose language use errors were corrected in the second grading were compared with the *expected total scores* previously calculated through the total scores of the essays that the experienced teachers marked in the second grading. For this purpose, an independent *t*-test was applied. Table-28 below presents the mean of the expected total scores of both groups of teachers.

Table-28. The mean scores of the expected total scores of inexperienced and experienced teachers for the second grading

	Teachers	N	Mean	Std. Deviation	Std. Error Mean
Second Grading	Inexperienced	20	72,65	3,0047	,6719
	Experienced	20	71,14	3,1733	,7096

Looking at the mean scores, it seems that inexperienced teachers assigned higher marks than their experienced colleagues. On the other hand, the independent *t*-test results did not prove this as shown in table-29 below. The *p* value was found to be 0,129, which is above the significance level of 0,05. This shows that there was no significant difference between the two groups' *expected* total scores that were obtained through the total scores assigned in the second grading.

Table-29. ***t*-test results of the comparison of the expected total scores of experienced and inexperienced teachers in the second grading**

Second Grading	Experienced and Inexperienced Teachers	*t*	*df*	*p*	Mean Difference	Std. Error Difference	95% Interval Confidence of the Difference	
							Lower	Upper
		1,550	38	,129	1,5145	,9772	-,4637	3,4928

4.4. Summary of the Results

The statistical results are interpreted in two parts. The first part concerns the effect of accurate use of language on the total scores, as well as on the sub-scores assigned to each sub-component, which provides answers to the first two research questions of this study. The next part is related to the background of the teachers, that is experience in teaching, which will highlight the last research question directed in the present study through the discussion of the statistical results.

4.4.1 Summary of the Results in terms of the Effect of Accurate Use of Language on the Total Scores And on the Sub-scores of Inexperienced And Experienced Teachers

The scores were analyzed for both groups, experienced and inexperienced teachers, individually. For the inexperienced group of 11 teachers, two sets of scores were initially taken into account: The first set of scores consisted of the total scores assigned to the original 20 essays in the first grading; the second set of scores were the expected total scores obtained by some calculations of the total scores assigned to the same set of 20 essays, yet whose sentence-level errors were corrected for the second grading. As mentioned at the beginning of this chapter, this second set of expected total scores were obtained as follows (see table-7 for a sample calculation):

First, the sub-score assigned in the first grading to the language use component of the analytic scoring criteria was subtracted from the sub-score assigned to the same component in the second grading. Next, the value obtained through this was also subtracted from the total score assigned in the second grading. Eventually, with these calculations, we obtained the expected total scores which are not thought to involve any extra marks given due to the correction of sentence-level grammar errors.

When the first set of total scores was compared with the second set of expected total scores, the mean score for the first grading was found to be 69,12 for the 11 inexperienced teachers. On the other hand, the mean of the expected total scores obtained from the total scores of the same group of teachers in the second grading was 72,65. This shows that inexperienced teachers assigned higher total marks to the corrected-version essays in the second grading than they did in the first grading. In addition, the difference between these two sets of scores was statistically significant with a *p* value of 0,000 at the significance level of 0,05.

In the next step, the sub-scores assigned in the first and second gradings by the inexperienced group of teachers to each of the four sub-components, namely content, organization, vocabulary use and mechanics were compared and analyzed. For a summary of the results, see table-30 below.

Table-30. The mean sub-scores of inexperienced teachers for each sub-component for the first and second gradings and the mean differences with *p* values

Inexperienced Teachers	First Grading	Second Grading	Mean Difference	*p*
Content (out of 30)	22,34	22,98	0,64	,029*
Organization (out of 20)	13,85	15,97	2,12	,000*
Vocabulary (out of 20)	14,70	15,31	0,61	,001*
Mechanics (out of 5)	3,90	4,05	0,15	,039*

* *p* value is significant at ,05 level

While the highest increase was observed in the sub-component of organization with a mean difference of 2,12, the lowest increase was observed in the mechanics component with a mean difference of 0,15. The content and vocabulary use components seem to have received higher sub-marks with almost a similar mean difference. An increase of 0,64 was observed in the former sub-component, and the latter sub-component was found out to increase with a mean difference of 0,61. Considering the paired t-test results, the increase observed in all the sub-components were statistically significant at the significance level of 0,05.

Moreover, the same process, which was followed for the comparison of the total scores and the expected total scores of the inexperienced group of 11 teachers for the two gradings as stated above, was also followed for the group of 10 experienced teachers. As a result, the mean of the total scores for the first grading was 67,88. However, the mean of the expected total scores was 71,14. Therefore, considering these mean scores, the experienced teachers were also observed to have awarded the papers more whose sentence-level grammar errors were corrected. That is, they assigned higher scores in the second grading than they did in the first grading. This increase was also found to be statistically significant with a *p* value of 0,000 (*p*<0,05).

As the next step, the sub-scores given in the first and second gradings by the experienced group of teachers to each of the four sub-components, namely content, organization, vocabulary use and mechanics were also compared and analyzed. See table-31 below for a summary of the results.

Table-31. The mean sub-scores of experienced teachers for each sub-component for the first and second gradings and the mean differences with *p* values

Experienced Teachers	First Grading	Second Grading	Mean Difference	*p*
Content (out of 30)	22,33	23,28	0,95	,029*
Organization (out of 20)	13,52	15,21	1,69	,000*
Vocabulary (out of 20)	14,15	14,68	0,53	,055
Mechanics (out of 5)	3,90	3,98	0,08	,214

* *p* value is significant at ,05 level

As can be seen from table-31 above, the organization sub-component was observed to have increased in the second grading with a mean difference of 1,69. The other sub-component that was found to increase in the second grading was content with a mean difference of 0,95. The other two sub-components which received higher sub-marks in the second grading were vocabulary use with a mean difference of 0,53 and mechanics with a mean difference of 0,08. Considering the paired *t*-test results, among all the four sub-components which seem to have been awarded with higher grades in the second grdaing if we just look at the mean sub-scores, the increase found in two sub-components, organization and content, was statistically significant at the significance level of 0,05. As for the mean differences found for the other two sub-components, vocabulary use and mechanics, the increase observed in the second grading was not found to be statistically significant.

In the following part, both groups of teachers are compared in terms of their experience in teaching writing in relation to the last research question asked in this study.

4.4.2. Summary of the Results in terms of Raters' Teaching Experience

Comparison of the expected scores of both groups of teachers in the second grading would provide us with an answer to the question of which group of teachers were more affected by the sentence-level grammar error correction, or were they equally affected? In other words, this would reveal whether experience in teaching writing has an effect on scores.

To do this, the expected scores of both groups of teachers in the second grading were taken into account. The mean of the expected total scores for the inexperienced group was 72,65 and the mean of the expected total scores for the experienced group was 71,14. If we look at the mean scores, it seems that inexperienced teachers assigned higher scores when compared with that of the experienced teachers. Nevertheless, statistically speaking, the difference was not significant. That is, experienced and inexperienced teachers were both observed to increase their scores in the second grading, yet to the same extend.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1. Introduction

This chapter starts by giving a brief summary of the study with conclusions drawn from the results. Following this, the findings of this study are compared to the other outstanding previous research results. Next, some pedagogical recommendations are made. At the end, suggestions for further research are provided.

5.2. Summary of the Study

The focus of this study was to first find out whether teachers' scores were influenced when they faced papers with accurate use of language or not. It was also intended to see if experience in teaching writing makes a difference in the scores assigned by teachers while assessing students' written texts. Consequently, this study, basically, aimed to answer the following questions:

1. When compared with the total scores assigned by the inexperienced teachers in the first grading, do the total scores assigned by the same group of teachers increase in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?

1.1.If so, in which of the four sub-components (content, organization, vocabulary, and mechanics) does a significant change occur between the first and second gradings?

2. When compared to the total scores assigned by the experienced teachers in the first grading, do the total scores assigned by the same group of teachers increase in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?

   2.1. If so, in which of the four sub-components (content, organization, vocabulary, and mechanics) does a significant change occur between the first and second gradings?

3. Is there a significant difference between the two groups of teachers' total scores assigned to the corrected-version essays in the second grading?

As the subjects of this study, a total of 28 English language teachers participated in this study. 18 of the teachers were inexperienced teachers, who had taught writing skill for varying periods of time, from one term to three terms. 10 teachers were considered to be experienced in teaching writing as they had taught this skill for three years or above. As for the materials used in the study, a total of 40 essays were graded by all the teachers by using ESL Composition Profile, which has analytic criteria for writing assessment. All the teachers were asked to grade these essays twice, having a period of one month between the two gradings. However, due to some reasons, 4 of the inexperienced teachers did not return the papers either in the first grading or in the second grading. Eventually, the scores of a total of 24 teachers, 14 of whom were inexperienced and 10 of whom were experienced, were taken into consideration for the statistical analysis.

As mentioned previously, these 40 essays were divided into two halves (see chapter 3 for details): The first set of 20 essays were used to see whether inexperienced and experienced teachers re-assign scores consistent with the previous scores that they had assigned a month before. As for the second set, these 20 essays were used to find out answers to the research questions above.

The statistical results of the scores of the first set of essays revealed that all the experienced teachers assigned consistent scores across the two gradings. However, this was not the case for the inexperienced teachers. 3 of the 14 teachers in the inexperienced group were statistically found to assign scores in the first grading inconsistent with the scores that they later assigned in the second grading. That is, when they were asked to grade the same set of essays on a different occasion, they did not re-assign scores similar to the scores that they had previously assigned. This fact led us to exclude these 3 teachers from the statistical analyses of the scores of the second set of essays, which were used as a material to answer the research questions in this study. Consequently, 11 inexperienced and 10 experienced teachers' scores were taken into account for the rest of the statistical analyses.

In order to be able to find out if 11 inexperienced and 10 experienced teachers' scores were influenced by students' accurate use of grammar, the sentence-level grammar errors of the second set of 20 essays were corrected before the second grading (see chapter 3 for further details about the procedure of sentence-level error correction). Hence, the teachers marked the original essays in the first grading and marked the corrected versions of the same essays in the second grading.

Statistical tests were applied for each group of teachers individually, and the results showed that inexperienced teachers assigned higher marks in the second grading. This was also found to be true for the experienced teachers. In other words, when teachers, either experienced or inexperienced, faced papers which included perfect grammar, they assigned higher total marks.

Seeing that both groups of teachers increased their total scores in the second grading when compared with the previous total scores that they had assigned in the first grading, our next concern was to see to which of the other 4 sub-components the increase in the total scores was distributed in the second grading. Statistically speaking, the results indicated that, in terms of the inexperienced teachers, in all of the 4 sub-components (content, organization, vocabulary use and mechanics) there was a significant increase in the second grading. On the other hand, the experienced teachers were statistically found to increase only the two sub-components, content and organization.

Lastly, we wanted to find out if experienced and inexperienced teachers differ from each other in terms of their total scores. That is, this study also tried to answer the question of whether experience in teaching writing had an effect on the scores assigned. For this purpose, both groups of teachers' expected total scores obtained from the second grading were compared across the two groups of teachers. The statistical results of the scores assigned in the second grading revealed that experienced and inexperienced teachers did not significantly differ from each other. Both groups of teachers were able to judge the papers with the same standards.

5.3. Assessment of the Study

The results of this study revealed two very important points in terms of the assessment of writing skills not only specifically for the preparatory school where this study was carried out but also in general for anywhere writing ability is assessed: First, all the teachers were found to increase their total scores if they assess a paper which has an accurate use of grammar. However, what is important here, from the result just stated above, one should never understand the very fact that we simply claim through the results of this study that a paper with accurate grammar receives a higher total mark. Reasonably, it is certainly a typical case for any paper to receive a higher total mark if the rater assigns a higher sub-mark to the sub-component of language use, which, of course, contributes to the total mark.

Instead, what we found out through the results of this study is that if the quality of the language use of a paper were improved yet the other 4 sub-components remaining the same, all the teachers were found to increase their sub-scores that they assigned to the other sub-components of the same paper and thus now observed to increase their total scores.

In terms of the increase in the total scores, it was only 3,53 for the inexperienced teachers and 3,26 for the experienced ones. In writing assessment, the difference between any two teachers' scores are tolerable if there is at most 10-point difference in-between. The increase of a mean of 3,53 and 3,26 in the total scores for both groups of teachers can therefore be considered to be tolerable in terms of writing assessment when %10 tolerable agreement between raters' scores is taken into consideration. On the other hand, it shouldn't be forgotten that the increase in the total scores in this study was observed in an end-of-year exam where 1-point difference causes a student to fail or pass the preparatory class. Therefore, in this study, the total mean score was found to be 69,12 in the first grading and 72,65 in the second grading for the inexperienced teachers, and 67,88 in the first grading and 71,14 in the second grading for the experienced teachers. As can be seen, the mean scores for both groups of teachers were over 70 in the second grading, which is the passing grade. On the other hand, the mean scores were below the passing grade in the first grading.

As for the statistical results of the analyses of the sub-scores assigned to the sub-components, the results were striking. The increase in the sub-scores assgined to the sub-components of content, organization, vocabulary use and mechanics by 11 inexperienced teachers was found to be statistically significant. This was also partly true for the experienced group of 10 teachers since they were observed to increase only the two sub-components, content and organization. However, if we take the ESL Composition Profile into consideration, only in the organization component was a considerable increase observed according to the profile's descriptors for each sub-component (see appendix A for the descriptors for each sub-component). That is, both groups of teachers assessed the organization component of the original essays considering the descriptors of the third bend, whereas they judged the organization sub-component of the grammatically-corrected versions, this time, taking the descriptors of the second bend into consideration. All the other 3 sub-components, content, vocabulary use and mechanics, were observed to remain in the same bend across the two gradings.

Secondly, with respect to whether there was any difference between the scores assigned by experienced and inexperienced teachers, experience in teaching writing did not play a significant role on the scores. That is, both groups of teachers were consistent in-between.

What's more, the results of this study seem to be very much in line with the results of previous research findings. As an example for the first point concluded above, Sweedler-Brown (1993) investigated whether experienced English instructors who are not yet trained to teach English as a second language are influenced by grammatical features of English found in students' papers. The researcher asked 6 instructors to mark 6 student essays twice. All the participating subjects in the study collectively averaged 10 years' experience in teaching writing and had spent at least 5 years evaluating essays. Before the second grading, the researcher corrected the sentence-level errors found in 6 papers. Consequently, the results of a paired *t*-test applied to the scores assigned in two gradings indicated that a significant difference was found between two sets of scores (p=,004).

As for the second point reached in this study, Shohamy et. al. (1992), in their comprehensive study, thought a number of factors to be the source of error of measurement. Among these factors was teaching background of graders. As a result of the statistical analyses, they claim that the graders participated in their study were capable of assigning consistent scores regardless of their teaching backgrounds. This finding is similar to that of ours in that experience in teaching hasn't got an effect on scores.

5.4. Pedagogical Implications

In Turkey, in most educational settings, for instance especially in secondary and high schools, students are commonly taught English as a foreign language on the basis of traditional grammar-focused language teaching. This is also true for most language teaching programs at university level. Furthermore, in general, governmental tests do examine mostly grammatical and vocabulary knowledge besides reading skills. Among such tests, the most outstanding is the language test of university entrance exam, which high school students take in order to attend ELT departments of universities. This traditional way of grammar-focused language teaching/learning (and testing as well) might naturally lead foreign language teachers to keep grammatical correctness at least in mind probably since, previously, they were fundamentally expected to use the language accurately.

This might also be the case in this study where teachers were found to assign higher scores to grammatically-accurate papers though they had followed an analytic writing assessment procedure. However, this is, of course, not a desirable situation especially where students fail or succeed a language learning program at the end of an academic year as in the case of the school of foreign languages at Anadolu University. Even 1-point difference is important in terms of students' failure or success in this program. In this respect, assessment of writing skill, especially for the end-of-year exam, is of great concern for students. This fact, in the light of the results of this study, makes it necessary to draw some pedagogical implications so as to increase the consistency of scores.

First of all, what is important as a result of this research is the consequence that experienced and inexperienced teachers at the preparatory school, at least who participated in this study, have proved to be consistent in-between irrespective of years of teaching experience. This is important because the preparatory school of foreign languages at Anadolu University accommodates a very large number of teachers (over a hundred) with varying teaching backgrounds such as teachers who are in their first or second year of teaching language and those who teach English language over a decade. Hence, the practical implication of this result suggests that the administrators in the preparatory school can select their raters without being concerned about their teaching backgrounds as this variable doesn't seem to increase or decrease the consistency of scores.

Another implication concerns the fact that the teachers participating in this study assigned higher total scores when they faced papers which had accurate use of grammar. This could be possibly also because the teachers did not refer to the analytic writing criteria while assigning their scores to other sub-components such as vocabulary, content or organization. That is, accuracy in language use in papers might have caused the graders to regard grammar as more important than the other features and thus to regard accuracy in grammar as the basis for a paper to deserve a higher mark. This consideration of teachers may be due to the traditional way of grammar-focused language teaching in Turkey.

The effect of accurate grammar use on higher scores can also be related to the possible fact that teachers might have assigned scores on a holistic-basis though they had an analytic writing assessment profile in hand. That is to say, the teachers might have assigned their total scores with the overall impression of accurate language use found in papers.

All of these stand just as possibilities for the source of higher scores assigned to the papers that have accurate use of grammar. Meanwhile, though the possible sources of such errors are hard to prove, it is not rare to hear some teachers say that they do not want to assign a high mark to a paper (which might actually deserve a higher mark that could even be the true score of that paper) just because the paper has some errors in some simple sentence structures such as the subject-verb agreement for the verb 'to be'.

In order to overcome such sources of errors, all teachers, either experienced or inexperienced, could attend training sessions so as to use such an analytic writing assessment profile effectively. In these sessions, teachers should be made more conscious of the importance of referring to the profile while assigning their scores. This is especially important in large-scale testing situations since all teachers are supposed to consider the same standards in order to avoid unfair assignment of scores to different students. Therefore, in the training sessions, teachers should also be made aware of the fact that it is impossible to reach the *true score* for a paper, yet they should be reminded of the fact that the more they take the assessment criteria into consideration in the grading process, the more consistent scores will be achieved among raters. What's more, it would be better if we could statistically investigate the consistency of the scores assigned by teachers who participate in a scoring session. According to these results, some precautions can be taken so as to increase consistency.

Another possible reason for the effect of accurate use of language on scores found in this study could be the fact that the sub-components in the assessment criteria might not have met the expectations of the teachers about what should be involved in a good writing. In order to deal with such a possible expectation, the descriptors in the criteria could be improved or the weightings of some of the components could be increased and some could be decreased by asking the comments of the teachers.

However, though this could be a solution, a more effective suggestion can be to ask two raters to mark different aspects of the same paper independently. That is, one of the teachers can mark the sub-components of vocabulary, language use and mechanics of a paper, while the other can mark the content and organization components of the same paper. The rationale for such a suggestion is that, in this way, the raters will not be allowed to know the total score of that paper while assigning their sub-scores. This will not cause them to think about the total score of the paper, and thus help them avoid re-considering their ratings to the sub-components, which will also avoid another factor called halo-effect.

One more suggestion can be to ask teachers to first correct grammar errors in a paper and assign a sub-score to the language use component. Following this, they can carry on with assessing the other components.

Furthermore, if it is a large-scale testing, as in the preparatory school at Anadolu University, teachers should not be allowed to know the proficiency level of the students (nor their names) while grading the papers. This is especially important when teachers are asked to mark essays written by students at lower levels and then asked to mark those of advanced students or vice versa.

In addition, two different teachers can grade the papers, and the two scores can be averaged unless a big discrepancy occurs in-between, as in the case of the preparatory school of foreign languages at Anadolu University. However, one of the raters can assess the paper using an analytic writing assessment profile, yet the other rater can assign an impressionistic mark on holistic-bases.

5.5. Suggestions for Further Studies

The research presented in this study investigated the potential impacts of students' accuracy in language use in papers and the possible effects of teaching experience of raters on the scores assigned to essays using analytic writing criteria. Consequently, no significant difference was found between the raters in terms of their professional background, yet both groups of teachers' sub-scores and total scores were influenced by the students' accurate use of grammar and thus assigned higher marks.

First and most important of all, one issue that can stimulate further research is the results itself found in this study. It could be investigated why an increase occurs in especially two of the sub-components, content and organization when teachers mark grammatically accurate papers. Thereby, an answer could be provided to the question of whether the increase was due to the teachers themselves, or if there is a positive correlation between the quality of language use and the quality of organization found in a paper.

Apart from this, with a similar research design, the errors that belong to vocabulary use could be corrected and the quality and variety of vocabulary used in papers can be improved so that it would be possible to see if the quality of vocabulary use has an effect on raters' scores. Furthermore, in a different study with a similar research design, the organization of ideas in papers could be improved.

Moreover, whether similar results would be obtained can be investigated if this study were replicated by using papers with a different rhetorical structure, such as an argumentative essay.

Next, since the study included 20 essays for the determination of the possible impact of students' accurate use of grammar, it would be worthwhile examining whether a larger amount of essays under exam conditions reveal different results in terms of other factors like fatigue, stress and time constraints which might influence raters' marking performance.

Another concern for further research may entail the replication of the present study involving native speakers of English in the grading process as the third group of subjects so as to see whether nationality of raters is also a factor that influences the scores.

Lastly, if this study were replicated at other institutions with different teachers as subjects with different demographic background, it is a question open to discussion whether the results would be similar or not.

Appendix A

ESL Composition Profile

ESL Composition Profile

RANGE	*CONTENT* CRITERIA
30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail
21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic
16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate

RANGE	*ORGANIZATION* CRITERIA
20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ supported • succinct • well-organized • logical sequencing • cohesive
17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing
13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development
9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate

RANGE	*VOCABULARY* CRITERIA
20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register
17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage but meaning not obscured
13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • meaning confused or obscured
9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate

Appendix A
(Continued)

RANGE	*LANGUAGE USE* CRITERIA
25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles,pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured
10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate

RANGE	*MECHANICS* CRITERIA
5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing
4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured
3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • meaning confused or obscured
2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate

ESL Composition Profile developed by Jacobs et al (1981)

Appendix B

User Guide for the ESL Composition Profile

THE ESL COMPOSITION PROFILE
- A GUIDE TO THE PRINCIPLES OF WRITING -

The Extended Profile Criteria

Since the criteria descriptors are only shorthand reminders of larger concepts in composition, a clear understanding of them is essential for effective use of the PROFILE. The concepts embody the essential principles of writing -- the rules, conventions,and guidelines -- that writers must observe to create a successful piece of writing. This section presents a detailed description of the concepts represented by the PROFILE criteria descriptors at the *Excellent to Very Good* mastery level. The other three levels of competence should be thought of as varying degrees of these extended criteria for excellent writing, with the primary distinguishing factor being the degree to which the writer's intended *meaning* is successfully delivered to the reader or is diminished or completely lost by insufficient mastery of the criteria for excellence. The PROFILE's first two mastery levels in each component (*Excellent to Very Good* and *Good to Average*) both indicate that successful communication has occurred (although differing in degree), whereas the two lower levels (*Fair to Poor* and *Very Poor*) suggest there is a communication breakdown of some sort -- either partial or complete. *Effect on meaning* thus becomes the chief criterion for distinguishing the degree to which the writer has mastered the criteria for excellent writing.

CONTENT

30-27	**EXCELLENT TO VERY GOOD: knowledgeable*substantive*thorough development of thesis* relevant to assigned topic**
26-22	**GOOD TO AVERAGE: some knowledge of subject* adequate range* limited development of thesis* mostly relevant to topic, but lacks detail**
21-17	**FAIR TO POOR: limited knowledge of subject* little substance* inadequate development of topic**
16-13	**VERY POOR: does not show knowledge of subject* non-substantive* not pertinent * OR not enough to evaluate**

Appendix B
(Continued)

DESCRIPTOR	CRITERIA
Knowledgeable	Is there understanding of the subject? Are facts or other pertinent information used? Is there recognition of several aspects of the subject? Are the interrelationships of these aspects shown?
Substantive	Are several main points discussed? Is there sufficient detail? Is there originality with concrete details to illustrate, define, compare, or contrast factual information supporting the thesis?
Thorough development of thesis	Is the thesis expanded enough to convey a sense of completeness? Is there a specific method of development (such as comparison/contrast, illustration, definition, example, description, fact, or personal experience)?
Relevant to assigned topic	Is all information clearly pertinent to the topic? Is extraneous material excluded?

ORGANIZATION

20-18	EXCELLENT TO VERY GOOD: fluent expression* ideas clearly stated/supported* succinct*well-organized*logical sequencing*cohesive
17-14	GOOD TO AVERAGE: somewhat choppy*loosely organized but main ideas stand out*limited support* logical but incomplete sequencing
13-10	FAIR TO POOR: non-fluent* ideas confused or disconnected* lacks logical sequencing and development
9-7	VERY POOR: does not communicate* no organization*OR not enough to evaluate

Appendix B
(Continued)

DESCRIPTOR **CRITERIA**

Fluent expression Do the ideas flow, building on one another? Are there introductory and concluding paragraphs? Are there effective transition elements -- words, phrases, or sentences -- which link and move ideas both within and between paragraphs?

Ideas clearly stated/supported Is there a clearly stated controlling idea or central focus to the paper (a thesis)? do topic sentences in each paragraph support, limit, and direct the thesis?

Succinct Are all ideas directed concisely to the central focus of the paper, without digression?

Well-organized Is the overall relationship of ideas within and between paragraphs clearly indicated? Is there a beginning, a middle, and an end to the paper?

Logical sequencing Are the points logically developed, using a particular sequence such as time order, space order, or importance? Is this development indicated by appropriate transitional markers?

Cohesive Does each paragraph reflect a single purpose? Do the paragraphs form a unified paper?

VOCABULARY

20-18	**EXCELLENT TO VERY GOOD: sophisticated range* effective word/idiom choice and usage* word form mastery * appropriate register**
17-14	**GOOD TO AVERAGE: adequate range* occasional errors of word/idiom form, choice, usage *but meaning not obscured***
13-10	**FAIR TO POOR: limited range* frequent errors of word/idiom form, choice, usage* *meaning confused or obscured***
9-7	**VERY POOR: essentially translation* little knowledge of English vocabulary, idioms, word form* OR not enough to evaluate**

Appendix B
(Continued)

DESCRIPTOR **CRITERIA**

Sophisticated range Is there facility with words and idioms: to convey intended information, attitudes, feelings? to distinguish subtleties among ideas and intentions? to convey shades and differences of meaning? to express the logic of ideas? Is the arrangement and interrelationship of words sufficiently varied?

Effective word/idiom choice and usage In the context in which it is used, is the choice of vocabulary accurate? Idiomatic? Effective? concise? Are strong active verbs and verbals used where possible? Are phrasal and prepositional idioms correct? Do they convey the intended meaning? Does word placement give the intended message? emphasis? Is there an understanding of synonyms? antonyms? homonyms? Are denotative and connotative meanings distinguished? Is there effective repetition of key words and phrases? Do transition elements mark shifts in thought? pace? emphasis? Tone?

Word form mastery Are prefixes, suffixes, roots, and compounds used accurately and effectively? Are words correctly distinguished as to their function (noun, verb, adjective, adverb)?

Appropriate register Is the vocabulary appropriate to the topic? to the audience? to the tone of the paper? to the method of development? Is the vocabulary familiar to the audience? Does the vocabulary make the intended impression?

LANGUAGE USE

25-22	**EXCELLENT TO VERY GOOD: effective complex constructions* few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions**
21-18	**GOOD TO AVERAGE: effective but simple constructions* minor problems in complex constructions * several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured***
17-11	**FAIR TO POOR: major problems in simple/complex constructions* frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions * *meaning confused or obscured***
10-5	**VERY POOR: virtually no mastery of sentence construction rules* dominated by errors* does not communicate* OR not enough to evaluate**

Appendix B
(Continued)

DESCRIPTOR	**CRITERIA**
Effective complex constructions	Are sentences well-formed and complete, with appropriate complements? Are single-word modifiers appropriate to function? Are they properly formed, placed, sequenced? Are phrases and clauses appropriate to function? complete? properly placed? Are introductory *It* and *There* used correctly to begin sentences and clauses? Are main and subordinate ideas carefully distinguished? Are coordinate and subordinate elements linked to other elements with appropriate conjunctions, adverbials, relative pronouns, or punctuation? Are sentence types and length varied? Are elements parallel? Are techniques of substitution, repetition, and deletion use effectively?
Agreement	Is there basic agreement between sentence elements: auxiliary and verb? subject and verb? pronoun and antecedent? adjective and noun? nouns and quantifiers?
Tense	Are verb tenses correct? properly sequenced? Do modals convey intended meaning? time?
Number	Do nouns, pronouns, and verbs convey intended quality?
Word order/function	Is normal word order followed except for special emphasis? Is each word, phrase, and clause suited to its intended function?
Articles	Are *a, an,* and *the* used correctly?
Pronouns	Do pronouns reflect appropriate person? gender? number? function? referent?
Prepositions	Are prepositions chosen carefully to introduce modifying elements? Is the intended meaning conveyed?

Appendix B
(Continued)

MECHANICS

5	**EXCELLENT TO VERY GOOD: demonstrates mastery of conventions* few errors of spelling, punctuation, capitalization, paragraphing**
4	**GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured***
3	**FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing * poor handwriting* *meaning confused or obscured***
2	**VERY POOR: no mastery of conventions* dominated by errors of spelling, punctuation, capitalization, paragraphing* handwriting illegible* OR not enough to evaluate**

DESCRIPTOR CRITERIA

Spelling Are word spelled correctly?

Punctuation Are periods, commas, semicolons, dashes, and question marks used correctly? Are words divided correctly at the end of lines?

Capitalization Are capital letters used where necessary and appropriate?

Paragraphing Are paragraphs indented to indicate when one sequence of thought ends and another begins?

Handwriting Is handwriting easy to read, without impeding communication?

Jacobs, et al (1981)

APPENDIX C

6-Point Holistic Scoring Criteria

HOLISTIC CRITERIA

City University of New York Freshman Wills Assessment Program Evaluation Scale for Writing Assessment Test

6: The essay is completely organized and the ideas are expressed in appropriate language. A sense of pattern or development is present from beginning to end. The writer supports assertions with explanation or illustrations. Sentences reflect a command of syntax within the ordinary range of standard written English. Grammar, punctuation, and spelling are generally correct.

4-5: The writer introduces some point or idea and demonstrates an awareness that development or illustration is called for. The essay presents a discernible pattern or organization, even if there are occasional digressions. The essay demonstrates sufficient command of vocabulary to convey, without serious distortion or excessive simplification, the range of the writer's ideas. Sentences reflect a sufficient command of syntax to ensure reasonable clarity of expression. The writer generally avoids both the monotony of rudimentary syntax and the incoherence created by angled syntax. The writer demonstrates through punctuation an understanding of the boundaries of the sentence. The writer spells the common words of the language with a reasonable degree of accuracy. Exceptions can be made for the so-called spelling demons which frequently trouble even an advanced writer. The writer shows the ability to use regularly, but not necessarily faultlessly, the common forms of agreement and of grammatical inflection in standard written English.

Appendix C
(Continued)

2-3: An idea or point is suggested, but is underdeveloped or presented in a purely repetitious way. The pattern of the essay is somewhat random and relationships between sentences and paragraphs are rarely signaled. The essay is restricted to a very narrow range of language, so that the vocabulary chosen frequently does not serve the needs of the writer. The syntax of the essay is not sufficiently stable to ensure reasonable clarity of expression. The syntax is rudimentary or tangled. The writer frequently commits errors of punctuation which obscure sentence boundaries.The writer spells the common words of the language with only intermittent accuracy The essay reveals recurrent grammatical problems; if there are only occasional problems, this may be due to the extremely narrow range of syntactical choices the writer has used.

1: The essay suffers from general incoherence and has no discernible pattern of organization. It displays a high frequency of error in the regular features of standard written English. Lapses in punctuation, spelling and grammar often frustrate the reader. Or: The essay is so brief that any reasonably accurate judgment of the writer's competence is impossible.

Gray and Slaughter (1980, in Perkins 1983)

Appendix D

Sample Original Version Essays Used in the First Grading

(Original Essay Number-1)

Working Too Much

People need more workers day by day. Therefore, working too much is important part of modern life. Nevertheless, working too much brings about some Problems which are social, healty and Family for several reasons.

Admittedly, working too much causes social Problems because of lack of time. Hardworking people work hard inspite of going to Cinema, or disco because working steals a lot of time From their lives. They can't separate time to different areas owing to working too much. They usually work from early mornings to midnights.

Secondly, working too much brings about healty problems for their body. Some people's body doesn't like working hard. If their bosses order to work hard, they will be sick. They suffer from this situations. Their mental activity doesn't work well because their brains don't rest, Maybe in the Future, they will be sick in sixty or seventy ages because of working too much.

Lastly, working too much causes Family Problems such as Shy children, or wives problem. Hardworking people don't consider their Children because of working lot. Besides this situation, they don't consider their wives because they think that if they consider their wives, their wives want new things after that their consantrate is spoilt. Because of working hard, their children don't grow well and than they become shy, stable etc. Because of working too much, they don't go anywhere with their family For vacation.

In summary, day by day workers becomes very important around the world, but few people think worker. They don't think their own situation as well. Therefore, some problems occur in their lives like social, healty and Family problems. I think, if they thought their own position, they will be better than this time.

Appendix D
(Continued)

(Original Essay Number-2)

WORKING TOO MUCH

Up to now, people and animals have worked too much. Nowadays, everybody is financially strained. Because of this reason, we have got a lot of responsibilities, so we must work. Working too much has got three main effect on our life.

The first and most important effect is that your health is effected badly. Especially, if you work with computer all day, you may have to wear glasses. For example, my sister have been working for three years and she is always opposite computer. We took her to hospital two months ago and now she wears glasses. Not only your eyes have got problems when you work too much, but also you can be very tired all day.

The second effect is that you don't find any time for social activities. If you work too much, you must forget parties, theatre and something else like this. You may have an appointment at 8:00 with your friends but you have to work until at 7:30. Because of working too much you can't go out with your friends. You can't find any time for you because of working too much.

The third effect is that you have a lot of problems in terms of your psycologhy. If you work too much, you can feel bad. You want to get rid of complain about everything, especially your job. According to research of American hospital, working too much make people agressive.

In conclusion, health is effected badly, you haven't got social life and your problems in terms of psycologhy gets worse day by day when you work too much.

Appendix D
(Continued)

(Original Essay Number-3)

WORKING TOO MUCH

People born, live and die in their life. They need something for their life. They work, because necessities are important. These necessities are meal, cleaning machine etc... These are very important necessities. People work too much for these things. It has got some effect working too much.

The first and major effect is healty. Some people want more than one. So they work too much. They don't think their life. People need relax, sleep etc... If they don't do these necessity, they will lost their healty. For example there will be some psycologic and body problems.

The second effect is family. People who are working too much don't interested in their family. If they have children, children will not like their father or mother. Every children want to go to somewhere or speak with their family. Children can be unhappy. Also husband and wife can lost their happy.

The last effect is social life. People want to go to somewhere, they want to watch movie or go to theater. They work always. We can't wait happness in such life. These are very important necessities and every people must do these activities. If they don't do these things, they will be unhappy.

In conclusion, these effects are reason of people's happness or unhappness. People must think that ''how can be life better than now''. We must go on way of the happness. Working too much can be bad both our own healty and our life, we must live happy because we live only once.

Appendix E

Sample Corrected-Version Essays Used in the Second Grading

(Corrected Essay Number-1)

Working Too Much

People need more workers day by day. Therefore, working too much is **an** important part of modern life. Nevertheless, working too much brings about some Problems which are **about** social, healty and Family for several reasons.

Admittedly, working too much causes social Problems because of lack of time. Hardworking people work hard instead of going to **the** Cinema, or **to** disco because working steals a lot of time From their lives. They can't separate time to different areas owing to **their** working too much. They usually work from early **morning** to **midnight**.

Secondly, working too much brings about healty problems for their **bodies**. Some people's **bodies don't** like working hard. If their bosses order **them** to work hard, they will be sick. They suffer from this **situation**. Their mental activity doesn't work well because their brains don't rest, Maybe in the Future, they will be sick **at the age of sixty or seventy** because of working too much.

Lastly, working too much causes Family Problems such as Shy children, or **problems with wives**. Hardworking people don't consider their Children because of working **a** lot. Besides this situation, they don't consider their wives because they think that if they consider their wives, their wives **will** want new things after that their consantrate **will be** spoilt. Because of **their** working hard, their children don't grow well and than they become shy, stable etc. Because of working too much, they don't go anywhere with their family For vacation.

In summary, day by day workers **become** very important around the world, but few people think **about workers**. They don't think **about** their own **situations, either**. Therefore, some problems occur in their lives like social, healty and Family problems. I think, if they thought **of** their own **positions**, they **would** be better than this time.

Appendix E
(Continued)

(Corrected Essay Number-2)

WORKING TOO MUCH

Up to now, people and animals have worked too much. Nowadays, everybody is financially strained. Because of this reason, we have got a lot of responsibilities, so we must work. Working too much has got three main **effects** on our **lives**.

The first and most important effect is that your health is **badly** effected. Especially, if you work with **computers** all day, you may have to wear glasses. For example, my sister **has** been working for three years and she is always **in front of computers**. We took her to hospital two months ago and now she wears glasses. Not only **do** your eyes have ~~got~~ problems when you work too much, but also you can be very tired all day.

The second effect is that you don't find any time for social activities. If you work too much, you must forget **about** parties, **theatres** and something else like this. You may have an appointment at 8:00 with your friends but you have to work until ~~at~~ 7:30. Because of working too much you can't go out with your friends. You can't find any time for **yourself** because of working too much.

The third effect is that you have a lot of problems in terms of your psycologhy. If you work too much, you can feel bad. You want to get rid of **complaining** about everything, especially **about** your job. According to research of American hospital, working too much **makes** people agressive.

In conclusion, health is **badly** effected, you haven't got **a** social life and your problems in terms of psycologhy **get** worse day by day when you work too much.

Appendix E
(Continued)

(Corrected Essay Number-3)

WORKING TOO MUCH

People **are** born, live and die in their **lives**. They need something for their life. They work, because necessities are important. These necessities are **meals**, cleaning **machines** etc... These are very important necessities. People work too much for these things. **Working too much** has got some **effects**.

The first and major effect is **on** healty. Some people want more than one. So they work too much. They don't think **of** their **lives**. People need **to** relax, **to** sleep etc... If they don't do these **necessities**, they will **lose** their healty. For example there will be some psycologic and body problems.

The second effect is **on** family. People who are working too much **are not** interested in their **families**. If they have children, **the** children will not like their father or mother. Every **child wants** to go ~~to~~ somewhere or **to** speak with his or her family. Children can be unhappy. Also **husbands** and **wives** can **lose** their happy.

The last effect is **on** social life. People want to go ~~to~~ somewhere, they want to watch **movies** or go to **the** theater. They **always** work. We can't wait **for** happness in such **a** life. These are very important necessities and **all** people must do these activities. If they don't do these things, they will be unhappy.

In conclusion, these effects are **the reasons for** people's happness or unhappness. People must think that ''how can life **be** better than now''. We must go on **the** way of ~~the~~ happness. Working too much can be bad **for** both our own healty and our life, we must live happy because we live only once.

Appendix F

The Scoring Rubric Used for the Assessment of Essays in the Previous Year's Final Exam

FINAL EXAM GRADING STANDARDS

TASK ACHIEVEMENT

40- 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 with 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 ORGANIZATION

40- The paragraphs of the essay are clearly and logically organized.
The text is organized into a clear introduction, body and conclusion.

30- The paragraphs of the essay are not logically organized.
Some part of the introduction, body and/or conclusion is incomplete.

20- The paragraphs of the essay are not organized.
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- 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.

PENALTY -10 FOR NOT ANSWERING THE QUESTION

A scoring rubric developed by Oruç (1999).

Appendix G

Scores Assigned to the Original 20 Essays in the First Grading And to the 20 Corrected Versions in the Second Grading by Inexperienced And Experienced Teachers

Inexperienced Rater NUMBER 1

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	27	25	14	19	13	18	5	5	18	23	77	83	5	78
2	22	20	13	16	13	15	3	5	11	19	62	77	8	69
3	24	21	12	14	13	18	4	5	13	18	66	77	5	72
4	19	22	18	19	12	16	4	3	16	22	69	77	6	71
5	25	18	11	16	16	18	4	4	12	19	68	79	7	72
6	18	24	18	15	16	14	3	3	11	19	66	75	8	67
7	18	28	15	18	16	16	5	4	15	21	69	83	6	77
8	25	28	11	15	14	13	4	3	20	24	74	79	4	75
9	21	18	12	17	14	17	4	5	14	21	65	79	7	72
10	24	22	15	16	11	17	3	5	11	20	64	80	9	71
11	19	21	14	17	15	13	4	5	12	18	64	73	6	67
12	27	20	12	19	14	18	3	4	16	21	72	78	5	73
13	23	24	18	13	18	15	5	5	12	20	76	86	8	78
14	20	21	17	19	17	14	5	4	19	23	78	83	4	79
15	27	26	15	16	18	16	5	5	18	24	83	87	6	81
16	22	21	16	11	18	17	4	4	11	20	71	82	9	73
17	23	28	18	19	18	17	5	4	14	19	78	88	5	83
18	21	21	15	12	16	15	3	3	15	21	70	78	6	72
19	27	25	16	16	12	18	4	5	17	21	76	81	4	77
20	26	27	18	14	16	16	3	3	12	20	75	85	8	77

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 2

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	16	21	16	15	18	17	2	5	14	19	66	75	5	70
2	23	22	10	15	11	14	2	3	13	18	59	70	5	65
3	20	25	11	14	11	11	5	5	14	21	61	75	7	68
4	21	21	14	13	16	18	3	4	10	19	64	76	9	67
5	21	23	13	19	14	12	3	5	17	23	68	78	6	72
6	21	20	11	18	10	11	2	3	10	18	54	71	8	63
7	15	21	16	16	17	17	5	4	13	20	66	79	7	72
8	21	28	16	14	15	15	2	5	17	22	71	85	5	80
9	24	25	15	14	11	17	4	3	10	19	64	80	9	71
10	18	20	17	12	15	18	2	4	10	16	62	76	6	70
11	17	21	14	18	11	11	2	3	11	18	55	71	7	64
12	21	25	14	18	15	12	3	3	21	23	74	77	2	75
13	16	26	17	15	12	15	4	5	20	23	69	76	3	73
14	22	23	12	19	14	17	2	5	21	25	71	80	4	76
15	23	25	17	12	13	16	2	5	15	22	70	83	7	76
16	21	27	12	12	11	12	5	5	12	21	61	79	9	70
17	19	25	14	18	10	17	4	5	18	22	65	80	4	76
18	21	22	12	13	10	17	3	3	13	19	59	74	6	68
19	24	22	15	15	14	18	3	3	10	19	66	80	9	71
20	21	23	14	14	12	17	3	4	19	21	69	77	2	75

Lu1&Lu2: The sub-scores assigned to the language use sub-component in the first and second gradings

T2: The total scores assigned in the second grading

Df: The difference between Lu2 and Lu1

E: The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 3

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	23	17	10	19	12	14	3	4	13	18	61	70	5	65
2	22	28	17	10	18	14	4	4	12	19	73	82	7	75
3	18	25	13	19	16	15	5	4	19	22	71	77	3	74
4	20	22	15	17	18	18	5	3	17	21	75	77	4	73
5	22	23	10	18	14	16	5	4	18	23	69	77	5	72
6	22	23	10	10	11	14	3	3	11	20	57	70	9	61
7	26	23	11	13	18	17	3	5	12	19	70	78	7	71
8	27	27	11	17	12	13	4	4	18	21	72	76	3	73
9	25	20	12	19	15	15	4	3	12	18	68	78	6	72
10	19	17	15	12	13	18	5	4	13	19	65	77	6	71
11	25	23	17	12	13	18	3	5	15	21	73	84	6	78
12	19	24	16	18	17	18	4	5	19	23	75	82	4	78
13	27	28	10	19	17	13	4	5	16	20	74	80	4	76
14	22	17	10	19	13	17	4	3	16	21	65	73	5	68
15	23	28	12	15	15	13	4	4	17	23	71	80	6	74
16	18	25	12	18	15	11	5	4	19	22	69	78	3	75
17	27	24	12	18	17	12	4	4	11	18	71	77	7	70
18	23	25	12	13	18	15	4	5	12	17	69	80	5	75
19	20	26	14	13	17	15	4	4	13	19	68	80	6	74
20	25	26	14	16	18	15	5	5	17	20	79	83	3	80

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 4

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	25	26	13	15	16	11	3	5	13	17	70	76	4	72
2	19	22	14	17	18	18	4	4	13	19	68	81	6	75
3	20	21	10	17	18	14	3	3	11	20	62	76	9	67
4	18	22	13	18	12	11	5	3	13	19	61	77	6	71
5	20	27	12	19	18	16	5	4	19	19	74	79	0	79
6	26	25	15	17	14	15	3	3	12	18	70	77	6	71
7	19	28	17	15	15	13	5	4	17	18	73	77	1	76
8	20	28	16	18	18	11	5	3	17	23	76	84	6	78
9	24	27	12	19	12	13	3	4	18	21	69	78	3	75
10	23	20	12	14	17	18	5	4	17	20	74	76	3	73
11	22	19	10	16	12	15	4	3	11	20	59	73	9	64
12	19	27	12	17	15	17	5	3	15	19	66	79	4	75
13	27	21	10	19	16	16	4	5	12	18	69	79	6	73
14	27	24	14	19	15	15	3	5	21	21	80	84	0	84
15	23	19	16	17	14	18	5	3	11	20	69	84	9	75
16	21	25	18	19	14	15	5	3	12	23	70	85	11	74
17	23	23	15	18	17	15	4	5	14	19	73	83	5	78
18	20	25	13	17	12	11	3	4	18	21	66	73	3	70
19	25	27	17	15	18	18	3	5	15	20	78	90	5	85
20	26	22	18	15	17	18	4	5	11	20	76	89	9	80

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 5

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	25	18	11	19	17	13	5	4	11	18	69	79	7	72
2	22	25	10	19	17	14	5	5	19	21	73	79	2	77
3	19	24	13	17	17	18	4	3	19	19	72	75	0	75
4	21	19	15	15	12	16	5	3	12	18	65	72	6	66
5	26	20	12	16	11	18	3	4	14	19	66	74	5	69
6	18	17	14	15	13	16	3	3	12	21	60	71	9	62
7	23	19	12	15	16	15	3	3	16	20	70	73	4	69
8	25	28	15	10	12	12	4	4	12	21	68	81	9	72
9	19	22	16	12	11	18	4	3	12	17	62	71	5	66
10	21	20	12	15	12	16	3	4	16	21	64	72	5	67
11	21	20	13	19	17	16	5	5	14	20	70	79	6	73
12	23	17	13	17	15	14	3	5	11	19	65	78	8	70
13	23	22	14	19	15	16	4	5	18	22	74	81	4	77
14	27	28	14	18	15	17	5	3	18	20	79	84	2	82
15	19	22	18	15	17	14	5	3	11	20	70	81	9	72
16	22	22	16	19	16	18	5	4	19	22	78	82	3	79
17	22	25	18	19	17	16	5	5	18	23	80	87	5	82
18	23	23	10	16	14	13	3	5	15	18	65	72	3	69
19	25	23	12	15	13	16	5	5	15	17	70	74	2	72
20	21	22	18	19	16	18	5	5	14	19	74	84	5	79

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 6

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	25	19	11	14	12	15	4	3	11	17	63	72	6	66
2	25	23	13	13	17	18	5	5	12	19	72	82	7	75
3	23	26	12	17	18	13	5	3	12	20	70	78	8	70
4	26	28	13	15	14	13	4	5	15	21	72	79	6	73
5	19	28	16	13	16	16	4	4	16	20	71	81	4	77
6	24	23	11	16	18	18	5	3	11	16	69	79	5	74
7	22	18	16	19	14	18	3	4	12	19	67	79	7	72
8	24	26	14	14	12	12	3	5	13	19	66	74	6	68
9	26	25	10	12	12	15	3	4	15	18	66	72	3	69
10	22	28	13	15	16	15	5	4	17	19	73	77	2	75
11	26	18	10	15	13	18	3	3	13	18	65	71	5	66
12	19	17	10	19	14	17	5	3	18	23	66	77	5	72
13	27	17	12	17	13	17	3	3	17	19	72	73	2	71
14	26	28	10	18	16	14	5	5	20	21	77	84	1	83
15	18	20	17	19	14	15	4	5	21	21	74	75	0	75
16	27	28	12	13	18	18	5	3	13	20	75	85	7	78
17	24	28	18	15	18	16	5	5	13	21	78	90	8	82
18	26	20	10	14	18	16	3	5	11	19	68	78	8	70
19	25	24	16	19	16	15	5	5	15	19	77	82	4	78
20	27	28	16	17	15	14	5	3	11	18	74	84	7	77

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 7

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	24	25	10	12	14	18	3	5	19	19	70	72	0	72
2	25	22	15	14	14	17	5	4	13	19	72	80	6	74
3	17	21	18	14	18	18	5	5	15	21	73	81	6	75
4	26	24	12	11	12	12	3	4	11	17	64	72	6	66
5	21	24	18	19	11	12	5	3	13	19	68	80	6	74
6	27	22	13	19	13	16	5	4	14	18	72	79	4	75
7	24	24	15	12	13	16	5	3	13	21	70	80	8	72
8	26	24	10	18	12	12	5	4	18	22	71	78	4	74
9	21	26	13	13	13	16	5	3	20	21	72	75	1	74
10	26	24	12	12	12	14	4	3	13	17	67	72	4	68
11	23	19	10	17	11	15	3	4	16	19	63	69	3	66
12	25	19	10	19	15	15	3	5	18	21	71	78	3	75
13	20	23	11	18	12	13	3	3	17	21	63	74	4	70
14	19	18	13	13	16	16	3	3	13	20	64	73	7	66
15	20	20	11	18	11	11	3	4	19	20	64	69	1	68
16	19	24	14	19	17	13	3	3	17	22	70	75	5	70
17	24	19	14	18	17	18	3	4	14	19	72	80	5	75
18	20	22	15	13	16	15	5	5	12	18	68	75	6	69
19	27	25	14	12	14	18	4	4	12	19	71	81	7	74
20	23	23	12	17	18	15	3	5	19	22	75	81	3	78

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 8

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	22	24	17	12	11	14	3	3	11	19	64	74	8	66
2	18	20	14	19	18	16	5	5	11	17	66	77	6	71
3	23	24	13	14	11	17	3	4	13	21	63	78	8	70
4	23	22	17	18	14	18	3	5	14	20	71	83	6	77
5	25	20	16	19	14	13	4	4	11	18	70	77	7	70
6	17	21	10	13	16	15	3	3	15	21	61	72	6	66
7	17	21	17	12	15	12	3	4	11	19	63	72	8	64
8	23	18	14	19	12	14	3	4	13	19	65	72	6	66
9	21	21	15	13	13	17	3	3	11	19	63	75	8	67
10	17	23	13	17	17	12	4	3	13	18	64	71	5	66
11	24	23	10	13	17	13	3	4	11	21	65	76	10	66
12	20	19	13	17	16	18	3	4	16	22	68	78	6	72
13	19	20	16	19	12	16	4	5	14	19	65	78	5	73
14	22	28	15	18	18	12	5	4	14	21	74	84	7	77
15	26	19	12	18	13	17	4	5	16	19	71	76	3	73
16	21	24	11	14	13	16	4	4	17	20	66	77	3	74
17	27	24	15	17	15	18	5	5	15	20	77	87	5	82
18	21	22	18	13	14	15	5	4	11	17	69	79	6	73
19	25	19	16	18	12	15	3	5	16	20	72	77	4	73
20	26	18	15	18	15	16	3	5	12	21	71	84	9	75

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 9

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	25	22	17	17	13	16	4	3	11	19	70	80	8	72
2	25	22	10	15	14	15	3	5	12	20	64	76	8	68
3	18	21	15	17	13	18	4	3	18	21	68	73	3	70
4	24	24	17	19	17	17	4	4	15	19	77	84	4	80
5	22	28	14	17	18	16	5	3	12	18	71	81	6	75
6	26	17	10	17	15	16	3	5	11	19	65	74	8	66
7	21	23	12	14	18	17	4	5	12	20	67	78	8	70
8	20	24	17	15	18	18	5	5	17	22	77	84	5	79
9	19	20	12	14	17	13	3	5	12	19	63	76	7	69
10	24	18	13	13	13	14	3	3	11	17	64	73	6	67
11	18	23	17	13	12	18	4	3	15	22	66	75	7	68
12	22	19	13	19	11	13	3	3	16	21	65	74	5	69
13	20	28	17	13	16	15	5	5	14	20	72	78	6	72
14	24	23	13	19	15	17	4	4	18	23	74	81	5	76
15	21	27	14	18	17	13	3	3	14	21	69	82	7	75
16	21	27	17	15	17	18	5	5	15	19	75	84	4	80
17	26	28	16	19	18	18	3	5	17	22	80	89	5	84
18	22	19	13	18	12	12	5	3	15	21	67	74	6	68
19	24	24	16	16	11	13	4	3	16	19	71	76	3	73
20	26	27	16	14	16	15	5	5	11	20	74	86	9	77

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 10

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	22	27	15	18	18	12	3	3	11	19	69	79	8	71
2	21	20	16	16	14	18	3	5	14	18	68	79	4	75
3	25	22	10	19	15	13	3	3	11	20	64	77	9	68
4	25	20	13	16	13	15	3	4	12	17	66	74	5	69
5	23	25	12	16	16	14	3	3	14	18	68	73	4	69
6	23	20	12	19	14	16	4	3	13	17	66	73	4	69
7	24	25	13	11	13	13	3	5	11	19	64	73	8	65
8	19	21	16	15	13	12	4	4	12	20	64	73	8	65
9	19	21	16	17	16	15	3	3	11	20	65	79	9	70
10	18	19	14	16	18	17	5	5	15	21	70	80	6	74
11	18	21	12	18	17	14	4	3	14	20	65	73	6	67
12	26	23	13	14	15	18	3	5	14	19	71	78	5	73
13	22	28	16	14	16	13	4	3	14	19	72	78	5	73
14	25	24	14	16	14	18	5	5	17	21	75	84	4	80
15	21	20	11	11	16	16	3	5	13	19	64	72	6	66
16	26	21	10	19	16	14	4	4	14	21	70	79	7	72
17	18	23	18	18	17	18	5	5	12	18	70	85	6	79
18	24	21	11	16	16	17	5	5	13	19	69	76	6	70
19	25	23	14	15	13	17	5	5	17	23	74	81	6	75
20	21	23	18	19	18	18	5	5	17	22	79	85	5	80

Lu1 & Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Inexperienced Rater NUMBER 11

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	22	19	18	18	17	17	5	4	12	18	74	82	6	76
2	22	28	18	14	15	15	4	5	11	19	70	81	8	73
3	22	25	15	16	15	12	5	5	18	22	75	80	4	76
4	20	24	14	19	17	17	4	4	15	21	70	83	6	77
5	19	27	16	13	15	13	3	3	12	20	65	75	8	67
6	19	25	16	18	13	11	3	5	18	20	69	78	2	76
7	25	28	13	15	11	12	3	4	11	19	63	82	8	74
8	22	27	17	16	15	14	3	4	13	21	70	81	8	73
9	21	26	16	17	13	11	4	3	12	18	66	74	6	68
10	19	19	12	13	14	18	3	4	16	19	64	73	3	70
11	24	25	12	13	12	11	3	3	11	20	62	72	9	63
12	25	23	10	19	16	14	3	5	12	18	66	80	6	74
13	24	23	12	19	13	13	4	3	13	19	66	78	6	72
14	19	21	13	15	13	14	5	5	19	24	69	75	5	70
15	25	24	16	16	12	15	3	3	15	19	71	76	4	72
16	17	23	16	13	15	18	5	4	12	19	65	80	7	73
17	25	27	18	16	17	18	5	4	12	17	77	87	5	82
18	22	23	11	13	12	18	4	4	20	20	69	70	0	70
19	21	27	16	18	17	17	5	4	18	19	77	79	1	78
20	26	22	16	16	16	18	5	5	13	17	76	83	4	79

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 1

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	22	24	11	11	13	12	3	4	11	19	60	71	8	63
2	20	21	14	15	10	13	3	3	12	18	59	69	6	63
3	17	23	13	14	11	15	5	3	16	19	62	70	3	67
4	21	24	18	15	14	13	3	4	12	17	68	77	5	72
5	18	21	15	18	17	14	5	5	11	17	66	79	6	73
6	17	23	14	11	13	11	3	3	11	19	58	68	8	60
7	19	18	14	18	13	15	5	4	14	19	65	77	5	72
8	20	26	16	17	17	15	4	3	11	20	68	84	9	75
9	27	28	12	13	14	15	5	4	13	21	71	84	8	76
10	22	19	14	15	10	14	5	5	11	16	62	72	5	67
11	17	24	14	12	15	12	3	4	12	16	61	67	4	63
12	24	19	18	19	13	15	3	4	12	19	70	83	7	76
13	25	28	13	13	15	18	3	4	13	17	69	79	4	75
14	27	28	12	19	17	18	5	5	16	19	77	89	3	86
15	22	22	16	14	11	11	3	3	11	18	63	72	7	65
16	24	27	14	10	10	16	3	4	17	19	68	76	2	74
17	27	25	16	18	13	18	5	3	13	17	74	86	4	82
18	20	17	11	16	14	13	3	5	13	19	61	74	6	68
19	21	24	15	16	18	16	3	5	17	21	74	81	4	77
20	26	23	10	16	16	15	3	5	11	18	66	79	7	72

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 2

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	23	23	11	18	13	13	3	3	14	18	64	74	4	70
2	27	23	10	15	11	18	3	3	15	21	66	78	6	72
3	20	21	15	16	18	17	4	4	11	17	68	82	6	76
4	27	28	13	14	12	15	5	4	12	19	69	82	7	75
5	21	27	16	11	13	17	4	4	12	18	66	76	6	70
6	25	21	10	15	10	11	3	3	12	21	60	74	9	65
7	26	24	10	17	12	11	3	4	12	19	63	74	7	67
8	21	21	11	19	16	17	5	4	19	23	72	80	4	76
9	21	19	13	12	12	15	3	4	13	20	62	72	7	65
10	21	22	16	13	11	14	5	3	11	18	64	74	7	67
11	19	23	12	14	17	15	5	3	12	16	65	71	4	67
12	24	24	14	17	18	15	5	5	11	17	72	82	6	76
13	18	24	12	18	18	12	3	5	19	20	70	72	1	71
14	19	22	12	17	15	11	3	3	12	19	61	71	7	64
15	18	18	10	12	16	16	5	5	15	18	64	69	3	66
16	21	18	10	18	14	11	3	3	11	18	59	68	7	61
17	20	22	13	17	16	18	3	5	16	19	68	76	3	73
18	27	23	10	12	13	15	3	3	11	15	64	71	4	67
19	23	22	17	16	14	15	5	5	16	19	75	80	3	77
20	21	21	17	14	12	18	4	3	17	19	71	75	2	73

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 3

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	23	22	12	17	13	13	3	3	12	20	63	74	8	66
2	23	24	10	15	12	11	3	3	14	19	62	69	5	64
3	17	27	11	15	17	12	3	3	11	18	59	75	7	68
4	18	22	13	12	17	14	3	5	13	17	64	74	4	70
5	21	25	17	14	12	18	4	5	18	23	72	82	5	77
6	24	20	10	14	16	18	3	5	11	19	64	76	8	68
7	20	22	18	16	15	16	5	5	13	18	71	80	5	75
8	20	24	14	18	17	14	3	5	18	22	72	79	4	75
9	21	25	12	13	13	13	4	4	12	21	62	81	9	72
10	27	27	10	10	13	15	3	4	11	18	64	75	7	68
11	21	27	14	17	17	11	3	5	15	20	70	76	5	71
12	23	21	13	17	10	14	4	3	12	18	62	76	6	70
13	22	24	10	15	17	15	4	5	14	19	67	76	5	71
14	26	23	16	19	13	14	3	5	11	21	69	85	10	75
15	22	24	15	10	15	18	3	5	15	19	70	78	4	74
16	22	25	15	15	11	18	4	4	16	18	68	75	2	73
17	26	24	11	18	16	18	5	5	15	20	73	84	5	79
18	21	21	12	18	13	14	4	5	17	19	67	72	2	70
19	26	21	15	19	13	18	4	5	14	17	72	77	3	74
20	19	23	16	17	12	16	5	5	18	19	70	76	1	75

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 4

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	24	26	12	15	14	12	4	4	11	19	65	78	8	70
2	26	28	15	12	14	14	3	4	11	18	69	81	7	74
3	20	27	13	10	11	11	3	4	12	21	59	73	9	64
4	20	21	12	13	10	13	4	5	18	20	64	68	2	66
5	26	18	14	13	10	17	3	5	11	21	64	79	10	69
6	19	20	10	17	16	11	5	3	11	20	61	71	9	62
7	19	19	14	18	12	15	3	3	16	19	64	69	3	66
8	20	21	10	15	14	12	5	5	15	17	64	69	2	67
9	24	22	10	12	14	16	4	3	12	19	64	76	7	69
10	25	25	16	14	11	15	3	3	11	19	66	76	8	68
11	17	18	10	19	18	12	4	3	11	18	60	70	7	63
12	26	26	18	15	11	17	4	5	13	20	72	82	7	75
13	25	23	13	14	17	17	4	4	16	22	75	83	6	77
14	21	26	13	19	18	18	5	4	19	20	76	79	1	78
15	19	24	11	12	15	11	5	4	11	18	61	74	7	67
16	26	25	10	14	16	15	4	4	11	19	67	79	8	71
17	25	20	12	19	18	18	3	5	15	18	73	80	3	77
18	19	22	17	16	13	12	5	3	11	17	65	75	6	69
19	18	23	14	19	13	11	4	3	13	18	62	72	5	67
20	24	27	11	14	15	13	5	3	16	19	71	76	3	73

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 5

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	23	27	11	16	18	13	3	3	11	17	66	77	6	71
2	21	23	15	12	14	15	3	5	11	19	64	76	8	68
3	22	21	18	17	11	15	4	5	17	18	72	76	1	75
4	17	24	16	12	16	11	3	5	15	19	67	74	4	70
5	23	26	10	12	13	13	4	3	11	21	61	76	10	66
6	17	25	12	16	13	13	4	4	18	24	64	75	6	69
7	17	22	10	15	17	14	5	3	15	18	64	69	3	66
8	18	24	11	13	15	11	5	3	14	17	63	70	3	67
9	22	24	10	18	17	12	5	5	14	19	68	75	5	70
10	18	24	16	16	15	16	5	3	17	18	71	74	1	73
11	20	28	13	10	16	15	4	3	12	19	65	74	7	67
12	23	19	12	18	15	11	3	5	11	19	64	76	8	68
13	26	28	16	14	10	11	3	3	11	18	66	76	7	69
14	21	28	13	15	18	15	3	5	17	20	72	82	3	79
15	18	26	17	14	16	15	3	5	17	18	71	74	1	73
16	20	24	15	18	15	17	5	3	15	17	70	76	2	74
17	26	26	16	18	18	18	5	5	12	18	77	89	6	83
18	17	23	17	12	10	16	5	3	15	19	64	71	4	67
19	21	25	13	13	17	13	4	4	17	19	72	75	2	73
20	26	25	14	17	16	14	4	4	11	20	71	84	9	75

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 6

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	26	28	10	11	18	17	3	3	11	18	68	77	7	70
2	17	22	13	16	10	11	3	4	16	18	59	67	2	65
3	21	18	10	12	16	16	5	4	12	19	64	75	7	68
4	27	27	17	17	18	17	5	4	14	17	81	83	3	80
5	26	20	11	14	10	14	5	4	16	17	68	71	1	70
6	24	21	16	17	10	15	3	3	12	19	65	75	7	68
7	19	25	15	12	12	13	3	3	15	20	64	74	5	69
8	21	21	12	14	12	13	3	3	13	21	61	73	8	65
9	22	21	11	13	16	17	3	4	11	18	63	76	7	69
10	22	28	10	10	15	12	3	3	11	19	61	72	8	64
11	26	24	12	10	10	18	3	3	14	21	65	74	7	67
12	17	21	17	16	14	16	4	3	11	20	63	77	9	68
13	23	21	10	10	12	18	3	5	18	22	66	74	4	70
14	21	21	10	13	15	15	3	3	11	21	60	74	10	64
15	23	18	16	17	15	13	3	5	11	19	68	79	8	71
16	19	27	10	13	12	11	3	3	11	17	55	71	6	65
17	22	27	17	18	18	15	5	4	13	18	75	84	5	79
18	18	20	14	13	13	14	5	3	17	19	67	71	2	69
19	27	22	15	17	16	18	3	5	13	20	74	84	7	77
20	27	25	17	17	18	18	5	5	17	21	84	87	4	83

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 7

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	27	18	11	13	10	14	4	4	11	18	63	71	7	64
2	26	18	10	18	10	11	4	3	14	19	64	70	5	65
3	19	21	13	18	16	11	4	4	13	17	65	74	4	70
4	25	23	16	14	13	16	5	5	14	17	73	76	3	73
5	27	19	11	15	10	18	4	3	15	18	67	74	3	71
6	21	21	10	15	13	11	3	5	14	19	61	69	5	64
7	22	27	16	13	14	16	3	3	15	20	70	77	5	72
8	25	27	18	19	18	16	4	5	14	21	79	88	7	81
9	22	26	15	17	17	13	5	5	14	20	73	84	6	78
10	24	21	10	17	15	15	4	3	11	17	64	73	6	67
11	24	22	13	11	12	18	4	4	11	15	64	71	4	67
12	21	23	18	19	14	16	5	5	17	19	75	79	2	77
13	18	27	13	12	16	15	5	4	15	21	67	77	6	71
14	19	28	16	14	12	12	3	3	18	22	68	75	4	71
15	18	22	11	13	14	11	5	3	16	20	64	69	4	65
16	26	19	16	16	10	18	3	4	12	19	67	77	7	70
17	26	25	16	14	15	16	3	5	14	18	74	83	4	79
18	19	27	16	11	18	16	4	5	15	19	72	79	4	75
19	26	28	16	16	18	18	5	5	15	18	80	86	3	83
20	25	25	17	19	18	15	5	5	16	20	81	84	4	80

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 8

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	25	20	16	14	12	12	3	4	12	19	68	76	7	69
2	27	27	11	15	14	12	5	3	15	20	72	79	5	74
3	26	28	11	17	16	11	3	3	15	20	71	77	5	72
4	19	25	18	17	17	14	5	5	15	19	74	81	4	77
5	21	25	14	13	16	16	4	4	13	16	68	73	3	70
6	22	21	15	17	12	11	4	5	14	19	67	74	5	69
7	27	23	13	18	18	18	5	4	12	18	75	84	6	78
8	27	23	11	16	18	15	3	3	11	16	70	75	5	70
9	18	25	13	18	17	14	4	4	19	24	71	77	5	72
10	22	24	14	12	11	13	3	4	14	19	64	71	5	66
11	24	25	15	18	15	13	4	5	11	16	69	79	5	74
12	26	18	11	17	14	15	4	4	16	22	71	76	6	70
13	26	20	12	17	10	16	3	3	16	19	67	73	3	70
14	21	23	17	18	17	14	5	5	17	20	77	81	3	78
15	25	26	10	12	10	11	3	3	13	16	61	66	3	63
16	19	23	15	15	16	11	3	3	11	16	64	72	5	67
17	21	28	18	15	17	18	5	5	17	20	78	81	3	78
18	22	26	16	16	14	12	5	3	14	20	71	80	6	74
19	23	20	14	12	15	17	3	4	15	18	70	74	3	71
20	21	19	14	13	17	17	5	5	13	17	70	76	4	72

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 9

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	27	22	10	13	11	12	3	5	13	17	64	73	4	69
2	23	28	13	12	13	18	4	4	17	20	70	77	3	74
3	21	19	12	19	16	18	5	5	19	22	73	78	3	75
4	24	26	18	19	17	18	4	3	11	19	74	88	8	80
5	18	24	13	15	13	16	3	3	20	21	67	73	1	72
6	26	24	11	16	11	13	5	3	13	20	66	74	7	67
7	22	25	13	19	12	11	5	4	13	19	65	78	6	72
8	21	26	11	14	17	15	3	3	15	19	67	75	4	71
9	20	24	16	14	17	14	4	5	11	20	68	78	9	69
10	21	20	13	19	13	15	5	5	15	19	67	74	4	70
11	24	19	10	13	15	16	3	4	16	20	68	71	4	67
12	27	23	11	18	14	18	3	5	17	20	72	78	3	75
13	21	20	16	13	14	17	5	3	15	21	71	77	6	71
14	21	28	17	18	17	17	5	4	18	23	78	85	5	80
15	24	21	14	19	15	18	5	3	13	19	71	78	6	72
16	26	25	15	15	11	16	3	4	15	19	70	75	4	71
17	27	23	16	16	10	14	4	4	15	17	72	74	2	72
18	20	20	15	16	17	18	3	5	19	20	74	76	1	75
19	17	26	17	13	17	17	3	4	18	20	72	76	2	74
20	27	25	16	17	10	16	5	3	17	22	75	82	5	77

Lu1&Lu2:	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df:	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

Appendix G

(Continued)

Experienced Rater NUMBER 10

PAPER NUMBER	CONTENT		ORGANIZATION		VOCABULARY USE		MECHANICS		Language Use		Total		Df=Lu2-Lu1	E=T2-Df
	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING	1ST GRADING	2ND GRADING
1	24	20	10	11	13	15	3	3	11	20	61	70	9	61
2	20	19	16	19	14	15	5	5	11	18	66	78	7	71
3	21	19	14	16	13	16	5	5	14	19	67	72	5	67
4	27	22	18	19	12	18	4	4	12	16	73	81	4	77
5	27	20	12	18	13	17	3	5	16	19	71	78	3	75
6	19	20	10	11	15	12	3	3	11	21	58	67	10	57
7	26	24	15	19	14	13	5	5	14	20	74	81	6	75
8	25	26	18	19	18	15	5	5	11	19	77	86	8	78
9	25	24	11	14	12	16	4	3	19	24	71	75	5	70
10	22	24	17	14	10	16	4	3	17	21	70	72	4	68
11	27	21	10	10	10	17	3	4	14	18	64	70	4	66
12	18	28	17	16	14	11	3	3	17	21	69	73	4	69
13	20	26	15	18	17	14	5	3	14	19	71	77	5	72
14	24	24	17	19	18	14	5	5	11	20	75	84	9	75
15	22	25	15	18	12	15	5	5	19	22	73	77	3	74
16	22	26	11	19	18	14	4	5	20	22	75	77	2	75
17	19	28	16	18	16	15	3	3	17	20	71	78	3	75
18	25	23	11	12	12	14	5	5	17	21	70	76	4	72
19	25	24	15	14	10	17	3	3	15	17	68	73	2	71
20	22	19	10	17	13	17	5	4	19	22	69	78	3	75

Lu1&Lu2	The sub-scores assigned to the language use sub-component in the first and second gradings
T2:	The total scores assigned in the second grading
Df	The difference between Lu2 and Lu1
E:	The expected total scores calculated through the subtraction of Df from T2

REFERENCES

Bachman, L. F. (1990). **Fundamental Considerations in Language Testing**. New York: Oxford University Press.

Bahçe, Aysel. (1992). **The Effect of Background Knowledge on the Global Writing Proficiency of EFL Students**. Master Thesis. Anadolu University, Eskişehir.

Baker, David. (1989). **Language Testing: A Critical Survey And Practical Guide**. Great Britain: British Library Cataloguing in Publication Data.

Brossell, Gordon. (1996). **Writing Assessment in florida: A Reminiscence**. In White, E. M. et. al. (Ed), **Assessment of Writing: Politics, Policies, Practices** (pp. 25-33). New York: The Modern Language Association of America.

Brown, J. D. (1988). **Understanding Research in Second Language Learning**. USA: Cambridge University Press.

____ (1996). **Testing in Language Programs**. New Jersey: Prentice-Hall Inc.

Camp, Roberta. (1996). **Response: The Politics of Methodology**. In White, E. M. et. al. (Ed), **Assessment of Writing: Politics, Policies, Practices** (pp. 97-105). New York: The Modern Language Association of America.

Connor-Linton, J. (1995). ''Looking behind the Curtain: What Do L2 Composition Ratings Really Mean?'', *TESOL Quarterly*, 29 (4), 762-765.

Eames, K. and Loewenthal K. (2001). ''Effects of Handwriting And Examiner's Expertise on Assessment of Essays'', *The Journal of Social Psychology*, 130 (6), 831-833.

Elbow, Peter. (1996). **Writing Assessment: Do it better, do it less**. In White, E. M. et. al. (Ed), **Assessment of Writing: Politics, Policies, Practices** (pp. 120-135). New York: The Modern Language Association of America.

Engber, Cheryl A. (1995). ''The Relationship of Lexical Proficiency to the Quality of ESL Compositions'', *Journal of Second Language Writing*, 4 (2), 139-155.

Ferris, D. and Hedgcock, J. S. (1998). **Teaching ESL Composition: Purpose, Process, and Practice**. New Jersey: Lawrence Erlbaum Associates Publishers.

Gay, L. R. (1985). **Educational Evaluation And Measurement: Competencies for Analysis And Application (Second Edtion)**. Ohio: Bell & Howell Company.

Gronlund, N. E. (1988). **How to Construct Achievement Tests**. New Jersey: Prentice-Hall Inc.

Hamp-Lyons, Liz. (1990). **Second Language Writing: Assessment Issues**. In Kroll, B. (Eds), **Second Language Writing: Research Insights for the Classroom**. USA: Cambridge University Press.

____ (1995). ''Rating Non-native Writing: The Trouble with Holistic Scoring'', *TESOL Quarterly*, 29 (4), 759-762.

____ (1996). **The Challenges of Second-Language Writing Assessment**. In White, E. M. et al (Ed), **Assessment of Writing: Politics, Policies, Practices** (pp. 226-241). New York: The Modern Language Association of America.

Hamp-Lyons, Liz and Kroll, B. (1996). ''Issues in ESL Writing Assessment: An Overview'', *College ESL*, 6 (1), 52-71.

Harrison, Andrew. (1983). **A Language Testing Handbook**. Hong Kong: Macmillan Publishers Ltd.

Heaton, J.B. (1975). **Writing English Language Tests**. Hong Kong: Longman.

Henning, Grant. (1986). **Twenty Common Testing Mistakes for EFL Teachers To Avoid**. In Newton, A. C. (Eds). **A Forum Anthology: Selected Articles from the English Teaching Forum**. Washington D.C.: English Language Programs Division Bureau of Educational Affairs United States Information Agency.

Homburg, T. J. (1984). ''Holistic Evaluation of ESL Compositions: Can it be validated objectively?'', *TESOL Quarterly*, 18 (1), 27-45.

Hughes, Arthur (1989). **Testing for Language Teachers**. Great Britain: Cambridge University Press.

Jacobs, Holly L. et. al. (1981). **Testing ESL Composition: A Practical Approach**. Boston: Newbury House.

Janopoulos, M. (1992). "University Faculty Tolerance of NS And NNS Writing Errors: A Comparison'', *Journal of Second Language Writing*, 1 (2), 109-121.

Johnson, D. W. and Johnson, R. T. (2002). **Meaningful Assessment: A Manageable And Cooperative Process**. Boston: Pearson Education Company.

Johnson, R. L., Penny, J. and Gordon P. (2000). ''The Relation between Score Resolution Methods And Interrater Reliability: An Emprical Study of an Analytic Scoring Rubric'', *Applied Measurement in Education*, 13 (2), 121-138.

Kobayashi, H. and Rinnert, C. (1996). ''Factors Affecting Composition Evaluation in an EFL Context: Cultural Rhetorical Pattern And Readers' Background'', *Language Learning*, 46 (3), 397-437.

Kroll, Barbara. (1991). **Teaching Writing in the ESL Context**. In Celce-Murcia M. (Eds), **Teaching English As a Second And Foreign Language**. (pp. 245-263). Boston: Heinle & Heinle Publishers.

Kubiszyn, Tom and Borich, Gary. (1990). **Educational Testing And Measurement: Classroom Application And Practice**. United States of America: Library of Congress Cataloguing-in-Publication Data.

Leki, Ilona. (1992). **Understanding ESL Writers: A Guide for Teachers**. Portsmouth: Heineman.

Lumley, Tom. (2002). ''Assessment Criteria in a Large-Scale Writing Test: What Do They Really Mean to the Raters?'', *Language Testing*, 19 (3), 246-276.

Madsen, Harold S. (1983). **Techniques in Testing**. United Kingdom: Oxford University Press.

Nitko, A. J. (1996). **Educational Assessment of Students (Second Edition)**. New Jersy: Prentice-Hall Inc.

Omaggio, A. (1986). **Teaching Language in Context: Proficiency Oriented Instruction**. Boston: Heinle & Heinle.

Oruç, Nesrin. (1999). **Evaluating the Reliability of Two Grading Systems for Writing Assessment at Anadolu University Prepoartory School**. Bilkent University. Ankara.

Perkins, Kyle. (1983). ''On the Use of Composition Scoring Techniques, Objective Measures And Objective Tests To Evaluate ESL Writing Ability'', *TESOL Quarterly*, 17 (4), 651-671.

Polat, Murat. (2003). **A Study on Developing a Writing Assessment Profile for English Preparatory Program of Anadolu University School of Foreign Languages**. Unpublished M.A. Thesis. Eskişehir.

Porte, G. and Inglesa, de Filologia (1999). ''Where To Draw the Red Line: Error Toleration of Native And Non-Native EFL Faculty'', *Foreign Language Annals*, 32 (4), 426-434.

Ruth, L. and Murphy, S. (1988). **Designing Writing Tasks for the Assessment of Writing**. New Jersey: Ablex Publishing Cooperation.

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

Santos, T. (1988). ''Proffessors' Reactions to the Academic Writing of Non-Native Speaking Students'', *TESOL Quarterly*, 22 (1), 69-90.

Shale, Doug. (1996). **Essay Reliability: Form And Meaning**. In White, E. M. et. al. (Ed), **Assessment of Writing: Politics, Policies, Practices** (pp. 76-96). New York: The Modern Language Association of America.

Shohamy et. al. (1992). ''The Effect of Raters' Background And Training on The Reliability of Direct Writing Tests'', *Modern Language Journal*, 22 (1), 69-90.

Sweedler-Brown, C. O. (1993) ''The influence of Sentence-Level And Rhetorical Features'', *Journal of Second Language Writing*, 2 (1), 3-17.

Thorndike, R. M., Cunningham, G. K., Thorndike. R. L. and Hagen, E. P. (1991). **Measurement And Evaluation in Psychology And Education**. Republic of Singapore: Macmillan Publishing Company.

Unat, Hale. (1999). **An Evaluation of Analytic Writing Criteria from the Perspectives of Native-Nonnative And Novice-Expert Teachers**. Unpublished M.A. Thesis. Middle East Technical University. Ankara.

Vann, R. J., Meyer, D. E. and Lorenz, F. O. (1984). ''Error Gravity. A Study of Faculty Opinion of ESL Errors'', *TESOL Quarterly*, 18 (4), 427-440.

White, E. M. et. al. (1996). **Assessment of Writing: Politics, Policies, Practices**. New York: The Modern Language Association of America.