Yüzüncü Yıl Üniversitesi, Eğitim Fakültesi Dergisi. Haziran 2009. Cilt:V1, Sayı:I, 85-103
http://efdergi.yyu.edu.tr

YABANCI DİL ÖĞRENCİLERİNİN YAZILI ANLATIM BECERİLERİNİN DEĞERLENDİRİLMESİ: DOĞRU DİLBİLGİSİ KULLANIMININ, DENEYİMSİZ ÖĞRETMENLERİN VERDİKLERİ NOTLARA ETKİSİ

Okt. Mehmet DURANLIOĞLU
Anadolu Üniversitesi Yabancı Diller Yüksek Okulu
mduranlioglu@anadolu.edu.tr

Yrd. Doç. Dr. Hasan ÇEKİÇ
Anadolu Üniversitesi Eğitim Fakültesi İngilizce Öğretmenliği Bölümü
hcekic@anadolu.edu.tr

ÖZET

Son zamanlarda, birçok İngilizce yabancı dil öğretimi programında, öğrencilerin yazılı anlatım kağıtları analitik yöntemle değerlendirilmektedir. Bu yöntemle, değerlendirmeyi yapan öğretmenlerin objektifliği ve güvenilirliğinin artırılması amaçlanmaktadır. Analitik değerlendirmede, öğrenci kompozisyonlarının farklı bileşenlerine (gramer/dil kullanımı, içerik, organizasyon, kelime ve noktalama işaretleri kullanımı) ayrı ayrı notlar verilmesi beklenmektedir. Bu ise, öğretmenlerin, farklı özelliklere birbirinden bağımsız olarak not vermede tutarlı olup olmadıkları sorusunu gündeme getirmektedir. Dolayısıyla, bu çalışmada, kompozisyonların bir bileşeninde (gramer/dil kullanımı) yapılan düzeltmelerin, değerlendirme işinde deneyimsiz Türk İngilizce öğretmenlerinin, kompozisyonların diğer bileşenlerine verdikleri notlara etkisi incelenmiştir. Bu amaçla, Anadolu Üniversitesi Hazırlık Okulu'nda çalışan 14 öğretmenden, bir aylık arayla 20 öğrencinin yazılı anlatım kağıtlarını iki kez değerlendirmeleri istenmiştir. İkinci değerlendirmeden önce, kağıtlardaki gramer hataları bir İngiliz öğretmen tarafından düzeltilmiştir. Birinci ve ikinci notların istatistiki karşılaştırılması, bir bileşendeki iyileştirmenin, diğer bileşenlere verilen notlarda önemli artışa yol açtığını göstermiştir. Doğal olarak, her bir bileşene verilen notlardaki artış, kağıtların bütününe verilen notları da anlamlı biçimde artırmıştır.

Anahtar Kelimeler: Yazılı anlatım, Analitik değerlendirme, Doğru dilbilgisi kullanımı.

ASSESSMENT OF EFL LEARNERS' WRITING SKILLS: THE IMPACT OF GRAMMATICAL ACCURACY ON NOVICE TEACHERS' SCORING

ABSTRACT

In many English language teaching programs, assessment of students' written products has usually been done analytically to increase raters' objectivity and reliability. In analytic assessment, raters are expected to assign separate scores for a certain number of components of students' compositions. Assessing in this way, however, raises the question of whether raters consider each component independently in assigning their scores. This study, therefore, examined the probable effects of improvement in one component (grammar/language use) upon novice Turkish English language teachers' grading the whole composition and the other four components (content, organization, vocabulary, and mechanics). For this purpose, 14 teachers, teaching at Anadolu University Prep School, were asked to grade 20 student essays twice over a one-month interval. However, before the second grading, the errors in language use were corrected by a native speaker instructor of English. The results of statistical analysis regarding the first and second gradings revealed that the teachers tend to assign considerably higher scores in the second grading for all of the other four components. This led, in turn, to significant increases in the students' overall grades.

Key Words: EFL writing, Analytic assessment, Grammatical accuracy.

INTRODUCTION

In the process of assessing students' writing ability, teachers have always had to deal with the difficult task of grading the written texts produced by students. What makes an essay test difficult or somehow problematic is not the development of the testing instrument but the quality of information it yields. How much can it be trusted?

The scoring procedure of written texts produced by students is considered by a number of teachers to be difficult and complex since it is potentially a very subjective process that may involve low scorer reliability (Baker, 1989; Gay, 1985; Harrison, 1983; Huang, 2008; Henning, 1986; Kubiszyn and Borich, 1990; Huang, 2008). However, despite the probability of such a problem, Gay (1985) reminds that ''the degree of subjectivity can considerably be minimized by careful planning and scoring'' (p. 226). Therefore, adopting an analytic assessment procedure which provides raters with detailed criteria that will lead them to focus their attention on some common standards in the marking process might be a way of decreasing this potential subjectivity.

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 sub-scores and where students are able to see what constitutes their total scores. Hughes (1989) explains these advantages in terms of both students and teachers as follows: ''First, it disposes of the problem of uneven development of sub-skills 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).

Despite all these advantages of using an analytic scoring guide, Hughes (1989) is still concerned about whether scorers can judge each of the sub-components of a scoring guide independently of the others (which is called *halo effect*). Perkins (1983) also holds a similar view 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).

Despite using the same analytic scoring criteria, different teachers' gradings for the same student's composition may result in inconsistent scores which, in turn, leads to a high subjectivity. Such a case is likely to be encountered due to various factors such as the different background of raters (e.g. teachers' experience in teaching writing), or as the varying quality of certain features of students' written texts (e.g. students' competence in grammatical structures). There are several studies that investigate whether teachers with different teaching backgrounds and/or students' competence in grammatical structures play a role in the distribution of the scores assigned to the papers.

In literature, there are a number of studies that investigated whether the scores assigned by scorers to the written texts of students are influenced by several factors. 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 grade ESL essays. The results indicated that even experienced teachers but not trained to teach ESL paid more attention to the students' grammatical accuracy. They assigned higher scores to the papers, whose grammar errors were corrected, when compared to the scores assigned to the original ones. Therefore, the researcher concludes that teachers cause 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 their papers.

Vann, Meyer and Lorenz (1984) also revealed in their study that errors at sentence-level are judged by different standards. That is, some errors were considered by some respondents to be less acceptable, while the same type of errors were 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.

Janopoulos (1992) searched for the possible tolerance of native speaker and non-native speaker writing errors. The results indicated that there generally occurred more

tolerance of non-native speaker errors than that of errors perceived to be made by native speakers.

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. The results indicated a 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. A 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.

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. No effect was observed in relation to the raters' background. However, with respect to whether training affects raters' judgements, the results indicated that training significantly influenced ratings.

It would be better here to remember that, 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 students' writing skills in the way as follows: Any student taking a writing course is expected to express himself effectively and efficiently in his written production so that he can successfully convey his message to others using the language he has 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 should pay attention to the use of transitions that help coherence and cohesion in a text).

To restate this in a single sentence, a well-written text is expected to involve a rich content with well-organized ideas, a wide range of vocabulary and grammatically correct sentences. In this sense, 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 all the above features, 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 beginner to advanced) take a writing test as a part of the end-of-year achievement test. In this writing test, students are asked to write an essay. Due to the examinees at various levels, it is always highly likely that teachers in the marking-committee will end up with papers of varying qualities in terms of these 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 another student of the same proficiency level may fail, say to organize his/her ideas even though s/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 with appropriate word-choice and/or with the accurate use of grammar. These are all just possible illustrations that teachers may encounter when assessing student papers. Thus when an analytic scoring scale is used for the assessment of written productions, as in the case of this study, one should never forget that each of these five components (content, organization, vocabulary use,

language use and mechanics) refers to a different feature of a text and should be assessed independently from one another.

As a productive skill, writing occupies an important place in all foreign language teaching syllabuses. Parallel to this significance of the place that writing skill takes in language teaching programs, reliable assessment of students' written performance is also of great importance. Then 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. In some cases, a one-point difference in students' scores may determine their success of failure in the program. This is also true for the students from all proficiency levels (from beginner to advanced) who attend the Preparatory School of Anadolu University where about 150 instructors with different teaching backgrounds teach English. Therefore, this study aimed at finding out whether grammatical accuracy (or errors) in student essays influences the scoring of teachers. For that purpose, the teachers participating in this study were asked to grade the same paper twice as follows: In the first grading, they were requested to mark the original papers and in the second grading they were asked to mark the same paper, yet the sentence-level grammar errors of which were corrected. This would help us to see if there would be 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 would occur.

Depending on the research problem, the following questions were set forth.

1. Do the total scores assigned by novice teachers in the first grading differ from those assigned in the second grading, in which they re-marked the same set of papers, yet whose sentence-level grammar errors were corrected?
2. In case of an increase in the total scores in the second grading, which of the other four components is affected by the improvement in the language use component?

METHOD

Subjects

The subjects of this study were 14 English language teachers, teaching at Anadolu University Prep School. They were all non-native speakers of English with different teaching backgrounds. The language teaching experience of all the participants varied from six months to two years, and they had taught writing for only one or two

academic years. Therefore, all the participating teachers were considered to be inexperienced writing teachers since according to Johnson et. al. (2000), a teacher who has taught writing for at least three years and above is considered to be experienced.

Data Collection

Selection of Student Essays

The materials used in this study consisted of 40 essays written by lower intermediate students under exam conditions the previous 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. 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 among the five different essay topics. All the five topics required students to write a cause-and-effect essay. Here, 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 essays written on the same topic. Therefore, by random selection, one topic was determined which 97 students chose to write essays about. Among these essays, a total of 40 essays were chosen for use in this study again on random basis.

Preparation of the essays

These 40 essays were divided into two sets. The first set of 20 essays was used to find answers to the research questions directed in this study. All the participating teachers were asked to mark this set of 20 essays twice. The second grading was held a month after the first grading, and before the second grading, the sentence-level grammar errors found in this first set of 20 essays were corrected. All the essays were typed in computer. This revealed the range of the number of words used in these essays, which was between 252 and 345 with a mean of 303.

As for the second set of 20 essays, they served as a tool to see if the teachers were consistent in their markings over time in terms of their intra-rater reliability. This was considered necessary for this study since the teachers were asked to grade the first set of essays (which were used to find answers to the research questions of the present study) twice over a certain period of time, which was a month. Therefore, the intra-rater

reliability coefficients would help see whether the teachers participating in this study were capable of re-judging papers considering similar standards across two gradings at different times.

ESL Composition Profile

The teachers received both sets of essays, 40 in total, together, and they were asked to mark them by using a well-known analytic scoring guide, the ESL Composition Profile. This profile was developed by Jacobs et al (1981), and it consists of five features such as content, organization, vocabulary, language use and mechanics, each of which has different weights of scores (for the profile, see Appendix 1). It is important to mention here that the grammar errors found in the first set of 20 essays were corrected by a native speaker of English, who had taught writing at Anadolu University for five years at the time of the study. 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 belonging to the other components (content, organization, vocabulary, spelling, mechanics) could be avoided. Here, it would be better to remember an important characteristic of sentence-level errors referring to Leki (1992, p. 105), "problems at the discourse level are often fairly subtle, leaving the reader with the feeling that something is not quite right within the text but with no clear picture of where the problem lies. At the sentence level, however, errors are relatively obvious."

Data Collection Procedures

As mentioned before, the teachers were asked to grade the two sets of essays twice at different times. For the first grading, all the teachers were asked to grade these two sets of essays (40 essays in total).

Before the second grading, as mentioned above, the sentence-level grammar errors of the first set of 20 essays were corrected. However, the second set of 20 essays remained the same for the second grading. In other words, the same second set used in the first grading was used again for the second grading without making any changes on the papers.

A further point to mention here for the second grading is that the teachers were not informed about the fact 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 first set of 20 corrected-papers. Another point important for the second grading is that all the 40 papers in both sets were put into the packs in a random order so that the sequence of essays would not have an impact on markings.

Determining the Internal Consistency of Raters' Scores across Two Gradings

In order to be able to compare the scores of the second set of 20 essays marked in the first grading with the scores of the same second set of 20 essays re-marked in the second grading, the Spearman-Brown Corelation was run. This statistical tool would provide us with correlation coefficients of the internal consistency of the scores assigned by 14 teachers. For this purpose, as mentioned above, the scores obtained in the first and second gradings were compared. These scores were assigned to the second set of 20 essays which remained the same (no change was made in papers) in the first and second gradings. The correlation coefficients were found for each of the 14 inexperienced as shown in table-1 below.

Table 1. Correlation Coefficients of Internal Consistency

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

* Coeffients found to be below 0,70

The results indicated that the scores assigned by three inexperienced teachers had correlation coefficients below 0,70, which is not desirable in terms of the consistency of the scores (Baker, 1989). Therefore, these three inexperienced teachers with low correlation coefficents were not included in the data analysis procedure described in the following part. The reliability correlation coefficients for the remaining 11 inexperienced teachers were found to be between 0,77 and 0,95 with a mean of 0,85.

Analysis of the Data

For the statistical analysis of the scores obtained through the grading of the first set of 20 essays, paired *t*-test was run. This would help see whether the teachers' scores were influenced by the correction of sentence-level grammar errors. To do this, two sets of scores assigned by 11 inexperienced teachers were taken into account: *the total*

scores of the first set of 20 essays marked in the first grading and the *expected total scores* of the corrected set of the first 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.

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

FINDINGS AND DISCUSSION

Analyses of the Total Scores of the Original Essays in the First Grading and the Expected Total Scores of the Corrected Versions in the Second Grading for Inexperienced Teachers

For the purpose of the observation of whether teachers' scores were influenced by the language use corrections, paired *t*-test was run. The reason for applying paired *t*-test was that the same teachers marked the essays twice at different times. To do this, as mentioned above, two sets of scores were compared: the total scores that the teachers assigned to the original 20 essays and the total scores that the teachers re-assigned to the same 20 essays a month later, yet language use errors of which were corrected before the second grading.

However, before the comparison of the two sets of scores obtained in the first and second gradings, it was necessary to arrange the total scores that the teachers assigned in the second grading. The reason for such an arrangement was that these total scores assigned in the second grading were thought to include extra marks that might 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 ended up with the *expected* total scores from the second grading as shown in table-2 below, which illustrates a sample calculation of an expected total score for one of the inexperienced teachers' scores.

Table 2. A Sample Calculation of an Expected Total Score

	Language Use Component (out of 25)		TOTAL (out of 100)		Calculation of an Expected Total Score in the Second Grading	
Inexperienced Teacher	First Grading	Second Grading	First Grading	Second Grading	Increase in Language Use in the Second Grading	The Expected Total Score in the Second Grading
Teacher Number:10 Paper Number:7	11	19	63	**82**	(19-11)= **8**	(82-8)= **74**

These calculated scores are called *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 or at least close to the total score assigned by him or her in the first grading. In table-2 above, sample scores of an inexperienced teacher who participated in this study is shown. The teacher assigned 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 this 8-point is subtracted from the total score of the second grading (82), a score of 74 (82-8 = 74) is obtained, 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 procedure for calculations above, the expected total scores were found for each inexperienced teacher and for each of the 20 papers for the second grading. Table-3 below 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 inexperienced teachers. It also presents the mean of the expected total scores for the second grading.

Table 3. The Mean of the Expected Total Scores Calculated for Inexperienced Teachers

Mean of the Scores Assigned to the Language Use Component (out of 25)		Mean Total Scores (out of 100)		Mean of the Increase in Language Use in the Second Grading	Mean of the Expected Total Scores in the Second Grading
First Grading	Second Grading	First Grading	Second Grading		
14,3272	19,859	69,1272	78,1863	(19,859-14,3272) = 5,5318	(78,1863-5,5318) = **72,6545**

Eventually, the total scores assigned in the first grading and the *expected total scores* obtained from the second grading were statistically compared. The results are shown in table-4 below.

Table 4. Comparison of the First And Second Gradings for Inexperienced Teachers

Inexperienced Teachers	Mean Scores		Mean Dif.	Paired t	p
	First Grading	Second Grading			
Content	22,3410	22,9863	,6453	-2,354	,029*
Organization	13,8590	15,9772	2,1182	-6,975	,000*
Vocabulary	14,7000	15,3136	,6136	-3,950	,001*
Mechanics	3,9000	4,0502	,1502	-2,223	,039*
TOTAL	69,1272**	(Expected Total) 72,6545	3,5273	-23,485	,000*

* p value is significant at ,05 level
** Language use in the first grading included

The Effect of Accurate Use of Language on the Total Scores And on the Sub-scores of Inexperienced Teachers

The scores were analyzed for inexperienced teachers. 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 inexperienced teachers to each of the four sub-components, namely content, organization, vocabulary use and mechanics were compared and analyzed.

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; the former 0,64 and the latter 0,61. Considering the paired t-test results, the increase observed in all the sub-components was statistically significant at the significance level of 0,05.

CONCLUSION AND SUGGESTIONS

The results of this study revealed two very important points in terms of the assessment of writing skills. First, all the teachers were found to increase their total scores if they assessed a paper which had an accurate use of grammar. However, what is important to state here is that, from the result just stated above, one may think that it is already a typical case for any paper to receive a higher total mark when the rater assigns a higher sub-mark to the sub-component of language use when corrected, which in turn contributes to the total mark. Instead, what we found out through the results of this study was 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 can now be observed to increase their total scores.

In terms of the increase in the total scores, it was only 3,53 for the inexperienced teachers. In writing assessment, the difference between any two teachers' scores is commonly considered tolerable if there is at most a 10-point difference in-between. However, it shouldn't be forgotten that the increase in the scores stated above is not the one observed between two teachers who scored only one paper. What is important is that the increase was observed for a total of 11 inexperienced teachers each of whom scored 20 papers in total. It should neither 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. As can be seen, the mean scores of the inexperienced 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 assigned to the sub-components of content, organization, vocabulary use and mechanics by 11 inexperienced teachers was found to be statistically significant. 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 for the descriptors for each sub-component). That is, all the inexperienced 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.

What's more, the results of this study seem to be very much in line with the results of previous research findings. In one study, 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).

Although there are possible sources for such errors and they 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, especially inexperienced ones, can attend training sessions so as to use such an analytic writing assessment profile more 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 unfortunately 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

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 or decreased by asking for the comments of 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 lead 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.

Furthermore, if it is a large-scale testing, 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 upper students or vice versa.

Suggestions for Further Studies

The research presented in this study investigated the potential impacts of students' accuracy in language use in papers on teachers' scores assigned to essays using analytic writing criteria. Consequently, 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 whether there is a positive correlation between the quality of language use and the quality of organization of ideas 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' sub-scores and thus on their total scores. Furthermore, in a different study with a similar research design, the organization of ideas in papers could be improved.

Another study could employ this time experienced teachers to investigate whether the results of the study with a similar research design will differ from those of the present study, in which only inexperienced teachers participated.

Furthermore, it could also be investigated in another study whether there will occur any difference between the scores of two groups of teachers, experienced and inexperienced, in order to see if experience in assessing written products of students plays a role in the scores assigned to student written products.

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

REFERENCES

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

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

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

Henning, G. (1986). **Twenty Common Testing Mistakes for EFL Teachers to Avoid. 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.

Huang, J. (2008). *How Accurate Are ESL Students' Holistic Writing Scores on Large-Scale Assesssment? A Generalizability Theory Approach*. **Assessing Writing**, 13, 201-218.

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

Jacobs, H. 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, R., 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.

Kroll, B. (1991). **Teaching Writing in the ESL Context: Teaching English as a Second and Foreign Language**. Boston: Heinle & Heinle Publishers.

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

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

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

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

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

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. (1993) *The influence of Sentence-Level and Rhetorical Features*. **Journal of Second Language Writing**, 2 (1), 3-17.

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

Appendix 1.

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

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

	TOTAL SCORE OUT OF 100

ESL Composition Profile developed by Jacobs et al (1981)