International Journal of Languages' Education and Teaching
Volume 6, Issue 3, September 2018, p. 331-346

Received	Reviewed	Published	Doi Number
17.08.2018	01.09.2018	30.09.2018	10.18298/ijlet.3138

Is Cross-Marking A Way To Increase Rater Reliability?

Murat POLAT [1]

ABSTRACT

Most of the error correction research has focused on whether teachers should correct errors in student writing, how they should do it and how deep it should be. Recent research, thus, has mostly focused on pedagogical merits of error correction and its possible benefits for student learning. However, in some particular contexts where graders make multiple scorings on the same paper, not much has been investigated to see if those corrections manipulate other graders or whether the writing teachers' corrections on students' papers have a positive or negative impact on the reliability of the scores when raters see the corrections of the other graders on the papers they mark. This study intended to explore whether corrections made by the graders affect the scores of colleagues who are scoring the same papers second time to gain more accurate results and to ensure the rating reliability. To do that, 12 writing teachers graded 20 student essays written by intermediate level English learners. The participants were first asked to grade 10 papers without doing error correction and those papers were re-scored after 3 weeks by the same graders, inter-rater and intra-rater reliability computations were carried out for this set of papers to see the actual reliability levels of the raters under normal circumstances. In the second stage, the graders were asked to score the other 10 papers, but this time they also made error corrections on the papers and after 3 weeks, the same teachers graded the same papers that were corrected by their pair graders. The scores assigned each time to these papers by the same raters, were compared statistically and the effect of error correction was investigated on their scores. In conclusion, the results revealed that error marking and grader comment on writing papers may have a negative effect on raters' intra-rater reliability levels whereas it could have a positive effect on raters' inter-rater reliability levels when a pool of raters grade the same papers.

Key Words: Cross-marking, inter-rater reliability, intra-rater reliability, grading writing, language testing.

1. Introduction

The achievement scores assigned to students' works matter significantly, that's why any educational institution no matter if it is a secondary school or a college, a private or a state school must in a way assure its students and all the involved parties (parents, employers, decision makers etc.) that not only its exams but also results of these exams are accurate, valid and reliable since these scores undoubtedly affect their students' chances of gaining scholarships, further education, educational awards and finally employment (Rust, 2007). Considering the fact that most of the decisions taken on

[1] Dr., Anadolu University, mpolat@anadolu.edu.tr.

behalf of the learners are based on the grades they take, assuming those grades reflect their academic performance fairly objectively is indispensable. As Rom (2011) states in his study, even a 4-point-difference out of a hundred-point-scale could make a caretaking difference on a student's final score and his/her success ranking compared to a friend who had relatively similar grades. Many researchers (Bloxham, 2009; Ecclestone, 2001; Fleming, 1999; Guanxin, 2007; Yorke, 2008) thus believe that obtaining accurate and reliable results from assessment is somehow problematic and requires a lot of effort. Most of the time, academicians feel that whenever they grade a paper, they assign the true grades and those graders feel that they are mostly reliable within themselves. This is considered true most of the time since there is only one assessor to shed the lights on but what if there were more assessors to assess the same student's performance. Which quality in such settings should be prioritised; the genuine scoring done by the grader him/herself or the reliable scoring done by a pair of graders that might be influenced by one of the raters in the jury?

2. Ways to increase rater reliability

To evaluate learners' academic writing skills, using the essay writing has been a common application in most academic settings and is highly valued in undergraduate education (Price et al., 2011; Robson et al., 2002). However, many studies (Bloxham, 2009; Johnson et al., 2009; Laming, 1990; Oruc, 2015; Shaw, 2008; Yorke, 2008) revealed that assessment of student writing is harshly criticised in terms of reliability and accurate marking. Needless to say, even if subjective evaluations are done, a considerable degree of reliability and accuracy should be guaranteed since those qualities are highly critical if standard writing tests are used to determine learners' success in educational settings. To have reliable and accurate scoring Brown, Glaswell and Harland (2004) propose a number of ways including the use of valid and explicit marking guides (either a holistic or analytic, depending on the needs of the institution), systematic evaluation of students' writing, cross-marking students' essays, training the graders regularly and finally using expert and trained markers in assessment.

Stemler (2004) highlights three major approaches in determining the reliability and accuracy of grading students' writing. Of these three, the first one consensus estimates shows us the exact level which graders assign similar scores to the same piece of work, the second consistency estimates reveals the degree to which graders agree on the band of high or low scores given to successful or unsuccessful pieces of writings and the last one measurement estimates shows us the degree to which grades could be attributed to common scoring rather than to error components. In another study, Brown (2009) defends the idea that gaining a high level of inter-rater reliability is crucial and requires al these three approaches. Since the nature of grading a writing paper involves rater subjectivity, the scores that a pool of raters assign to a student work should be similar and crosschecking; therefore, grading a paper by two or more raters could increase the scoring accuracy and reliability positively (Shavelson & Webb, 1991). On the other hand O'Hagan and Wigglesworth (2015) state that there are a number of factors which could lead to inconsistencies in the grades given to the same written work even if it was double-marked. As to the reasons for these inconsistencies, the marking procedures, expertise of the graders and the moderation of training for marking could be listed.

Researchers including Bell (1980), Wood and Quinn (2006) claim that at least two graders should score an essay for a better and reliable scoring, others including Caryl (1999) and Kuper (2006) believe that multiple markers (in some cases where there is score disagreement three or more markers) could be used for essay scoring. Though involving more graders may reduce the random error in judging the

quality of the written work, it may not mean that more accurate or reliable scores will be taken. Read et al. (2005) stated that two different essays received six different grade classifications (from fail to successful) when graded by 50 trained assessors. Therefore, what might be more important to have more accurate results may not be involving more graders but determining clear instructions on how the assessment would be done (on what principles and in which methods).

Brown (2009) asserts that reliability and legitimacy of essay marking is doubtful since there are various sources of error in essay scores. These could stem from the writing task, grading procedure or graders' personal attitudes on scoring. To lessen those errors, most of the time, graders' formal meetings so called as *"norming or standardisation sessions"* are organized for the moderation of the participants. In some cases, graders are asked to mark errors on essays and they assign their scores based on those errors in terms of content, organisation, grammar, vocabulary and the mechanics of the written work. This kind of error marking or commenting on the essay may help the graders to see the picture in detail for a better personal evaluation and may be a useful tool to justify his/her scoring and even in some cases, this kind of error marking could help and ease the work of the other grader(s) where multiple scorings are done on the same paper. However, it should also be taken into account that such error marks or corrections may affect the grade of the other graders and may manipulate them while assessing a student's performance. Unfortunately, there is almost no research examining the effect of grader comments or corrections on the other assessors which they see on writing papers.

Cross-marking is offered as a useful tool to increase rater reliability and grading accuracy in writing assessment, however, the principles of how this marking should be carried out or if error corrections and/or grader comments should be made are still unclear. On one hand, students want to see their errors on their papers after the exams and error correction and rater comments may help them in this sense, on the other hand the role of these corrections and comments on the decisions of the other grader(s) is not clear and worth studying since there is considerable evidence of inconsistency among graders scores. Clearly, the question of the role of cross-marking on the reliability of scores and how it should be done (with or without error correction and grader comments) has not been clearly and fully resolved since related literature reflects no significant evidence. Thus, in this study it was aimed to identify the role of cross-marking in the effort to increase scoring reliability in writing assessment and drive some implications on how error-correction should me made. The study reported here will also address the question if cross-marking harms true grading and will check if it really increases inter-rater reliability and finally will present evidence on how it affects intra-rater reliability levels on grading writing.

3. Method

The aim of this study is to investigate the role of error correction and grader comments on exam papers on graders' judgements concerning the reliability of essay scores; thus, an exploratory design was implemented. The following are the research questions of the study:

1. Do error correction and grader comments affect the other graders' judgements significantly while doing cross-marking?
2. Do error correction and grader comments affect intra-rater reliability?
3. Do error correction and grader comments affect inter-rater reliability?

3.1. Participants

20 writing graders who had a working experience more than 7 years and working at a state university in Turkey were randomly selected (their gender, prior education on language teaching, master professions were not taken into acoount) and invited to contribute to this research via email. Of these 20 teachers, 12 (8 female, 4 male graders) participants accepted to take part in the study voluntarily. Total year of experience of these teachers in grading writing ranged between 8-23 years (6 teachers were experienced between 8-15 years and 6 teachers were experienced more than 15 years up to 23 years), and all of them had a previous experience not only in scoring writing but also in scoring oral performances as they are mostly assigned to grade midterms or proficiency exams of their institution. Those teachers were paired as one grader with an experience more than 15 years and the other grader who is less experienced than 15 years. They never came together nor shared opinions or commented on their scores while grading. All the scores they gave throughout the study were kept confidential and they were especially warned not to share their comments or scores they gave with their partners.

3.2. Instruments

20 essays written on the topic "What has been the most influential invention for humankind?" were used in the study for the graders to grade. They were all written by intermediate level English language learners whose ages ranged from 17-21. Only intermediate level papers were included to lessen the number of variables like language level, the amount of correction and feedback etc. It was an opinion essay and students wrote about 250 words in each essay. The ESL Composition Profile was used as the scoring rubric and a simple correction code was adopted by the graders like "gr" for grammatical errors, "voc" on the choice of vocabulary, "org" for organisational problems or "mech" for the errors related to capitalisation, punctuation, handwriting etc. All the graders were given specific grader codes and related documents so as to file their grades after each grading session. To ensure that graders do not score the papers within the same line, in each scoring session the codes of the papers were changed (the first paper in the first grading was the fourth in the second grading, was the sixth in the third session etc.)

3.3. Procedure

Four grading sessions were held during the study to collect data. 20 essays (these were the exam papers which were scored beforehand by experienced graders of the same institution and they were intermediate level students' papers) were divided into two sets (Set A- Set B) including 10 papers in each. The original scores of the papers were ranging from 60-80 out of 100 points and they were randomly divided in to two separate sets. Two different sets were designed in order to avoid participants remembering the content or the score of the papers they grade. In the first grading session, the raters were given the first set of papers (Set A) and they were invited to grade those papers without writing any comments or making error correction on the papers, they used ESL Profile as the grading rubric and at the end of the grading session the scores they assigned on Set A papers were collected and computerized. After a 3 weeks, the raters were given Set A papers again and they re-scored the papers again without writing any comments or error corrections. After a week Set B papers were given to the graders and they were invited to grade them after they make error correction and if they feel necessary they were free to write their comments on organization, idea development, justification of their ideas, exemplification, vocabulary variety, mechanical quality etc. After 3 weeks for the final session, the papers which were corrected by the other member of the jury were switched

to other graders and each grader scored Set B papers again seeing their partners' corrections and comments but not their grades on them.

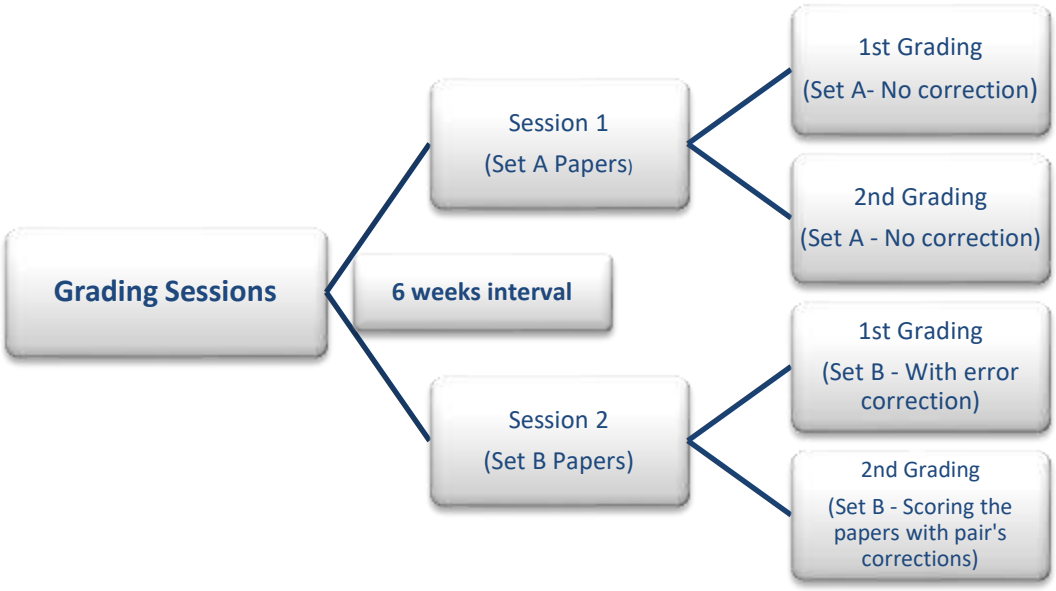

Figure 1: Conceptual framework of the study

3.4. Data Collection

The data gathered in this study consisted of the 12 graders' scores after 4 grading session and the recordings which include semi-structured interviews with 9 of the participants. The scores were collected to check if error correction makes a significant effect on raters' marks. Data Normality Test, Post Hoc Tests (for multiple comparisons), Pearson's Correlation Coefficient Tests (to indicate the strength and direction of a linear relationship between two random variables) were made using IBM SPSS version 20 software in order to identify the role of error correction in grading, to check the intra-rater and inter-rater reliability degrees of the rater pool in this study.

4. Findings

4.1. Data Distribution

This study aimed to find the effect of error correction on graders scores and to do this four sets of scores given by 12 graders to 20 papers were collected and computed. In order to select the statistical test types to implement (parametric or non-parametric tests), normality of the distribution of the scores was tested and the results were presented in Table 1.

Table 1. *Normality test results*

	Grading	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistics	df	Sig	Statistics	df	Sig
SET A	1	.106	120	.200*	.965	180	.628
	2	.109	120	.200*	.972	180	.910
SET B	1	.098	120	.200*	.970	180	.845
	2	.118	120	.200*	.981	180	.597

As it is vividly seen in Table 1, the scores assigned to each set of papers showed normal distribution according to the results of both Kolmogorov-Smirnov and Shapiro-Wilk tests ($p>0.05$). Thus, parametric tests were utilized to find answers to the aforementioned research questions on the role of error correction on scoring reliability.

4.2. Intra-Rater Reliability Findings

To check out the intra-rater reliability levels of each grader, 10 papers in Set A were scored by the graders in the first and the second grading sessions without any corrections and grader comments. To have an overall view, first, means of all these scores given by all these 12 graders were computed and given in Table 2.

Table 2. *Comparison of mean scores for Set A*

Descriptive Statistics					
	Grading	**Mean**	**N**	**sd**	**Std. Error Mean**
Set A	**1**	71.73	120	8.263	1.211
	2	73.27	120	8.461	1.259

The mean score of the first grading for Set A papers was found as 71.73 (mean score of 12 graders) and in the second grading which was done three weeks later the mean score was 1,54 more and it was 73,27 for the same papers. Considering a scale out of a 100 points a difference of 1,54 points in the mean scores looks acceptable but if it is not statistically significant. In Table 3, t-test results of these mean scores were given.

Table 3. *Paired samples t-test results for Set A*

Paired samples t-test							
	Grading	**Mean**	**sd**	**Std. Error Mean**	**t**	**df**	**Sig.(2-tailed)**
Set A	**1-2**	-1.540	5.217	1,402	-.730	119	.490

Table 3 results reveal that mean score difference in grading 1 and 2 by 12 graders does not show a significant difference according to t-test result ($t= -.730$, $df=119$, $p>0.05$). Thus, considering the mean scores comparison, it can be concluded that raters scored the same papers quite similarly although they had a 3-week interval between the scoring sessions and it can be a positive indicator of intra-rater reliability. Time variable is proven to be insignificant in this case, since the mean scores were just 1,54 different from each other.

After examining the mean scores of all the graders, now it is time to check each grader's own performance to have a better idea on intra-rater reliability levels. Pearson Correlation is a common statistic that is used to determine the amount of similarity between a grader's own scores assigned to the same papers, assigned in different times. The correlation value is a value between 0-1 and the higher the correlation the better it is. There are different considerations but Brown (2009) recommends a value of r=, 750 or more for acceptable intra-rater reliability estimates.

In Table 4, correlation of each grader's scores on Set A papers were given.

Table 4. *Paired Samples Correlations on Set A*

Paired Samples Correlations (Pearson's Product Moment)						
Set	**Rater**	**Scoring**	**N**	**df**	**P.**	**Correlation**
A	1	1 - 2	10	9	,013	,877
	2	1 - 2	10	9	,005	,937
	3	1 - 2	10	9	,003	,942
	4	1 - 2	10	9	,011	,887
	5	1 - 2	10	9	,009	,901
	6	1 - 2	10	9	,012	,883
	7	1 - 2	10	9	,014	,861
	8	1 - 2	10	9	,009	,904
	9	1 - 2	10	9	,028	,792
	10	1 - 2	10	9	,053	,741 *
	11	1 - 2	10	9	,007	,914
	12	1 - 2	10	9	,021	,806

It should be remembered that each grader scored the same set of papers twice and the analysis given in Table 4 shows the correlation of each graders' own scores which were produced with a three-week time interval. Findings revealed that all the graders except Grader 10 had considerably high score correlations (correlation of the scores of the same grader in 2 grading sessions on the same paper) even a 3-week interval was given in between two scorings. Grader 10's score correlation (r=,741) was not bad but below the average and Grader 2, 3, 5, 8, 11's score correlations were found to be over ,900 which is a good example of high intra-rater reliability. When all these grader scores checked if there was statistically, meaningful differences between two scoring sessions T-test results proved the findings of the Pearson test.

Table 5. *Paired samples t-test for each grader (Set A)*

Paired Samples T-Test										
						95% Confidence Interval				
			Mean	**Std. Dev.**	**Std. Error**	**Upper**	**Lower**	**T**	**df**	**Sig.**
A	**1**	**1 - 2**	-,300	6,567	2,077	-4,998	4,398	-,144	9	,788
	2	**1 - 2**	-,100	7,015	2,218	-5,118	4,918	-,045	9	,965
	3	**1 - 2**	-,100	6,691	2,116	-4,886	4,686	-,047	9	,963
	4	**1 - 2**	,270	6,766	2,140	-4,840	4,840	,160	9	,900
	5	**1 - 2**	-,700	7,334	2,319	-5,946	4,546	-,302	9	,770
	6	**1 - 2**	,900	7,125	2,253	-4,197	5,997	,399	9	,699
	7	**1 - 2**	-,320	6,673	2,124	-4,998	4,348	-,167	9	,763
	8	**1 - 2**	-,120	7,022	2,202	-5,212	4,902	-,063	9	,912
	9	**1 - 2**	3,18	8,401	2,703	-,2644	,9174	,1315	9	,244
	10	**1 - 2**	-6,60	8,852	2,799	-12,926	-,376	-2,390	9	*,041
	11	**1 - 2**	-,680	7,218	2,306	-5,936	4,566	-,286	9	,784
	12	**1 - 2**	3,22	8,599	2,719	-2,863	9,393	1,209	9	,254

Results of t-test given in Table 5 verify the situation noticed in Pearson's correlations. Only Grader 10's scores given to the same papers differed significantly (t= -2.390, df=9, p<0.05), the other 11 graders scored in a similar way that no statistically significant difference was found. To conclude, intra-rater reliability levels of all graders but Grader 10 were considerably high if the graders grade on their own and do not see the corrections or the comments of the other graders on papers they score.

However, the overall picture changes tremendously when error correction and grader comments intervene in Set B papers. It must be noted here that in Set B papers, graders were allowed to make error correction on students' mistakes in terms of grammar, vocabulary choice, word form, punctuation etc. and write their personal comments on the organisation, content of the paper and if they wish they were free to give possible feedback for the students as if these papers would be shown to the students. To have a look on what has changed in overall scores, means of all the total scores given by 12 graders to Set B were computed and given in Table 6.

Table 6. *Comparison of mean scores for Set B*

Descriptive Statistics					
	Grading	**Mean**	**N**	**sd**	**Std. Error Mean**
Set B	**1**	74.82	120	9.014	1.211
	2	67.27	120	8.302	1.259

The mean score of the first grading for Set B papers was found as 74.82 (mean score of 12 graders) and in the second grading which was done three weeks later the mean score was 7,55 points less and it was 67,27 for the same papers. The mean scores differed almost 10% and the only variable changed in this set of papers (compared to Set A papers) was rater comments and corrections on the papers. In Table 7, t-test results of these mean scores were given.

Table 7. *Paired samples t-test results for Set B*

Paired samples t-test							
Grading		**Mean**	**sd**	**Std. Error Mean**	**t**	**df**	**Sig.(2-tailed)**
Set B	**1-2**	7.550	13.620	2,971	-2.60	119	** .000

Table 7 results reveal that mean score difference in grading 1 and 2 by 12 graders shows a significant difference according to t-test result (t= -2.60, df=119, p<0.05). Thus, considering the mean scores comparison, it can be concluded that raters' scores changed meaningfully when they see the corrections and comments of their partners on the papers they score. It is clear that the scores assigned to the same papers reduce significantly (7,55 points a 10% deduction in mean scores) when errors are highlighted on papers.

In Table 8, correlation of each grader's scores on Set B papers were given.

Table 8. *Paired Samples Correlations on Set B*

Paired Samples Correlations (Pearson's Product Moment)						
Set	**Rater**	**Scoring**	**N**	**df**	**P.**	**Correlation**
B	1	1 - 2	10	9	,054	,737*
	2	1 - 2	10	9	,027	,821
	3	1 - 2	10	9	,083	,697 *
	4	1 - 2	10	9	,023	,840
	5	1 - 2	10	9	,052	,746 *
	6	1 - 2	10	9	,089	,682 *
	7	1 - 2	10	9	,081	,701 *
	8	1 - 2	10	9	,020	,803
	9	1 - 2	10	9	,014	,862
	10	1 - 2	10	9	,086	,689 *
	11	1 - 2	10	9	,056	,748 *
	12	1 - 2	10	9	,012	,885

Findings on Set B papers revealed that the graders score correlations (correlation of the scores of the same grader in 2 grading sessions on the same paper) reduced compared to their score correlations for Set A papers. Grader 1 (r=,737), Grader 3 (r=,697), Grader 5 (r=,746), Grader 6 (r=,682), Grader 7 (r=,701), Grader 10 (r=,689), and Grader 11 (r=,748) scored the same papers differently and their intra-rater values were found to be below the average when they score the same papers again seeing some corrections and rater comments. No rater's intra-rater reliability was found over ,900 and this might be an evidence that error correction on papers may have a negative influence over graders' judgements. Grader 3, 5 and 11 were found to have great score correlations and were highly reliable within their own scores when they were not influenced by their partners' comments and corrections in Set A scorings; however, all these raters scores changed significantly and their correlation values reduced from ,900s to ,600 which could be considered as a very striking change. When all these grader scores checked if this change between two scoring sessions is significant statistically, T-test results, again, proved the findings of the Pearson test.

Table 9. *Paired samples t-test for each grader (Set B)*

Paired Samples T-Test										
						95% Confidence Interval				
			Mean	**Std. Dev.**	**Std. Error**	**Upper**	**Lower**	**T**	**df**	**Sig.**
B	**1**	**1 - 2**	-6,70	8,845	2,797	-13,027	-,373	-2,395	9	*,040
	2	**1 - 2**	3,30	8,629	2,729	-2,873	9,473	1,209	9	,257
	3	**1 - 2**	5,70	5,012	1,585	2,114	9,286	3,596	9	*,006
	4	**1 - 2**	2,90	9,585	3,031	-3,957	9,757	,957	9	,364
	5	**1 - 2**	-5,90	8,144	2,575	-11,726	-,074	-2,291	9	*,048
	6	**1 - 2**	-5,50	3,504	1,108	-8,007	-2,993	-4,964	9	**,001
	7	**1 - 2**	-5,70	8,765	2,783	-8,618	-,112	-2,562	9	*,028
	8	**1 - 2**	3,24	8,623	2,648	-2,854	9,364	1,194	9	,252
	9	**1 - 2**	2,80	9,242	3,004	-3,916	9,210	,940	9	,388
	10	**1 - 2**	5,80	5,004	1,563	2,120	9,412	3,436	9	*,005

11	**1 - 2**	-5,80	8,096	2,536	-11,152	-,069	-2,282	9	*,049
12	**1 - 2**	,880	7,268	2,212	-4,162	5,986	,403	9	,701

As it can be noted in Table 9, Grader 1, 3, 5, 6, 7, 10 and 11 scored the same papers in Set B statistically differently ($p<0.05$) when they see their partners' corrections and comments while grading. It should be remembered that only Grader 10 was found to have statistically different scores in Set A papers, but this time in Set B papers where corrections and comments of the previous graders could be seen, 7 out of 12 graders were found to have low intra-rater reliability values.

To sum up the findings of this section, it could be stated that under normal conditions (when no artificial variables like corrected errors or grader comments intervene) all the participants except Grader 10 had high intra-rater values and were mostly scoring the papers similarly even after a considerable amount of time like three weeks. The same graders; however, tend to lower their grades like 10%, and 7 out 12 graders were found to have low intra-rater reliability values when they see their partners comments and corrections on the papers they grade.

4.3. Inter-Rater Reliability Findings

After finalising the analysis of the individual performances of the participants to have evidence on their intra-rater reliability degrees, it was aimed to examine their inter-rater reliability performances in this section. The graders were paired randomly (regardless of their experience or gender) as 1-7,2-8, 3-9, 4-10, 5-11 and finally 6-12 and it should be remembered that they hadn't come together to share their opinions or negotiate their scores on papers any time during the scoring sessions. In the first analysis given in Table 10, each pair's Set A score correlation results were presented.

Table 10. *Pearson's product moment test on correlations (Set A)*

Paired Samples Correlations (Pearson's Product Moment)						
Set	**Pair**	**Raters**	**N**	**df**	**P.**	**Correlation**
A	**1**	**1 - 7**	10	9	,059	,721
	2	**2 - 8**	10	9	,036	* ,798
	3	**3 - 9**	10	9	,097	,651
	4	**4 - 10**	10	9	,071	,703
	5	**5 - 11**	10	9	,099	,643
	6	**6 - 12**	10	9	,018	* ,855

Two graders in each pair graded the same papers in Set A twice individually and the analysis in Table 10 shows the results of the comparison of the scores of each pairs' grader scores considering the results of second grading (second grading results taken into account for each set since the independent variable "grader correction and comments" might be tested in Set B only in the second grading). It can be concluded that 2 out of 6 pairs were found to have acceptable correlation values over ,750 (Pair 2 r=,798; Pair 6 r=,855). In other words, Grader 2, 6, 8 and 12 could have high inter-rater reliability values under normal circumstances. When these pairs' scores were checked if there was statistically meaningful differences, T-test results in Table 11 proved the findings of the Pearson' Product Moment test.

Table 11. *Paired samples t-test for each pair of graders (Set A)*

Paired Samples T-Test										
			Mean	Std. Dev	Std. Err.	Upper	Lower	t	d	Sig.
A	1	1-7	-6,10	7,695	2,433	-11,605	-,595	-2,50	9	*,033
	2	2-8	-5,70	10,667	3,373	-13,331	1,931	-1,69	9	,125
	3	3-9	-10,10	7,671	2,426	-15,688	-4,712	-4,20	9	**,002
	4	4-10	-6,20	7,703	2,407	-10,402	-,528	-2,70	9	*,022
	5	5-11	-10,20	7,671	2,426	-15,688	-4,712	-4,20	9	**,001
	6	6-12	-2,60	8,622	2,648	-8,760	2,984	-,992	9	,362

T-test results revealed that Grader 1 and 7, 3 and 9, 4 and 10, 5 and 11 scored differently and this difference is statistically significant under normal conditions ($p<0.05$), whereas, Grader 2, 6, 8 and 12 assigned similar scores to the same papers and could have high inter-rater reliability values. In some pairs' scores (Pair 3&5) the mean difference in scores ranges up to 10 points and this is a remarkable difference considering that the overall mean was computed around 73 points in this set of papers.

The analysis presented in Table 12 reveals the correlation of the scores when graders see their partners' corrections and comments on the papers. In other words, up to this analysis, graders were totally unaware and unbiased of the reactions and judgements of their partners on the papers they both score, but this time they might have had the chance to see some clues of their partners' scoring.

Table 12. *Pearson's product moment test on correlations (Set B)*

Paired Samples Correlations (Pearson's Product Moment)						
Set	Pair	Raters	N	df	P.	Correlation
B	1	1 - 7	10	9	,023	* ,839
	2	2 - 8	10	9	,020	* ,849
	3	3 - 9	10	9	,003	* ,945
	4	4 - 10	10	9	,060	,719
	5	5 - 11	10	9	,037	* ,780
	6	6 - 12	10	9	,022	* ,842

The results presented in Table 12 show that inter-rater reliability levels increased and mostly passed over ,750 in the scores of all pairs but Pair 4. It should be noted that Grader 10 had low intra-rater reliability values in almost all the score computations in both Set A and Set B papers, thus, this grader might have had a negative attribution to its partner in this scoring considering the overall setting. Taking into account the role of error correction on graders' judgements, it can be seen that it had a positive effect on Pair 1's (r= ,839), Pair 2's (r= ,849), Pair 3's (r= ,945), Pair 5's (r= ,780), Pair 6's (r= ,842) score correlations and increased the inter-rater reliability levels of so called pairs' graders. The last analysis, t-test results prove this fact as well.

Table 13. *Paired samples t-test for each pair of graders (Set B)*

Paired Samples T-Test										
			Mean	Std. Dev	Std. Error	Upper	Lower	t	d	Sig.
B	1	1-7	4,50	8,593	2,717	10,647	-1,647	1,65	9	,132
	2	2-8	-2,70	8,744	2,765	-8,955	3,555	-,976	9	,354
	3	3-9	-1,00	3,801	1,202	-3,719	1,719	-,832	9	,427
	4	4-10	-6,40	7,726	2,465	-11,780	-,612	-2,44	9	* ,031
	5	5-11	4,80	8,495	2,416	-11,036	-1,796	-1,76	9	,103
	6	6-12	-2,90	8,736	2,788	-9,210	3,712	-,996	9	,338

According to the findings, it could be stated that only Pair 4's graders scored the papers significantly different ($p<0.05$) from each other and unfortunately Grader 10 again could be the reason of this difference. The other pairs all scored similarly and were found to have high inter-rater reliability levels. Remembering the fact that graders saw each others' error corrections and comments on the papers in this scoring session, the positive role of rater comments and corrections on papers to reduce the score gap between raters while doing cross-marking could be highlighted.

5. Discussion & Conclusions

This study aimed to analyse the role of error correction and rater comments on exam papers in such institutions where cross-marking is done to increase the reliability and accuracy of scoring. To be able to control the role of error correction on graders' scores two sets of exam papers were graded twice with a three-week time interval. In the first set (Set A) graders did not make error correction and they scored the papers twice to be able to check the intra-rater reliability degrees. Graders were free to write their comments and make error corrections on Set B papers while grading and the corrected papers were switched in each pair and graders were given the chance to see each others's comments and correction in the second grading of the same papers. The normality test results of both Kolmogorov-Smirnov and Shapiro-Wilk tests ($p>0.05$) revealed that the data had a normal distribution, therefore parametric statistical tests were implemented to see the role of error correction on graders' scores as the independent variable of this study.

The intra-rater reliability degrees of graders were estimated by comparing their first and second scorings of the same papers in both sets of papers. This scoring was done under actual conditions where graders do not know the other graders' comments, reactions and of course scores given to the same papers in Set A. The statistical analysis results which were made to explore the intra-rater reliability degrees of the participants were all over the expected degree ($r \geq ,750$) but Grader 10. Score correlation of this grader on its own gradings was found as ,741 (indeed this is not a very bad correlational degree and in some settings it could be accepted as reliable indeed). This finding revealed that the grader pool was made up of a successful pool of raters and they were mostly highly reliable in their own scorings (r=,942 for Grader 3, r=,937 for Grader 2, r=,914 for Grader 11, r=,904 for Grader 8, r=,901 for Grader 5). Likewise, paired samples t-test results justified the findings of the correlation test and all the participants except Grader 10 were proven to be scoring highly reliable when their own scores were compared. There was only 1,54 points difference between the mean

scores of the graders and this difference was not statistically significant (t= -.730, df=119, p>0.05). The good result driven from Set A papers changed remarkably when graders were allowed to make corrections and personal comments on students' essays in their first scoring of Set B papers. The graders were asked to switch the Set B papers with their partners and they were allowed to make their second scoring on Set B papers seeing their partners' corrections and comments. This time the mean score of the first grading for Set B papers was found as 74.82 (mean score of 12 graders) and in the second grading it was 7,55 points less and it was 67,27 for the same papers. The mean scores differed almost 10%. Score correlations reduced significantly for most graders (r=,697 for Grader 3, r=,682 for Grader 6, r=,737 for Grader 1, r=,746 for Grader 5). The only variable which was deliberately changed in this setting was error correction in cross-marking. As a result, considering all the findings related to the comparisons of participants own gradings, it can be concluded that error correction could have a negative impact on graders intra-rater reliability levels and can also cause graders score the essays 10% lower than what actually they would score when they take part in cross-marking sessions.

The next quality, which is sought in reliable assessment, is surely the inter-rater reliability. Before the inter-rater reliability degrees of graders were estimated, 12 graders were paired randomly and 6 pairs of grading teams were formed. Inter-rater reliability degrees of graders were estimated by comparing each graders' scores by their partners' scores assigned to the same set of papers. For the first set of papers (Set A) the scoring was done under actual conditions where graders do not know the other graders' comments, reactions and of course the scores given to the same papers. The statistical analysis results which were made to explore the inter-rater reliability degrees of the participants were not very high. Just 2 out of 6 pairs were found to have acceptable correlation values over 0,750 (Pair 2 r=,798; Pair 6 r=,855). The others were all below the expected level which was predetermined as 0.750 and also T-test results revealed that Grader 1 and 7, 3 and 9, 4 and 10, 5 and 11 scored differently and this difference is statistically significant under normal grading conditions (p<0.05), whereas, Grader 2, 6, 8 and 12 assigned similar scores to the same papers and could have high inter-rater reliability values. In some pairs' scores (Pair 3-5) the mean difference in scores ranges up to 10 points and this is a remarkable difference in mean scores. Nevertheless, the disappointing view of the graders' performance in Set A papers changed completely after they graded Set B papers in which they were able to see their partners' comments and corrections on them. Except the 4th pair, all the pairs' score correlations were over the critical limit 0,750 . Grader 10 ,who was found as the only grader in the rater pool whose intra-rater reliability was lower than the expected level, was one of the two members of this jury (Pair 4) and his/her grading performance might have affected the performance of the pair as well. Finally, taking into account the role of seeing their partners' error correction on graders' judgements, it can be seen that grader comment ad error correction had a positive effect on Pair 1's (r= ,839), Pair 2's (r= ,849), Pair 3's (r= ,945), Pair 5's (r= ,780), Pair 6's (r= ,842) score correlations and increased the inter-rater reliability levels of so called pairs' graders.

To summarise, the findings of this study revealed that graders could score more genuinely and can reflect their personal scoring performance much better if they do not see their colleagues comments and corrections on exam papers. In this way, their intra-rater reliability degrees remain high since they are not biased by an independent variable like the comments and corrections of the other raters. However, if a pool of raters is involved in a writing assessment process and if scoring reliably within the whole team is prioritised, making error correction on papers seems working in terms of increasing inter-rater reliability degrees. Thus, there appears a dilemma in which institutions have to decide

between the two, either let your graders score on their own and present higher intra-rater reliability or let yor graders effect each other by their comments and corrections, perform much better as a team and present higher inter-rater reliability degrees. This decision could be made by the faculties only after they decide what kind of assessment qualities they give primacy to: consensus among the raters or raters' free scoring and their own judgements. As for the implications of this study, the latter for sure seems more logical if hundreds or maybe thousands of papers are graded at once. Though making error correction on papers have both pros and cons, it seems to have a control on graders stay in the same line while grading and can enable them to justify or present their personal reasons on how they score or on the rationale they score on problematic papers.

As for the limitations and suggestions, the number of graders could have been more to have a better picture about the issue or the number of papers certainly, but the study had a voluntary basis that is why increasing the number of papers would mean far more workload, for that reason it was limited to ten papers to grade. Next, the grading purpose in any study is highly important. A study investigating the same issue on exam papers under real conditions could help a lot since the graders would perform differently knowing that those results would be real exam results. The amount of correction or the content of rater comment could be another variable that can be studied. In some studies (Janopoulos, 1992, Rust et al. 2003) even the colour of pen that was used in error correction was reported to be effective on graders' judgements while scoring the same paper, thus, various independent variables like the type of errors that could be corrected or not, grader comments on only organizational matters or the content of the essay etc. The last but not the least, in this effort in increasing the reliability of essay scoring, the importance of idea sharing and organising common norming sessions for scoring could be organised among institutions and the results and outcomes of these grading sessions could be reported in future studies to have more dimensions in assessment and gain further insights on what we are doing well and how we can do much better.

References

Bell, R.C. (1980). Problems in improving the reliability of essay marks. *Assessment & Evaluation in Higher Education*. 5(3): 254-263.

Bloxham, S. (2009). Marking and Moderation in the UK: False Assumptions and Wasted Resources. *Assessment & Evaluation in Higher Education*. 34(2):209-220.

Brown, G. (2009). The reliability of essay scores: The necessity of rubrics and moderation. *In Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice, and Research. Editors: Meyer L, Davidson S, Anderson H, Fletcher R, Johnston PM, Rees M. 43-50. Ako Aotearoa - The National Centre for Tertiary Teaching, Wellington, NZ 2009.*

Brown, G., Glasswell, K., Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*. 9 (2004):105-121.

Caryl, P. G. (1999). Psychology examiners re-examined: A 5-year perspective. *Studies in Higher Education*. 24 (1): 61-74.

Ecclestone, K. (2001). "I Know a 2:1 When I See It: Understanding Criteria for Degree Classifications in Franchised University Programmes." *Journal of Further and Higher Education*. 25 (3): 301-313.

Fleming, N.D. (1999). "Biases in Marking Students' Written Work: Quality?" *In Assessment matters in higher Education: Choosing and Using Diverse Approaches*, edited by S. Brown and A. Glasner, 83-92. Buckingham, UK.

Guanxin, R. (2007). The Reliability of Essay Marking in High-Stakes Chinese Second Language Examinations. *Academic Journal, Babel*. Vol 42(2):25-31.

Janopoulos, M. (1992). "University Faculty Tolerance of NS and NNS Writing Errors: A Comparison." *Journal of Second Language Writing*. 1 (2): 109–20. doi: 10.1016/1060-3743(92)90011

Johnson, M., Nádas, R., Shiell, H. (2009). An investigation into marker reliability and other qualitative aspects of on-screen essay. *Paper presented at the British Educational Research Association annual conference, Manchester University, September 2009*. Cambridge Assessment.

Kuper, A. (2006). Literature and Medicine: a problem of assessment. *Academic Medicine*: 81(10): 128-137.

Laming, D. (1990). "The Reliability of a Certain University Examination Compared with the Precision of Absolute Judgements: *Quarterly Journal of Experimental Psychology*. 42A (2):239-254.

O'Hagan, S.R., Wigglesworth, G. (2015). Who is marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context. *Studies in Higher Education*. 40:9, 1729-1747, DOI: 10.1080/03075079.2014.896890

Oruc, N. E. (2015). Testing your Tests: Reliability Issues of Academic English Exams. *International Journal of Psychology and Educational Studies*. 2015, 2 (2), 47-52.

Price, M.J., Carroll, B., O'Donovan, B., Rust, C. (2011). "If I Was Going There I Wouldn't Start From Here: A Critical Commentary on Current Assessment Practice. " *Assessment & Evaluation in Higher Education*. 36(4): 479-492.

Read, B., Francis, b, Robson, J. (2005). "Gender, Bias, assessment and Feedback: Analysing the Written Assessment of Undergraduate History Essays." *Assessment & Evaluation in Higher Education*. 30(3): 241-260.

Robson, J., Francis, B., Read, B. (2002). "Writes of Passage: Stylistic Features of Male Undergraduate History Essays. "*Journal of Further and Higher Education*. 26(4): 351-362.

Rom, M. C. (2011). Grading More Accurately. *Journal of Political Science Education*. 7: 208-223.

Rust, C. (2007). "Towards a Scholarship of Assessment." *Assessment and Evaluation in Higher Education*. 32 (2): 229–37. doi: 10.1080/0260293060080519

Rust, C., Price, M., O'Donovan, B. (2003). "Improving Students' Learning by Developing Their Understanding of Assessment Criteria and Processes." *Assessment and Evaluation in Higher Education*. 28 (3): 147–64.

Shavelson, R.J., Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shaw, S. (2008). Essay Marking On-Screen: implications for assessment validity. *E-Learning*. Vol: 5 (3). https://doi.org/10.2304/elea.2008.5.3.256

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment Research* & Evaluation. 9(4).

Wood, R., Quinn, B. (1976) Double Impression Marking Of English Language Essay and Summary Questions. *Educational Review*. 28 :(3): 229-246, DOI: 10.1080/0013191760280307

Yorke, M. (2008). *Grading Student Achievement in Higher Education: Signals and Shortcomings.* Abingdon: Routledge.

Yorke, M. (2011). "Summative Assessment: Dealing with the 'Measurement Fallacy'." *Studies in Higher Education*. 36 (3): 251–73. doi: 10.1080/03075070903545082.