Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Mustafa Kemal University Journal of Social Sciences Institute
Yıl/Year: 2011 ♦ Cilt/Volume: 8 ♦ Sayı/Issue: 16, s. 471-486

RELIABILITY OF RATERS FOR WRITING ASSESSMENT: ANALYTIC - HOLISTIC, ANALYTIC - ANALYTIC, HOLISTIC – HOLISTIC

Yrd. Doç. Dr. Yakup Çetin
Fatih Üniversitesi İngilizce Öğretmenliği, e-posta: ycetin@fatih.edu.tr

Abstract

One of the main concerns in writing assessment is the choice of reliable and valid rating criteria to decide on students' writing proficiency levels. For this purpose, holistic and analytic rubrics have been employed most commonly by EFL/ESL programs and specialists as essay scoring instruments. In this present study, based on their actual use of rubric types, 31 novice Turkish teachers of English who were responsible for rating 344 student essays were randomly appointed as either holistic raters or analytic raters. Given that there were two rater pairs per essay, three different conditions and approaches to essay scoring were realized and these were used for this correlational study: holistic versus holistic, holistic versus analytic, and analytic versus analytic. Inter-rater reliability was determined through the study of these three types of scoring conditions to find the amount of correlation between raters' scores. Results determined that the highest correlation occurred most strongly between two holistic raters followed respectively by two analytic raters. The study also revealed that inter-rater reliability is rather low in a condition when two different rater types – holistic versus analytic – score the same student essay.

Key Words: *Writing evaluation, holistic rubric, analytic rubric, correlation*

KOMPOZİSYON DEĞERLENDİRMESİNDE DEĞERLENDİRİCİLERİN GÜVENİRLİĞİ: ANALİTİK - HOLİSTİK, ANALİTİK - ANALİTİK, HOLİSTİK – HOLİSTİK

Özet

Öğrencilerin kompozisyonlarını değerlendirmede temel sorunlardan bir tanesi güvenilir ve geçerli bir değerlendirme kriterine karar vermektir. Bu amaçla analitik ve holistik değerlendirme kriterleri yabancı dil programlarında uzmanlarca yaygınca kullanılmaktadır. Bu çalışmada, değerlendirme kriter ve tercihlerine göre – analitik veya holistik - 31 tane yabancı dil öğretmeni 344 öğrenci kompozisyonu değerlendirmek üzere rast gele seçilerek görevlendirildi. Her bir öğrenci kompozisyonu iki öğretmen tarafından değerlendirildiği için kullanılan farklı kriterlerden – holistik/holistik, holistik/analitik, analitik/analitik dolayı oluşan farklı değerlerin korelâsyonları incelendi. Üç farklı durum dikkate alınarak öğretmenlerin kompozisyonlara verdikleri puanların korelâsyonları hesap edilerek değerlendiriciler arasındaki güvenirlik tespit edildi. Sonuçlara göre en yüksek korelasyon aynı kompozisyonu holistik kritere göre değerlendiren öğretmenler arasında ortaya çıktı. Bu çalışmada en düşük korelasyon aynı kompozisyonu farkı kriterlere – analitik veya holistik – göre değerlendiren öğretmenler arasında saptanmıştır.

Anahtar Kelimeler: *Kompozisyon değerlendirmesi, holistik kriter, analitik kriter, korelasyon*

1. Introduction

A rubric is generally defined as a scoring tool for grading assignments. Rubrics provide a point by point guide to analysis of a given text to help raters determine an overall score for assessment. They do this by assigning a range of points for each category to be assessed within the body of the written work. Because each category should be applicable to the work, each rubric has the potential to be unique if it is designed for its specific assessment program. Correspondingly, a number of researchers have pointed out that writing assessment is more reliable and professional if a rubric is employed (Jonsson & Svingby, 2007; Silvestri & Oescher, 2006). In line with this, many EFL/ESL teachers make use of rubrics in order to be as objective as possible in the evaluation of their students' writings (Spandel, 2006). In many cases, especially due to time constraints or lack of personnel, standard rubrics taken from various ESL assessment resources are widely used for scoring students' essays.

Diverse methods for producing or obtaining rubrics for careful study of students' writing samples have been devised so far. Categories for scoring include any number of parameters such as 'smoothness of transitions from one paragraph to the next', 'sophistication of composition structure' and 'proper use of details to support topic sentences'. For instance, according to Weigle (2002), there are three types of rubrics employed in the evaluation of written product: primary trait, holistic and analytic. Primary trait rubrics are mostly utilized to decide on students' fundamental writing skills with regard to specific writing tasks (Weigle, 2002). Holistic rubrics are used to rate the properties of students' written works using a score in-line with the determined properties, but define different levels of performance superficially such as grammar, spelling, and punctuation errors (Elbow, 2000; Gunning, 2006; Weigle, 2002). Analytic rubrics are more detailed guides for assessment used to clarify the level of skill in various areas of written expression.

However, given the fact that rubric systems are different, the system used should be appropriate for the purpose of the written exam administered (Bacha, 2000). It follows that not all rubrics are equally sufficient for use and that the best system should be discovered for each assessment task. Looking at the research on both holistic and analytic approaches to writing assessment, it is easily realized that analytic rubrics are largely found to be more useful in determining students' proficiency levels; helping students to improve their quality of writing by analyzing scoring feedback and self-correcting accordingly (Read, Francis,& Robson, 2005). The analytic rubric is preferred by so many teachers and language programs, for it includes multiple scales and thus offers a collection of scores instead of only one. In this sense, for example, in a study by Beyreli and Ari (2009), an analytic rubric was employed that consisted of 'three sections and ten properties: External structure (format, spelling and punctuation), language and expression (vocabulary, sentences, paragraphs, and expression), and organization (title, introduction, story, and conclusion)'. Accordingly, in support of analytic rubric, Ross-Fisher (2005) express that a learner's knowledge and use of writing conventions such as writing style, grammar, vocabulary, spelling, and punctuation may positively or negatively influence rater's grading.

Ross-Fisher (2005) and Tomkins (2003), however, draw attention that unprofessional use of analytic rubrics may fail to provide promising results. Similarly, Rezaei and Lovorn (2010) argue the validity of analytic rubrics despite their many advantages in the evaluation of students' writings. They state that although research leans more heavily in favor of an analytic approach through the use of a rubric, this system may not be the most appropriate for all tests. The results of their study revealed that raters who were required to use a rubric for writing assessment tended to overlook the content, for they were primarily concerned with the mechanical characteristics of students' writing. The same researchers reported as well that validity or reliability of writing assessment may not improve as long as raters are not well guided in devising and using rubrics effectively. Recently, the number of pedagogues cannot be underestimated who advocate that no rubric can truly evaluate a student's writing performance from diverse perspectives including writing idiosyncrasies and creativity. Wilson (2007), as an example, regards rubrics as obstacles to good writing, for they prevent students from expressing their unique approach to concepts.

The holistic approach to assessment, on the other hand, differs from the use of analytic rubric in that the body of work is assessed as a whole and not by breaking up various parts of a work to be scored individually: the final score being a summation of the collective individual scores (Finson, 1998). Moskal's (2000) view of holistic approach is noteworthy to mention: 'When there is an overlap between the criteria set for the evaluation of the different factors, a holistic scoring rubric may be preferable to an analytic scoring rubric. In a holistic scoring rubric, the criteria are considered together on a single descriptive scale' Also, this approach has come under increasing fire by researchers, critics and policy makers, who argue that a rubric is necessary to guarantee teachers' uniformity and objective fairness in the assessment process (Hamp-Lyons 1995; Mabry, 1999; Knoch, 2009). Similarly, according to Cumming and Riazi (2000) the holistic scoring combines multiple traits of students' essays into one single score, which does not contribute much to the learning research. However, proponents of the holistic approach still see its value.

'...holistic grading can lead to improvement in writing skills in spite of the lack of diagnostics because holistic grading will change the perceptions and attitudes of both students and instructors. Students have already been taught basic writing skills prior to beginning the accounting curriculum, but they often ignore them, forget them, or devote little effort in maintaining and improving them because of lack of use. Instructors do not require written papers and do not ask essay questions because objective and reliable evaluation is difficult and time-consuming.

Even when essay tests are used, many instructors grade only for content and not for writing proficiency. Holistic grading can solve the dilemma.' (Dyer and Thorne, 1994, p.226)

Furthermore, in a study done at the University of Alaska, findings showed that the holistic approach was beneficial and reliable for large classes.

'The data show that grading efficiency...is satisfactory....our holistic grading approach appears to have reasonable reliability, validity and cost efficiency. It has functioned well over three years, serving almost 600 students...student feedback about the course has been solicited each semester, and not one complaint has been directed at the philosophy or mechanics of the grading system' (Madigan and Brosamer, 1991, p. 94).

Teachers have rather a hard time in rating student essays either holistically or analytically, for they have to cope with several factors like objective scores, appropriate writing tasks and genre, timing, and clear essay prompts before they can decide on a final score. Even several writing specialists are skeptical about the employment and reliability of rubrics as a writing grading tool (Turley & Gallagher, 2008; Wilson, 2007). In respect of this, Kroll (1990) voiced concerns about the complication in determining criteria and norms according to which to evaluate student writings. The same researcher cautions teachers to be careful in selecting a writing rubric simply because there is no best one. Likewise, though Rezaei & Lovorn (2010) find rubrics highly helpful and effective for essay scoring, they cannot refrain themselves from questioning their appropriate use. In this regard, several studies have been conducted in the literature concerning the reliability and validity of writing assessment strategies- mostly targeting holistic and analytic rubrics (Hafner and Hafner, 2003; Eliot, 2005; Meier et al., 2006). Knoch et al. (2007) also stresses that so long as raters are not well instructed on writing assessment, the prevailing belief in EFL/ESL writing research that rubrics provide promising inter-rater reliability is rather thought provoking. Even though there are a significant number of studies connected with each rubric type – in particular holistic and analytic – there is an unfortunate scarcity of experimental research that aims to compare the effectiveness of these rubrics in the same context including inter-rater reliability (Brookhart, 2005; Andrade, Du, & Wang, 2008). Therefore, this study has been undertaken in order to contribute to the limited empirical research on comparison of rubric types – analytical versus holistic- and to discover whether they increase or decrease the reliability of teacher's evaluation of student essays. This present correlational study specifically aims to explore inter-rater reliability based on essay scoring by inexperienced teachers of English in relation to three different rubric conditions -analytic versus analytic; analytic versus holistic; holistic versus holistic. The actual purpose then is to see if there is correlation among novice raters who use either the same or different rubric types in evaluating students' final essays. In other words, answers were sought to the following questions:

1. What is the rate of correlation between two groups of raters who both use holistic rubrics to score the same student essays?
2. What is the rate of correlation between two groups of raters who both use different rubrics - one analytic and the other holistic - to score the same student essays?
3. What is the rate of correlation between two groups of raters who both use analytic rubrics to score the same student essays?

2. Method

2.1. Subjects

A total of 344 students with their essays, all of whom were attending a University Preparatory School in Istanbul at Fatih University for almost one year, participated in this study. 162 of them were males and 182 were females whose ages ranged from 18 to 21 (m=19). At the time of the experiment the participants had already completed the four Module marathon (A1= Beginner; A2= Elementary; B1= Pre-Intermediate; B2= Intermediate) as a requirement of the one year preparatory program. At the time of the study, the students were waiting for the upcoming Proficiency Exam so that the successful ones, who passed the test with 70 points, could pursue their academic study in the respective faculties whose medium of language is English. As a preparation for the forthcoming proficiency exam, they were willing to participate in the present study with their essays.

Besides the students, 31 teachers of English (8 males, 23 females) participated as raters in the evaluation of 344 student essays. The raters whose ages ranged from 24 to 26 (m= 25) were mostly novice teachers with little or no EFL teaching experience except for one or two participants who had taught English for a couple of years. The rater participants were all graduates from different well-known English Language Teaching (ELT) departments in Turkey; however, they had neither rated this particular type of test before nor received any essay rating training so far.

2.2. Instruments

The rubric used in this study was analytic in nature. It was very similar to the writing evaluation rubrics being used in the English preparatory school at Fatih University (where this study was conducted). It has been designed by the University of Texas at Austin to conform to the 24 point scale, and includes the following criteria:

(1) Introduction, (2) articulation of thesis, (3) paragraph development, (4) use of examples, (5) conclusion, (6) transitions, (7) variation of sentences, and (8) coherence. All eight components on the scale were further divided into 3 subdivisions in relation to described criteria. For example, in the introduction part

students received 1 point for ‘no or poor introduction’, 2 points for ‘some introduction; nothing beyond a forecast’, and 3 points for ‘introduction grasps reader’s attention (engages the reader) and forecasts major points’. The 18 analytical raters (4 males, 14 females) were expected to give a score between 1(lowest) to 3 (highest) points to every component (8x3 points) out of the 24 point scale. On the other hand, raters relying on the holistic approach 13 raters (4 males, 9 females) in this study were not given a written rubric, relying instead upon a mental rubric to assess the work. Therefore, holistic is defined within our study as ‘assessing the whole body of a written text without the use of a written rubric.’ Accordingly, holistic raters were asked to evaluate student essays for this research in terms of the overall quality of the paper and their personal impression by assigning a single score. Since two groups of raters were randomly appointed to score the same student essays, a scoring condition emerged as a matter of course in which either both groups of raters were holistic (4 raters + 4 raters), or analytic (9 raters + 9 raters), or one holistic (8 raters) and the other analytic (8 raters).

The Spearman Correlation Coefficient was preferred to learn about the strength of relationship between the two raters’ scores (inter-rater reliability) who were either analytical or holistic. Moreover, the Spearman rho was used for the nonparametric statistical calculations instead the Pearson, for the analytical and holistic scores of raters were ordinal and not normally distributed according to Kolmogorov-Smirnov Test (Table 1).

Table 1: *Tests of Normality*

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Score holistic	,108	142	,000	,976	142	,013
Score analytic	,089	142	,008	,980	142	,034

a. Lilliefors Significance Correction

2.3. The procedure

As a part of the experiment, 344 students in 20 classes were told to choose one of the four topics calling for different rhetorical styles (argumentative, cause-effect, compare & contrast, and problem solving) and write a well-organized essay within a time period of 50 minutes. Even though it is likely that both the rhetorical style and topic may impact writing performance (Sweedler-Brown, 1992; White, 1994), this risk was thought to be minimum due to students’ experience and practice with various essay types throughout the modules (A1, A2, B1, B2) in the preparatory program; the popularity and variety of the topics for optimal performance; and the assignment of two groups of raters to evaluate the same student essays. In accordance with confidentiality requirements to overcome subjectivity, the essays were coded and distributed to the 31 raters for the first evaluation by the Preparatory School Coordinators who assisted in the implementation of the research. Upon the completion of the first evaluation and following the rearrangement of student essays, the coordinators appointed the same raters for a second-evaluation in order to be as objective as possible with the essay scores. Namely, every student essay was scored by two different raters who both could be either both holistic-holistic or both analytic-analytic or one holistic and the other analytic. The two rater scores on the same student essays whose average determines the final writing score were obtained from the writing coordinators for statistical calculations.

3. Results and Findings

3.1. Condition 1: Holistic versus Holistic

Table 2 and Table 3 provide the necessary statistical calculations regarding the condition in which all raters were asked to adopt a holistic standpoint as to the scoring of student essays. That is to say, the eight raters were divided into two groups of four and were asked respectively to evaluate the same student essays based on their personal impression without a written rubric. In line with the study regulations in order to meet the reliability and validity criteria, the same 64 essays were scored first by four raters and then again by four other raters: in total by 8 holistic raters. Table 2 indicates the descriptive statistics that include the means for both conditions (M=16, 54 and M=15, 87 respectively) standard deviations (SD=3, 84 and SD =3, 79 respectively), and the number of participants (N=64). Table 2, furthermore, provides the correlation coefficient of scores by all holistic raters in two rating conditions: holistic versus holistic. As indicated in Table 2, despite the fact that the score means (M= 16,54 and M= 15,87 respectively) are relatively close to each other, the statistical calculations in Table 3 show a high correlation coefficient (r=.775) between the essay scores given by two groups of holistic raters whose *p* value is also statistically significant (p=.000). In other words, the Spearman Correlation Coefficient Statistical Test reveals a highly positive correlation (r=.775) and inter-rater reliability between first scoring and second scoring of essays by two groups of raters who adopted a holistic style towards writing evaluation. These results indeed report that holistic raters were in relative agreement with their scoring of the same student essays on two occasions: holistic versus holistic.

Table 2: *Descriptive Statistics of two raters: Holistic vs Holistic*

	Mean	Std. Deviation	N
Holistic	16,5469	3,84183	64
Holistic	15,8750	3,79013	64

Table 3: *Correlation coefficients between two raters: Holistic vs. Holistic*

			Grade1	Grade 2
Spearman's rho	Grade1- Holistic	Correlation Coefficient	1,000	,775(**)
		Sig. (2-tailed)	.	,000
		N	64	64
	Grade2- Holistic	Correlation Coefficient	,775(**)	1,000
		Sig. (2-tailed)	,000	.
		N	64	64

** Correlation is significant at the 0.01 level (2-tailed).

3.2. Condition 2: Holistic versus Analytic

The following tables and statistical data were obtained from the scoring condition where two groups of raters were asked to adopt different standpoints – *holistic versus analytic* – in scoring the same student essays. In the *holistic versus analytic* condition, 142 essays were first scored by 9 holistic raters and subsequently by 9 more analytic raters for a second check for objectivity concerns. The first nine holistic raters were asked to evaluate the student essays without the use of a written rubric out of a maximum score of 24. On the other hand, for the second check of essays, the nine analytic raters were required to use a written rubric containing 8 components with 3 properties as a criterion whose sum of individual scores equals also 24 points. Table 4 shows the descriptive statistics of the two groups of raters and their scores (M=13, 99 and M=3, 16 respectively; SD =3, 97 and SD=3, 16 respectively). Relatedly, the correlational analysis in Table 5 indicates a statistically significant correlation between both conditions because the *p* value is apparently less than .05. The presence of a correlation coefficient of r=.602 draws attention to a highly moderate correlation, for a correlation between 0.70-0.30 is considered as a medium correlation in social sciences (Büyüköztürk, 2009). Correspondingly, both the statistically significant *p* value (p=.000) and the correlation coefficient of r=.602 are reliable values in foreign language research, especially in evaluating qualitative data like student essays.

Table 4: *Descriptive Statistics of two raters: Holistic vs Analytic*

	Mean	Std. Deviation	N
Holistic	13,9930	3,97063	142
Analytic	13,2394	3,16436	142

Table 5: *Correlation coefficients between two raters: Holistic vs Analytic*

			Grade1	Grade 2
Spearman's rho	Grade 1 - Holistic	Correlation Coefficient	1,000	,602(**)
		Sig. (2-tailed)	.	,000
		N	142	142
	Grade 2 - Analytic	Correlation Coefficient	,602(**)	1,000
		Sig. (2-tailed)	,000	.
		N	142	142

** Correlation is significant at the 0.01 level (2-tailed).

3.3. Condition 3: Analytic versus Analytic

When the 8 raters who scored the first check and the other 8 raters who scored the second check (in total 16 raters) were all identified as analytic raters, who used the same written rubric, the below descriptive statistics and correlation coefficients were acquired from the 138 essay scores. Table 6 shows the score means (M= 12, 60 and M=13, 04 respectively) and the standard deviations (SD= 3, 39 and SD=3, 41 respectively) by two rating conditions: analytic versus analytic. In Table 7 the Spearman correlation coefficient displays a medium correlation of r=.683 with a statistically significant *p* value (p=.000). When we regard a correlation between 0.70 and 1.00 as a high correlation in social sciences (Büyüköztürk, 2009), it can be observed that the obtained correlation (r=.683) in Table 7 is very close to the high correlation range.

Table 6: *Descriptive Statistics of two raters: Analytic vs Analytic*

	Mean	Std. Deviation	N
Analytic	12,6087	3,39264	138
Analytic	13,0435	3,41715	138

Table 7: *Correlation coefficients between two raters: Holistic vs Analytic*

			Grade1	Grade 2
Spearman's rho	Grade 1 -Analytic	Correlation Coefficient	1,000	,683(**)
		Sig. (2-tailed)	.	,000
		N	138	138
	Grade 2 - Analytic	Correlation Coefficient	,683(**)	1,000
		Sig. (2-tailed)	,000	.
		N	138	138

** Correlation is significant at the 0.01 level (2-tailed).

4. Summary and Discussion

The appropriate scoring of student essays in order to determine their proficiency level has been the primary interest of many ESL/EFL specialists for many years (Kroll, 1998). Within this framework, the objectivity of rubrics – mostly their reliability and validity - has been investigated from different aspects (Breland, 1983; Wolfe, 1997; Mabry, 1999; Spandel, 2006; Kohn, 2006). Yet, the belief that rubrics do not guarantee objective and reliable scoring because of their reductive nature is prevalent among a considerable number of language specialists (Pula & Huot, 1993; Kohn, 2006). Therefore, this study whose aim was to examine different conditions of inter-rater reliability - analytic vs. analytic; analytic vs. holistic; holistic vs. holistic - was conducted in order to contribute to the sparse and inconsistent body of research in the field.

Even though the findings of this study were unforeseen and also require further inquiry, the researchers found them valuable to report because of the meaningful sample size and the statistically significant results. In the present study the correlation coefficients between essay scores from 31 novice raters in three conditions - (a) analytic vs. analytic (b) analytic vs. holistic (c) holistic vs. holistic – were statistically calculated to find out about inter-rater reliability in terms of a certain rubric and personal impression of student essays in relation to their overall quality. Correspondingly, the three scoring conditions - (a) analytic vs. analytic (b) analytic vs. holistic (c) holistic vs. holistic – revealed statistically significant correlations and noteworthy inter-rater reliability. These findings are discussed next in more detail according to the research questions.

Research Question 1: What is the rate of correlation between two groups of raters who both use holistic rubric to score the same student essays?

The data in the present study indicates that the correlation coefficient (r=.775) is highest between two pairs of novice raters when they use a holistic rubric; that is to say, a highly positive inter-rater reliability is provided when two groups of raters score the same essays in terms of both the overall quality and their personal impression. Specifically, in a condition in which 64 essays were scored first by four holistic raters and subsequently again by four more holistic raters for reliability and validity purposes in accordance with exam regulations; a high correlation of .775 emerged between these two rater pairs. Since a correlation value between .70 to .89 has a high level of statistical significance according to Cohen and Holliday (1982), the holistic versus holistic essay rating condition is strongly correlated. Relatedly, out of the three conditions - *(a) analytic vs. analytic (b) analytic vs. holistic (c) holistic vs. holistic* – according to the statistical results of this study, the *holistic versus holistic* condition produced surprisingly the highest level of correlation and accordingly inter-rater reliability.

This finding was unexpected, for the researchers, in line with the literature in the field (Mabry, 1999; Knoch, 2009), assumed that the *analytic versus analytic* rating condition - which included raters who favored detailed and comprehensive rubrics for essay scoring - would lead to more inter-rater reliability with grading by decreasing the range and variability among scores. Therefore, the results of this study contradict the common belief that the employment of analytic rubrics in essay scoring highly increases the reliability or validity of writing evaluation. Moreover, the statistical results of the holistic versus holistic condition are in support of Rezaei & Lovorn's (2010) study in which raters were more or less predisposed to score essays according to the overall quality as well as personal impressions instead of detailed rubrics. The implication of their study is that the quality of an essay should be rated as a whole since the sum of evaluation of individual components does not lead to a more reliable and objective scoring. Similarly, the findings of this study also strengthen the prevalent use of holistic scoring in large-scale testing contexts, particularly with novice raters, because of its practicality, affordability, and simplicity. As a matter of fact, further training of holistic raters on essay scoring could yield better inter-rater reliability correlations in various ESL/EFL test that include writing component.

Research Question 2: What is the rate of correlation between two groups of raters who both use different rubrics - one analytic and the other holistic - to score the same student essays?

The findings of this present study also reveal that the correlation (r=.602) is lowest between the scores to the same essays when there is a difference in terms of use of rubric between novice raters: *analytic versus holistic.* Even though the obtained correlation coefficient of r=.602 is lowest in comparison to the other conditions - *analytic versus analytic and holistic versus holistic* - it is found to be still statistically moderate and significant (Cohen and Holliday, 1982). In a condition where the same 142 student essays were first scored by nine holistic-referenced raters and afterwards by nine more analytic-referenced raters in accordance with the exam regulations for reliability and validity issues, the inter-rater reliability was lowest because of the variability of the scores. Namely, the data suggests that as far as essay scoring is concerned, if more than one rater (two novice raters for this condition) are assigned to the assessment of the same essays, then raters' random use of rubrics will most possibly result in rather low inter-rater reliability.

The findings of the second condition more or less confirms McNamara (1996), Weigle (2002), and Lumley (2005), who point out that inconsistency between essay scores and raters are experienced in different forms and scales depending on the degree of raters' compliance to rubrics. Relatedly, one reasonable explanation for lower inter-rater reliability and correlation can be that holistic raters concentrated mostly on the overall quality of the student essays because of the product-oriented characteristic of holistic rubric. On the other hand, analytic raters are thought of as process-oriented; therefore, they were more interested in the evaluation of the details of an essay such as creativity, structural organization, writing mechanics, etc whose sum of scores constitute the final score. Therefore, rater's difference in preference and perception of rubric - *analytic versus holistic* – in scoring essays seems to be the most plausible explanation for the lowest correlation coefficient for the second condition, though of medium significance value. Similarly, Cohen and Manion (1994) reported as well that variation of essay scores is unavoidable as long as raters employ different rubric types. In the same context, as far as essay scoring is concerned, they suggest that raters should share the same rating scale based on a mutual understanding and interpretation.

Research Question 3: What is the rate of correlation between two groups of raters who both use analytic rubric to score the same student essays?

Condition three- *analytic vs. analytic* - seems to be complementary to the perspective that if raters use a common rubric for essay scoring, their evaluation results are expected to be in close agreement with another. A correlation coefficient of r=.683, (very close to .70) from *analytic versus analytic* condition typically draws attention to significant inter-rater reliability, though not as high as *holistic versus holistic* condition. In this context, the 138 student essays for this research were scored by two groups of novice analytic raters (each consisting of 8 raters) who were given a detailed written rubric. Specifically, the 138 essays were first divided between 8 analytic raters so that each rater was assigned to score approximately 17 essays in accordance with a certain written rubric. Upon the completion of the first scoring, the same 138 essays were again divided between 8 other different analytic raters for a second scoring to help ensure reliability and validity standards. The statistical calculations of the final average scores provided by the two groups of analytical raters, contrary to the expectations, did not produce as high of a correlation coefficient as the *holistic versus holistic* condition. Even though analytic rubrics, because of their criterion-referenced feature, are considered to be superior, particularly for detailed scoring of an essay (Bacha, 2001), the results of this present study nevertheless point more to a moderate inter-rater reliability between analytic raters. Correspondingly, the results of this study apparently contradict East's (2009) argument that an analytic rubric, due to its variable and varying features, is more reliable and objective than a holistic rubric. It appears that, because of the lower inter-rater reliability, the noteworthy aspects of analytic rubrics, like offering comprehensive information about the test taker's performance, has been shadowed by holistic rubric. As emphasized by Kroll (1990), 'There is no single written standard that can be said to represent the 'ideal' written product in English. Therefore, we cannot easily establish procedures for evaluating ESL writing in terms of adherence to some model of native-speaker writing. Even narrowing the discussion to a focus on academic writing is fraught with complexity. (p. 141)' this medium correlation obtained in the third condition reminds us of the complexity of creating criteria and standards in order to assess student essay.

Before closing, it is necessary to mention that the study was done with limited number of raters (n=31) who were untrained in essay rating and student essays (n=344); thus, in order to generalize its findings, further studies that include more raters and essays are absolutely necessary. Specifically, the number of raters for each condition, for instance holistic, was solely limited to 4 male and 9 female teachers. It can be assumed that a study that includes a greater number of experienced raters for each condition – analytic vs. analytic, holistic vs. holistic, holistic vs. analytic – might either strengthen or contradict the findings of this study. Additionally, the majority of raters (7 males, 24 females) in this study were females; it is conceivable that other studies that focus only either on male or female raters to determine the inter-rater reliability or intra-rater reliability for diverse scoring conditions might bring about notably different results.

Lastly, it seems to be the fact that many ESL / EFL teachers who are appointed as raters to score student essays according to an analytic rubric may not actually do so; instead, as in this case, they may give a single score based on their personal impression and overall quality of the essay regardless of the instructions given to them. For this reason, complementary studies are called for which inquire about raters' knowledge and compliance to a given rubric in the evaluation of student writings. For example, a study that quantitatively and qualitatively investigates analytic raters' level of attachment to the rubric during essay scoring might come up with surprising results for writing research.

Because of differences in rubric use by novice raters, a rating condition may emerge that may seriously question inter-rater reliability and in turn essay scores (Lumley, 2005). In connection with this, the purpose of the present study was to investigate inter-rater reliability among novice raters in three different essay scoring conditions: (a) analytic vs. analytic (b) analytic vs. holistic (c) holistic vs. holistic. The correlational findings of the study indicated that inter-rater reliability is highest when novice raters use holistic rubric to evaluate the same student essays. Surprisingly, contrary to expectations, the condition in which all raters used an analytic rubric to evaluate the same student essays did not result in a high inter-rater reliability, though the correlation is of medium significance. The same study also showed that despite the moderate correlational coefficient, inter-rater reliability is lowest when different group of novice raters who evaluate the same student essays employ different rubrics such as holistic versus analytic. Thus, the study implies that in essay scoring higher inter-rater reliability is obtained when novice raters are required or tend to use to the same rubric whether this be holistic versus holistic or analytic versus analytic. Furthermore, the low inter-rater reliability which resulted from different rubric use by raters – holistic vs. analytic- for the evaluation of the same student essays carries many implications for ESL/EFL programs and exams.

References

Andrade, H., Du, Y., &Wang, X. (2008). “Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students’ writing.” *Educational Measurement: Issues and Practice,* 27 (2), 3–13.

Bacha, N. (2009). “Writing evaluation: what can analytic versus holistic essay scoring tell us?” *System,* 29, 371–383.

Bacha, Nahla. (2001). “Writing evaluation: what can analytic versus holistic scoring tell us?” *System*, 29, 3, 371-383.

Beyreli, L. & Ari, G. (2009). “The Use of Analytic Rubric in the Assessment of Writing Performance-Inter-Rater Concordance Study- Kuram ve Uygulamada Eğitim Bilimleri “, *Educational Sciences: Theory & Practice*, 9 (1), 105-125.

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review*. New York: College Entrance Examination Board.

Brookhart, S.M. (1999). “The Art and Science of Classroom Assessment: The Missing Part of Pedagogy.” *ASHE-ERIC Higher Education Report* (Vol. 27, No. 1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Büyüköztürk, Ş. (2009). *Sosyal Bilimler İçin Veri Analizi El Kitabı: İstatistik, Araştırma Deseni SPSS Uygulamaları Ve Yorum* (10.Baskı). Pegem Akademi.

Cohen, L. and Holliday, M. (1982) *Statistics for Social Scientists*, London: Harper & Row.

Cohen, L., and Manion, L. (1994). *Research methods in education* (4th ed.). Newyork: Routledge.

Cumming, A., & Riazi, A. M. (2000).”Building Models of Adult Second Language Writing Instruction”. *Learning and Instruction* 10, 55-71.

Dyer, Jack L., & Thorne, Daniel. (1994). “Holistic scoring for measuring and promoting improvement in writing skills”. *Journal of Education for Business*, 69/4, 226-231.

East, M. (2009). “Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing”. *Assessing Writing*, 14, 88–115

Elbow, P. (1999). Ranking, evaluating, and liking: sorting out three forms of judgments. In: Straub, R. (Ed.), *A Sourcebook for Responding to Student Writing*. Hampton Press, Inc, New Jersey, pp. 175–196.

Finson, K. D. (1998). "Rubrics and their use in inclusive science". *Intervention in School and Clinic*, 34 (2), 79–88.

Gunning, T. G. (2006). *Assessing and correcting reading and writing difficulties* (3th ed.). Boston: Pearson Education Inc.

Hafner, J. C., & Hafner, P. M. (2003). "Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating". International Journal of Science Education, 25, 12, 1509–1528.

Hamp-Lyons, L. (1995). "Rating non-native writing: the trouble with holistic scoring". *TESOL Quarterly* 29, 4:759-62.

Jonsson, A., & Svingby, G. (2007). "The use of scoring rubrics: Reliability, validity and educational consequences". *Educational Research Review*, 2, 130–144.

Knoch, U. (2009). "Diagnostic assessment of writing: A comparison of two rating scales". *Language Testing*, *26,* 20, 275–304.

Knoch, U., Read, J., & von Randow, J. (2007). "Re-training writing raters online: How does it compare with face-to-face training?" Assessing Writing, 12, 26–43.

Kohn, A. (2006). "The trouble with rubrics". *English Journal*, *95* (4), 12–15.

Kroll, B. (ed.) (1990). *Second Language Writing: Research Insights for the Classroom*. Cambridge University Press, Cambridge.

Kroll, B. (1998). "Assessing writing abilities". *Annual Review of Applied Linguistics* 18:219-40.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Lang.

Mabry, L. (1999). "Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment". *Phi Delta Kappan*, *80* (9), 673–679.

Madigan, R., & Brosamer, J. (1991). "Holistic Grading of Written Work in Introductory Psychology: Reliability, Validity, and Efficiency". *Teaching of Psychology*, 18/2, 91-94.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Moskal, B. (2000). Assessment Resource Page. http://www.mines.edu/Academic/assess/Resource.htm (November 20, 2010).

Pula, J. J., & Huot, B. A. (1993). "A model of background influences on holistic raters. In: M. M. Williamson & B. A. Huot (Eds.)", *Validating holistic scoring for* *writing assessment: Theoretical and empirical foundations* (pp. 237–265). Cresskill, NJ: Hampton Press.

Read, B., Francis, B., & Robson, J. (2005). “Gender, bias, assessment and feedback: Analyzing the written assessment of undergraduate history essays.” *Assessment and Evaluation in Higher Education*, 30 (3), 241–260.

Rezaei, A. R. and Lovorn, M. (2010). “Reliability and validity of rubrics for assessment through writing”. *Assessing Writing, 15,18–39.*

Ross-Fisher, R. L. (2005). “Developing effective success rubrics*”. Kappa Delta Pi*, 41 (3), 131–135.

Silvestri, L., & Oescher, J. (2006). “Using rubrics to increase the reliability of assessment in health classes”. *International Electronic Journal of Health Education*, 9, 25–30.

Spandel, V. (2006). “In defense of rubrics”. *English Journal*, *96* (1), 19–22.

Sweedler-Brown, C.O., 1992. “The effect of training on the appearance bias of holistic essay graders”. *Journal of Research and Development in Education* 26 (1), 24–29.

Turley, E. D. & Gallagher, C. G. (2008). “On the uses of rubrics: Reframing the great rubric debate”. *English Journal*, 79, (4), 87–92.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

White, E. M. (1994). *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance*. Jossey-Bass Publishers, San Francisco.

Wilson, M. (2007). “Why I won’t be using rubrics to respond to students’ writing”. *English Journal*, 96 (4), 62–66.

Wolfe, E. W. (1997). “The relationship between essay reading style and scoring proficiency in a psychometric scoring system”. *Assessing Writing*, *4* (1), 83–106.