# RATERS' FATIGUE AND THEIR COMMENTS DURING SCORING WRITING ESSAYS: A CASE OF IRANIAN EFL LEARNERS

Amir Mahshanian[1]
Abbas Eslami Rasekh[2]
Saeed Ketabi[3]
*University of Isfahan, Iran*
mshn_amir@yahoo.com[1]; abbasseslamirasekh@yahoo.com[2]; ketabi@fgn.ui.ac.ir[3]

First received: 28 January 2017 Final proof received: 22 September 2017

**Abstract**

Rating accuracy in writing among EFL learners is crucial in determining their English proficiency. Despite the importance of its accuracy, little is known about various factors may affect the accuracy of rating writing essays. This study examines how raters' comments on EFL writing tasks change as a result of fatigue. To this end, four raters were selected and each given 28 essays to score and comment on. Six general types of raters' comments (i.e., those on grammar, choice of words, organization, punctuation, dictation, and capitalization) were into focus in this study. Overall, results suggested that fatigue affects raters' frequency of comments on grammar, choice of words, and organization, and that raters' comments on punctuation, dictation, and capitalization do not seem to change significantly due to the effect of fatigue. Furthermore, this study revealed that the most and least frequent comments in 112 scored essays were those on grammar and dictation, respectively.

**Key words**: raters' fatigue; frequency of comments; type of comments; EFL writing tasks

Writing has always had a place in EFL curriculum. The ability to write in an L2; however, may be even more important recently. Today, the need to learn to write in a second or foreign language, whether to transact business, interact on social networking sites, or to pursue academic degrees seem to be an essential one. As a result, many teachers will find themselves in need of teaching and scoring writing tasks effectively and may not feel well-prepared in so doing (Celce-Murcia, Brinton, & Snow, 2014). Among teachers and raters, there are many who agree on the fact that rating and scoring writing tasks is challenging due to its subjectivity and if enough care is not exercised in this regard they might end in test bias.

Ling, Mollaun, and Xi (2014) assert that scoring quality is critical to the validity and fairness of any test. For tests with constructed responses, for example, essays or speaking tasks, they argue that human raters are often employed to determine a score and comment on the responses in which case, it is a challenge to ensure scoring quality. They go further to point out that while human raters are trained to provide exact, unbiased, and reliable ratings based on scoring protocols and guidelines, their performance may be negatively affected by construct-irrelevant factors other than the scoring protocols. For example, task complexity and difficulty, task type, examinees' characteristics, and raters' background and training experiences have been found to be related to rating accuracy (Brown, 1995; Caban, 2003; Shohamy, Gordon, & Kraemer, 1992).

Furthermore, assessing and evaluating EFL writing tasks involve both assigning a score or grade to an essay and importantly commenting on it (Ling et al. 2014). Many studies in the literature (e.g., Johnson, Penny, & Gordon, 2001; Liu, Allspach, Feigenbaum, Oh, & Burton, 2004) have indicated that at least 2 raters should score students' writing assessments to improve inter-rater reliability. However, even for assessments that characteristically demonstrate high levels of rater agreement, 2 raters scoring the same essay can occasionally report different, or discrepant, scores (Johnson et al., 2001).

Inconsistency of scoring criteria could closely be related to raters' fatigue and as a result would affect test takers scores and introduce assessment bias to the process of scoring. Fatigue is particularly important for professions in which judgment errors are costly. The issue of fatigue is essentially a time-based concept and when undertaking activities requiring concentration, the longer one takes a task, the more fatigue there would be (Drave, 2011). Thus, in rating EFL writing tasks, fatigue has come to be known as a significant factor to influence raters' judgment and scoring quality (Ling et al., 2014). With respect to fatigue, in the literature, a number of characteristics and definitions have been put forward by researchers (Geacintov & Peavler, 1974; McCormick, 1970). For example, fatigue can be seen as mental or physical signs such as tiredness, drowsiness, sleepiness, and lack of concentration (Cumming, 1954). Fatigue is also believed to be qualitative and quantitative output

 *doi:* dx.doi.org/10.17509/ijal.v7i2.8347

reduction (Anastasi, 1979). Drawing on Anastasi's definition (1979), Ling et al., (2014) argues that output reduction leads to the high frequency of errors. Ling et al., (2014), also suggest that these signs are subtler than the output indicators in that they provide researchers with more space for error recognition.

In the literature, a plethora of research has attempted to explore the impact of construct irrelevant factors (e.g., fatigue, raters' attitude, etc.) on raters' judgment (e.g., Bendig, 1955; Constable & Andrich, 1984; Cumming, Kantor, & Powers, 2002; Cumming, 1990; Weigle, 1994; McNamara & Wesley, 1996; Drave, 2011; Ling et al. 2014; Lumley & McNamara, 1995; Massey, 1977; Schumm & Vaughn, 1991). Some researchers such as Weigle (1994), and McNamara and Wesley (1996) agree on the fact that with careful monitoring and training of raters, scoring procedure might end in reliable results and unbiased judgments. Too, some scholars (e.g., Cumming, 1954; Bendig, 1955; Drave, 2011; Tucker, 1948; Massey, 1977; Wohlhueter, 1966; Liu et al., 2004) maintain that construct-irrelevant factors, fatigue in particular, do not significantly affect test test-takers' scores and test-givers' scoring method. Bendig (1955), as a case in point, investigated the reliability of rater scoring and its possible loss as a result of fatigue and suggested that judgment fatigue did not affect scoring reliability. Drave (2011), as another example, explored the fatigue issue in the context of rating essays displayed onscreen. His finding suggested that only a few raters were affected by fatigue. It should be noted that despite the fact that in most studies, there has not been observed a significant effect of fatigue, they used relatively simple tasks that required a minimum level of attention, demanded a low level of cognitive ability, and lasted for a relatively short period of time (Ling et al., 2014). This issue was addressed in the present study in the sense that participants were given a demanding task of scoring 28 EFL essays.

On the other hand, some other studies concluded that raters' fatigue can negatively influence raters' judgment, the reliability, and the consistency of language tests (e.g., Wohlhueter, 1966; Hiramatsu, 2000; Goodall 2011; Sprouse, 2007; Ling et al., 2014). In this respect, Hall and Sheyholislami (2013), argue that raters' comments and the way they change, their comprehension of the language, and their various biases are influential in language test scores and inferences. Moreover, Sprouse (2007) maintained that fatigue can cause variance increase and a decrease in the violations acceptability (Ling et al., 2014). It should be noted, however, that the focus of Sprouse's study (2007) was syntactic errors and considering the inconsistency of raters' judgment based on only one criterion (i.e., syntax) as a result of fatigue is an incomplete vision. Also, Ling et al., (2014) exploring the effect of raters' fatigue on scoring speaking test admitted that raters' fatigue affect their judgment in scoring constructed response in speaking tests. It should be pointed out, however, that results of his study on speaking tests cannot be generalized to scoring writing tasks, which was the focus of the present study.

By and large, scoring writing tasks may introduce construct-irrelevant factors to scoring and commenting, and affects validity and fairness of the test. Fatigue is one of the factors that can negatively affect human performance in general and scoring and commenting on essays, in particular. Although many studies have highlighted the effect of fatigue on test takers/givers' performance on language tests, (e.g., Bendig, 1955; Constable & Andrich 1984; Cumming et al., 2002; Cumming, 1990; Drave, 2011; Ling et al., 2014; Lumley & McNamara, 1995; Massey, 1977; Schumm & Vaughn, 1991), very little is known about its effects on a raters' scoring quality in speaking and writing tasks (Drave, 2011; Ling et al., 2014). Also, results of studies regarding the effect of fatigue in the literature were quite conflicting in the sense that some suggested that fatigue can affect human judgments' significantly in language tests (e.g., Tucker, 1948; Cumming, 1954; Bendig, 1955; Massey, 1977; Wohlhueter, 1966; Liu et al., 2004; Drave, 2011), whereas others argued that the effect of fatigue on test-takers', or raters' judgments is not significant (e.g., Wohlhueter, 1966; Hiramatsu, 2000; Goodall 2011; Sprouse, 2007).

Contrary to studies focusing on simple and/or short tasks to investigate the impact of fatigue on human judgment (e.g., Cumming, 1954; Bendig; 1955; Snyder, 2000;), the present study investigated the effect of fatigue on raters who were given the demanding task of scoring and commenting on EFL writing tasks in a 3-hour-session. Thus, in an attempt to fill the gap in the literature, the current study was designed to examine the effects of fatigue on the consistency of raters' types of comments in scoring EFL writing tasks.

The present study was an attempt to investigate the effect of fatigue on the consistency of raters' comments while scoring EFL writing tasks. In technical terms, the following research questions were intended to be addressed: (1) Does fatigue bring about changes in the way raters comment on EFL writing tasks (essays) while scoring them?, (2) How does raters' frequency of different types of comments change after scoring 28 EFL writing tasks (essays)?, and (3) What types of comments are the most, and least frequent ones among raters while scoring EFL writing tasks (essays)?

## METHOD

This study employs an ex-post-facto design which intends to explore how raters' fatigue relates to the

type of their comments while scoring EFL writing tasks, and the extent to which raters frequency and type of comments are affected by fatigue.

***Participants***

Four EFL raters, with more than 8 years of foreign language teaching experience, in 2 language schools in Iran were selected to take part in this study (see Table 1 below). Also, 28 upper intermediate EFL learners were a part of this study as they were given a writing task to complete before the scoring procedure.

**Table 1.** Participants (Raters)

| Rater | Order of Scoring | Number of Scored Essays | Gender | Age | Years of Experience |
|---|---|---|---|---|---|
| 1 | 1-28 | 28 | male | 40 | 20 |
| 2 | 28-1 | 28 | male | 28 | 8 |
| 3 | 1-28 | 28 | male | 42 | 18 |
| 4 | 28-1 | 28 | male | 27 | 8 |

**Instruments**

The materials used in this study were IELTS sample topics for writing taken from Brown and Richards (2011). Learners were taught, based on the 3rd unit of the course book (IELTS Advantage Writing Skill), how to write an opinion essay. Also, a list of do's and don'ts conducted by the researchers was then employed to instruct the learners how to write about the topics. Further, a random IELTS writing topic was given to the learners to complete in one hour. The task, as the testing material, asked the learners to write a 5-paragraph essay (250 words) regarding the given topic.

**Procedure**

In order to control for variables, other than fatigue, learners needed to be homogenized. Thus, 60 EFL learners were randomly assigned to complete a writing task on a given topic. Following this, 6 raters, were requested to score and comment on the essays taking into account categories for evaluating writing adapted from Brown (1991). To ensure that the learners have almost similar writing proficiency, 40 learners, with almost same scores, were chosen for the purpose of this study, and requested to complete a writing task on a second topic, and 20 learners were excluded.

Although learners were selected exercising a lot of care, the process of homogenizing learners, in terms of writing proficiency went on. Accordingly, some other learners (i.e., 12 learners), were excluded from the study. Based on learners' scores and raters' judgments on their essays, researchers and expert judges decided that these 12 learners were not suitable for the purpose of this study due to their incompatibility of writing proficiency with other 28 learners. Thus, 28 remained EFL learners were asked to write an essay for the 3rd time on a different topic (i.e., a compare/contrast essay entitled "homeschooling vs. going to school"). Also, it should be noted that to motivate learners to do their best in writing tasks, the tasks were introduced as a part of must-do activities of the course, for which the instructors assign scores, and without which learners may lose scores and fail the course.

As for learners, raters were homogenized and 4 raters all with more than 8 years of teaching/ rating experience in EFL contexts, all male, and all with non-significant mean of score difference, and non-significant difference in mean of total frequency of comments, were asked to score and comment on the essays. What is worth adding is that raters were not allowed to take any break intervals when scoring the tasks during which they were closely observed by the researchers. The process of scoring lasted almost 3 hours. Also, before scoring the essays, raters were provided with rubrics for evaluating writing (i.e., those adapted from Brown, 1991) and with sample scored essays including raters' comments to have a general overview of writing evaluation (e.g., Richards & Brown, 2011). Drawing on Brown's categories (1991), the comments were to be on the content, organization, discourse, syntax, vocabulary, and mechanism. For the purpose of this study, the most frequent types of comments including those on grammar, choice of words, punctuation, dictation, capitalization, and organization were into focus.

Raters also needed to be motivated in order for their judgments to be as accurate and precise as possible. Thus, a few rewarding actions were in order (i.e., a permanent pay rise, an option for choosing the level of the classes to teach for the next two terms, and a 500.000-Rls gift card) providing accurate scoring and careful comments based on the rubrics were practiced.

As regards fatigue measurement, Theander (2007), argues that although there are approximately 250 measurement methods, researchers do not agree on a unified definition for fatigue. The most widespread scale for fatigue measurement; however, is (MAF) which is used for self-reported fatigue estimation (Drave, 2011). Drave (2011) also asserts that in humanities fatigue is defined "as a loose set of deleterious physical, emotional, behavioral and cognitive symptoms which negatively impact human performance" (p. 4). For the purpose of this study, fatigue is measured by taking Ling et al. (2014)'s concept of "output reduction" - comments reduction in the case of this study, and self-reported symptoms of fatigue (Drave, 2011) into consideration. In other

words, the frequency of comments made by the raters, and self-reported symptoms of fatigue (based on the results of the interview) were into focus in measuring raters' fatigue.

Thus, in order to ensure that the inconsistency of comments is due to fatigue and to control for other variables (e.g., the order of essays), raters were asked to score comment on essays in an opposite order (i.e., rater1 and 3 scored essays from no.1 to no 28, whereas rater 2 and 4 scored essays from no.28 to no.1). Also, the frequency of their comments on the essays was precisely calculated, and they were interviewed after the scoring procedure. It should be noted that, in retrospective interviews, in the end, raters were asked whether they had suffered from fatigue and how it affected them (see table 8 for the results). Also, confidentiality of the interviews was taken into consideration. For the list of interview questions (i.e., yes-no and open-ended ones) see appendix 7.

## FINDINGS

### Comments on Grammar

In the interest of space, the descriptive statistics of the frequency of comments on grammar is shown in Appendix 1. As for all the other types of comments (see appendices1-6), in seven groups, the mean, standard deviation, standard error of measurement, within 95% confident interval, minimum, and maximum of the data are estimated. It should be noted that in these appendices, 28 papers were divided to groups of four for analysis. Thus, group 1 is the first four papers which were rated (i.e., essay number 1 to 4), group 2 is the second four papers being rated (i.e., essay number 5 to 8), and so on. As earlier mentioned, since there are 4 raters as subjects of this study and in each group, they rate 4 papers, the total number of the papers to be scored and commented on, in one group is 16 and the total number of all papers in all groups to be compared are 112. An ANOVA, also was run to show the meaningfulness of the relationship of the frequency of comments on grammar in groups (see Table 2).

**Table 2.** ANOVA for the comments on grammar

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 188.607 | 6 | 31.435 | 3.739 | .002 |
| Within Groups | 882.813 | 105 | 8.408 | | |
| Total | 1071.420 | 111 | | | |

According to the table in which the amount of p value is estimated at (0.002), there is a meaningful relationship between the frequency of comments on grammar in 7 groups. Thus, it can be argued that by passage of time, the frequency of comments on grammar differed due to the effect of fatigue. The degree to which these comments on grammar are different will be discussed below.

### Multiple Comparisons of Comments on Grammar

To compare each of these seven groups with one another regarding the frequency of comments on grammar, as well as other criteria, and find a meaningful relationship between any two of them a Post hoc LSD test was run and results suggested that first, there is no significant relationship between group 1, and group 2, 3, 4, and 5. This shows that the effect of fatigue is not significant on the frequency of comments on grammar when raters score the first 20 papers. Second, there has been observed a significant relation between group 1 and group 6 and 7. That is to say, fatigue starts to affect human raters significantly, regarding the comments on grammar, after scoring 20 essays. This suggests that the more paper the raters score and comment on, the more fatigued they become, and as a result the frequency of their comments on grammar would be minimized. Figure 1 shows the plot for the means of frequency of total comments on grammar which indicates that the most frequent ones are in the 2nd group (i.e., essay number 5 to 8) and the fewest comments on grammar are in the 6th group (i.e., essay number 21 to 24). Thus, as is clear in Figure 1 below, fatigue affects raters' comments on grammar significantly after scoring 20 papers.

### Comments on Choice of Words

As mentioned earlier, the descriptive statistics of the frequency of comments on choice of words is shown in appendix 2. An ANOVA was run to show the significance of the frequency of the comments on choice of words (see Table 3, below).

According to the table in which the p value is estimated at (0.046), there is a meaningful relationship between the frequency of comments on choice of words in 7 groups. Thus, one can argue that by passage of time, the frequency of comments on choice of words differed due to the effect of fatigue. The degree to which these comments on choice of words are different will be discussed in the next section. It should be noted, however, that mean difference in choice of words in 7 groups is not as much as that of grammar.

### Multiple Comparisons of Choice of Words in Seven Groups

Results of the post-hoc LSD test suggest that there is not any significant relationship between group 1,

and group 2, 3, and 4. This implies that the effect of fatigue is not significant on the frequency of comments on choice of words when raters score the first 16 papers. It should be noted, however, that group 1, has a significant relationship with group 5, 6, and 7. That is to say, fatigue starts affecting human raters significantly, regarding the comments on choice of words, after scoring 16 essays. This is despite the fact that the effect of fatigue on comments on grammar became significant after 20 papers. Thus, fatigue affects comments on choice of words sooner than those on grammar. Also, results suggest that the more papers the raters score and comment on, the more fatigued they become, and as a result the frequency of their comments on choice of words would be minimized. Figure 2 below clearly depicts the means of frequency of total comments on choice of words with the most frequent ones in the 1st group (i.e., essay number 1 to 4) and the fewest comments in the 7th group (i.e., essay number 21 to 24). Thus, as is clear in the figure, fatigue affects raters' comments on choice of words after scoring 16 papers.

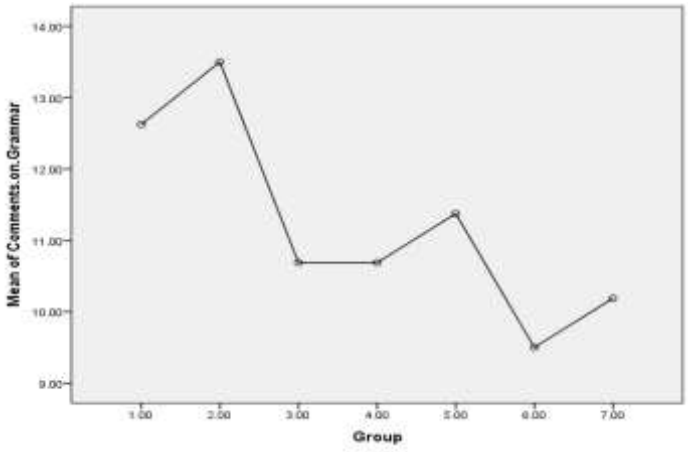


**Figure 1.** Means plot for the comments on grammar

**Table 3.** ANOVA for the frequency of comments on choice of words

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 16.554 | 6 | 2.759 | 2.229 | .046 |
| Within Groups | 129.938 | 105 | 1.238 | | |
| Total | 146.491 | 111 | | | |

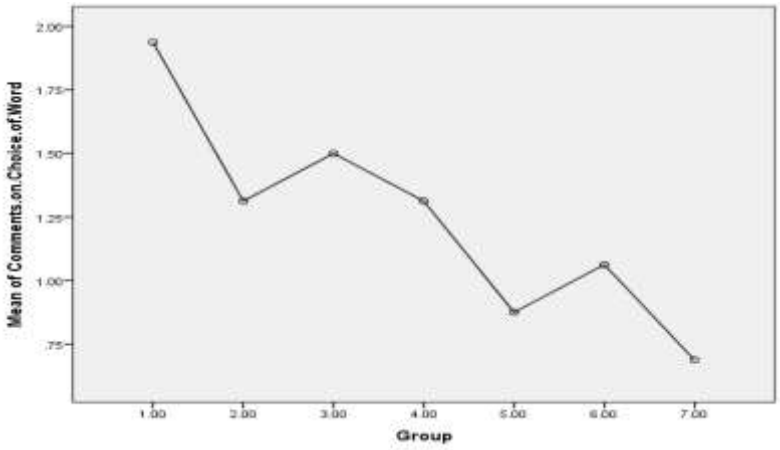


**Figure 2.** The mean of comments on the choice of words

**Comments on Punctuation**

As for the frequency of comments on grammar, and choice of words, and in the interest of space, the descriptive statistics of the frequency of comments on punctuation is shown Appendix 3. An ANOVA was run to investigate the significance of the relationship among frequency of comments on punctuation in 7 groups, and the results are shown in Table 4.

According to the estimated significance of the p-value in Table 4 (sig=0.032), it can be argued that although there is a significant relationship between the frequency of comments on punctuation in 7 groups, this is not as strong a relationship as it was for the comments on grammar. This, then, suggests

that by the passage of time, the frequency of comments on punctuation differed due to the effect of fatigue. This mean difference, however, is not as much as that of grammar. The degree to which these comments on punctuation are different will be discussed as follows.

**Table 4.** ANOVA for the frequency of comments on punctuation

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 20.089 | 6 | 3.348 | 2.410 | .032 |
| Within Groups | 145.875 | 105 | 1.389 | | |
| Total | 165.964 | 111 | | | |

**Multiple Comparisons of Comments on Punctuation in Seven Groups**

According to the results of post-hoc LSD test, there has not been observed any significant relationship between any 2 groups. This implies that the effect of fatigue is not significant on the frequency of comments on punctuation, and that it could be due to fact that frequency of comments on punctuation is very low (i.e., approximately fewer than 2 for each essay). Figure 3, is the means plot for the means of frequency of total comments on punctuation which shows the least frequent ones in the 2nd and 7th groups (i.e., papers 5 to 8 and 25 to 28, respectively) and the most comments on punctuation in the 6th group (i.e., papers 21-24).

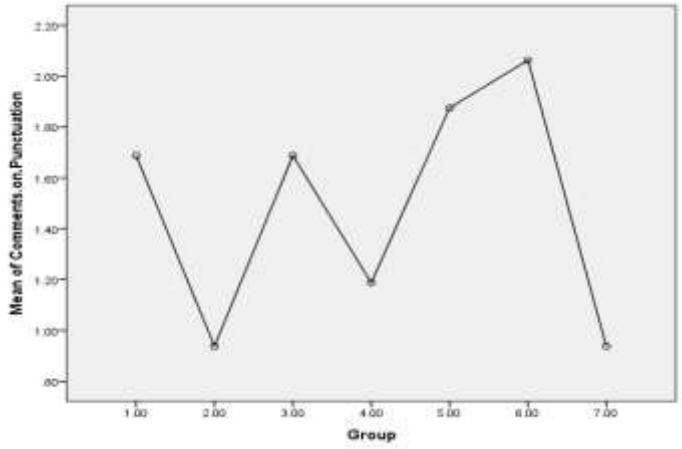


**Figure 3.** Mean of comments on punctuation

**Comments on Organization**

As for the frequency of comments on grammar, choice of words, punctuation, and in the interest of space, the descriptive statistics of the frequency of comments on organization is shown Appendix 4. An ANOVA was run to investigate the significance of the relationship among frequency of comments on organization in 7 groups, and the results are shown in Table 5.

**Table 5.** ANOVA for the frequency of comments on organization

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 32.804 | 6 | 5.467 | 4.256 | .001 |
| Within Groups | 134.875 | 105 | 1.285 | | |
| Total | 167.679 | 111 | | | |

Based on the table, the significance of the p-value is estimated at (0.001). Thus, it can be argued that there is a significant relationship between the frequency of comments on organization in 7 groups, and this is as strong a relationship as it was for the comments on grammar. This, then, suggests that by the passage of time, the frequency of comments on organization differed a great deal due to the effect of fatigue. The degree to which these comments on organization decreased will be discussed as follows.

**Multiple Comparisons of the Comments on Organization**

Results of the post-hoc LSD test suggests that there is a significant relationship between group 1, and group 2, 3, and 4. This implies that the effect of fatigue is not significant on the frequency of comments on organization when raters comment on the first 16 papers. However, group 1, has a significant relationship with group 5, 6, and 7. That is to say, fatigue starts to affect human raters significantly, regarding the comments on organization, after scoring 16 essays. This is in line with the effect of fatigue on comments on choice of

words as it became significant after commenting on 16 papers. Results, further suggest that the more papers the raters comment on, the more fatigued they become, and as a result the frequency of their comments on organization would be minimized. Figure 4 is the plot for the means of frequency of total comments on organization with the most frequent ones in the 1st group (i.e., essay number 1 to 4) and the lowest in the 7th group (i.e., essay numbers 25-28). Thus, as is depicted, fatigue affects raters' comments on organization significantly after scoring 16 papers.

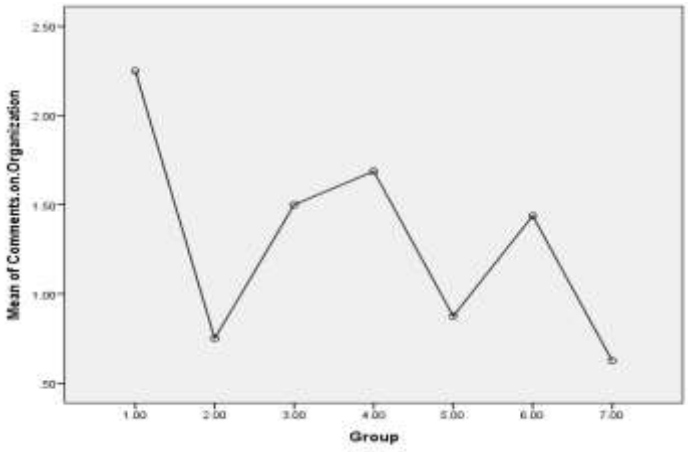


**Figure 4.** Mean of comments on organization

**Comments on Dictation**

As for the frequency of comments on grammar, choice of words, punctuation, organization, and in the interest of space, the descriptive statistics of the frequency of comments on dictation is shown in Appendix 5. An ANOVA was also run to explore the significance of the relationship among frequency of comments on dictation in 7 groups, and the results are shown in Table 6.

**Table 6.** ANOVA for the frequency of comments on dictation

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 17.089 | | 6 | 2.848 | .007 |
| Within Groups | 94.688 | | 105 | .902 | |
| Total | 111.777 | | 111 | | |

As shown clearly in the table, the significance of the p-value is estimated at (0.007). Thus, it can be argued that although there is a significant relationship between the frequency of comments on dictation in 7 groups, one cannot distribute this scattered change of means to the effect of fatigue or the number of essays to be rated (see below).

**Multiple Comparisons of the Comments on Dictation**

Results of post-hoc LSD test were run and shown in Appendix 5. According to data based on this test, there is no significant relationship among any of the seventh group. This could be caused by the format of writing which was a word document and that would correct almost all the dictation-related errors. Figure 5 is the plot for the means of frequency of total comments on dictation which shows the most frequent ones in the 4th group (i.e., essay no. 13 to 16) and the fewest comments in the 6th group (i.e., essay no. 21 to 24).

**Comments on Capitalization**

Descriptive statistics of the frequency of comments on capitalization is shown. In seven groups, the mean, standard deviation, standard error of measurement, within 95% confident interval, minimum, and maximum of the data is shown in Appendix 6. An ANOVA was also run to show the significance of the comments on capitalization, results of which are reported in Table 7.

Based on the significance of the p-value in Table 7 which is estimated at (.049), it can be argued that although there is a significant relationship between the frequency of comments on capitalization in 7 groups, this is not as strong a relationship as it was for the comments on grammar, organization, and dictation. Thus, it can be suggested that by passage of time the frequency of comments on capitalization differs due to the effect of fatigue. This mean difference however, is not as much as that of grammar, organization, and dictation. The degree to which these comments on capitalization are different will be discussed below

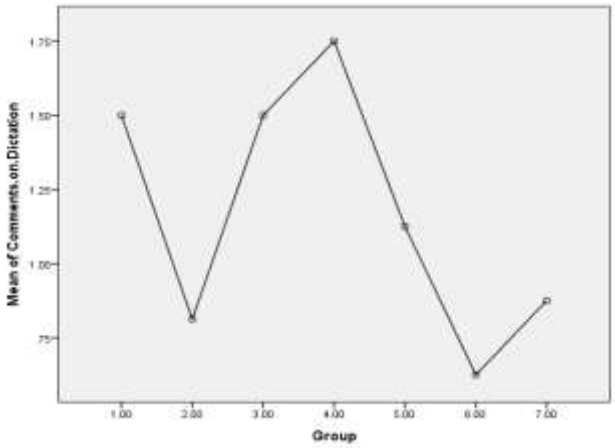


**Figure 5.** Mean of comments on dictation

**Table 7.** ANOVA for the frequency of comments on capitalization

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 9.839 | 6 | 1.640 | 2.218 | .049 |
| Within Groups | 83.938 | 105 | .799 | | |
| Total | 93.777 | 111 | | | |

**Multiple Comparisons of Comments on Capitalization in Seven Groups**

In an act of comparing seven groups with respect to the frequency of comments on capitalization, and investigate the relationship between the two of them, a Post hoc LSD was used. Based on the results of this test, there is not a significant relationship between any of the seven groups. This could have its roots in the format of writing which was a word document with its error-correction software which could underline errors related to capitalization. Figure 6 depicts the means of frequency of total comments on capitalization in which the most frequent ones have been observed in the 3rd group (i.e., essay no. 13 to 16) and the least in the 6th group (i.e., essay no. 21 to 24).

**Interviews**

There were 4 interviews, as mentioned above, which were recorded and transcribed, and finally reviewed and analyzed using, an emergent, constant-comparative method of grounded interpretation, (adopted from Cumming, 2011). The summary of subjects' responses to the interview questions (see appendix 7) is shown in Table 8.

In the interviews, all raters admitted that they had experienced fatigue during scoring the writing tasks. In addition, they all noted that their pain in their muscles, eyes, hands, necks, their distraction, sleepiness, dizziness, and unwillingness for giving more comments, were among the manifestations of fatigue and attributed these to the task of scoring essays for long hours.

## DISCUSSION

The present study was conducted in order to explore the effect of fatigue and the number of essays on raters' type and frequency of comments. This paper made an attempt to fill the gaps of the previous research studies carried out with its main focus on the discrepancy of comments made by raters when scoring EFL writing tasks.

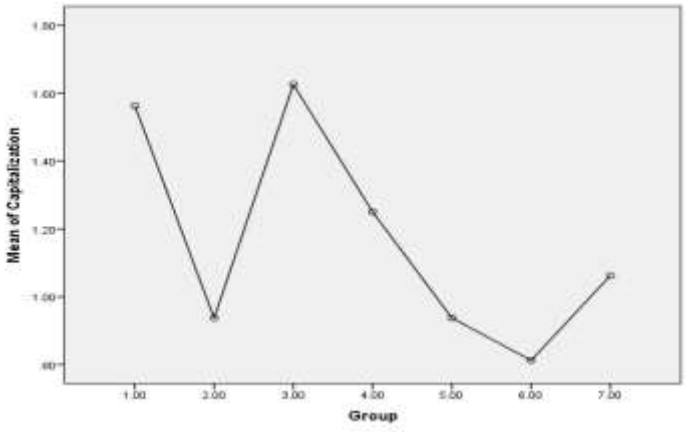


Figure 6. Mean of comments on capitalization

Table 8. Summary of the interviews

| Questions | Rater(s) with the same responses | Rater(s) with different responses |
|---|---|---|
| 1 | 4 | 0 |
| 2 | 4 | 0 |
| 3 | 4 | 0 |
| 4 | 4 | 0 |
| 5 | 4 | 0 |
| 6 | 4 | 0 |
| 7 | 4 | 0 |
| 8 | 4 | 0 |

As discussed earlier, there have been some studies with the focus on the effect of raters' fatigue on scoring speaking tests or in areas rather than language testing, few of which, however, dealt with its effect on scoring writing, in general, and differentiation in the type of the comments. One could mention its detailed analysis of the effect of fatigue on every type of comments as a strength of this study as opposed to rather general analysis of the effect of fatigue on scores given by raters (e.g., Guangming, Mollaun, & Xi, 2014). As another strength of this study, one could bring up its contribution to raters' writing assessment protocols in line with most recent studies (i.e., those conducted by Snyder, 2000; Sprouse, 2007; Hiramatsu, 2000; Goodall, 2011). The major findings of the present study can be summarized as follows;

1- Fatigue brings about changes in the way raters comment on EFL writing tasks while scoring them.
2- Fatigue affects raters' frequency of comments on grammar after scoring/commenting on 20 essay papers.
3- Fatigue affects raters' frequency of comments on choice of words, and organization after scoring/commenting on 16 essays.
4- Fatigue does not affect raters' comments on punctuation, dictation, and capitalization after scoring/commenting 28 essays.
5- The most, and least frequent comments in 112 scored essays were those on grammar and dictation (and capitalization), respectively.

In keeping with previous research studies (e.g., Massey, 1977; Wohlhueter, 1966; Weigle, 1994; McNamara, 1996; Hiramatsu, 2000; Liu et al., 2004; Goodall 2011; Sprouse, 2007; Drave, 2011; Ling et al., 2014), the present study highlighted the necessity of raters' training, and the importance of assessment protocols in order to avoid test bias. This study also suggested that fatigue can endanger even highly qualified raters' judgment in that the frequency and type of their comments on rated essays would change in an unfair manner from the first to the last few scored essays. Finally, the best break time for essay raters in order to have unbiased judgment when scoring and commenting on EFL writing tasks is the one after scoring 16 essays (5-paragraph essays with almost 250 words)..

There have also been some limitations in this study despite attempts to move through them. First, and foremost, the subjects of the study were only 4 Iranian raters, and one could question the size of the population. However, finding homogeneous raters who can participate in the present study and be observed during the process, was a painstaking task for the researchers taking almost 3 months. Furthermore, learners were asked to type their essays in a word-document for the ease of scoring, and similarity among essays. This, however, might endanger the authenticity of the task in the sense that some errors made by learners would have already been corrected by Microsoft Word Office's error-correction software. Too, raters were asked to score and comment on the essays within a 3-hour period having no break interval. This also questions the validity of the research in that one can argue that in normal situations, raters will never rate 28 essays in 3 hours, without any breaks.

A major goal of investigating factors which affect raters' judgments and consistency of scoring is to increase the level of test fairness and reliability, and to minimize test bias. It is of paramount significance for raters to apply the criteria of rating constantly with the maximum similarity. Also, examining these factors aims at understanding test constructs and test inference to define construct validity more precisely. Pinpointing the areas of inconsistency among raters and the criteria raters apply (those which are not included in the rating instructions), may provide test developers with more opportunities to reevaluate, refine, and develop the construct using rating criteria. Thus, investigating inconsistency among raters in scoring writing tasks is a practical function in the process of test validation.

## CONCLUSION

As discussed earlier, results of the data analysis suggested that there is a significant relationship among groups regarding the types of comments including the ones on grammar, choice of words, punctuation, dictation, capitalization, and organization. This implies that fatigue brings about changes in the way raters comment on essays from the first to the last few ones. Although a lot of

distinctions have been observed on the way raters commented on the essays from the first to the last few ones, one is considered the most significant and that is comments on grammar. Notwithstanding the p value which indicated the significant relationship among the frequency of all comments in groups, those on grammar were a lot more variable than the other types. That is to say, comments on the choice of words, punctuation, dictation, capitalization, and organization varied from (0) to (5) on each essay which is considered very few in number. This is in contrast with the frequency of comments on grammar which varied from (6) to (25) on each essay. Surprisingly, comments on grammar are a lot more in number, than the comments of different types. Although fatigue affects raters' frequency of comments on grammar, choice of words, and organization, it does not affect raters' comments on punctuation, dictation, and capitalization. This, further, raises the question why Iranian raters are not that severe when errors of the choice of words, punctuation, dictation, capitalization, and organization come into play. The question is beyond the scope of this paper and would be suggested for further research.

In summary, test bias is caused by a number of factors (e.g., those related to test method facet, raters/test-takers' educational/language background, raters/test-takers' fatigue, etc.). Commenting on and scoring a great number of writing tasks is a demanding task which causes fatigue, and as a result, a considerable decrease in the frequency of raters' comments. With fewer comments on the writing tasks, due to fatigue, raters' judgment in assigning a score can be negatively affected and test results would be endangered. This study argues that fatigue significantly affects EFL raters' judgments in commenting on and, as a result, scoring writing tasks which results in introducing construct irrelevant factors to test results and interpretations which can end in test bias.

**REFERENCES**


Anastasi, A. (1979). *Fields of applied psychology* (2nd Ed.). *New York: McGraw Hill*.

Bendig, A. W. (1955). Rater reliability and "judgmental fatigue. *Journal of Applied Psychology, 39*(6), 451-454.

Brown, J. D. (1991). Do English ESL faculties rate writing samples differently*? TESOL Quarterly,* 25(4), 578-603

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12, 1–15.

Brown, R., & Richards, L. (2011). IELTS Advantage Writing Skill, A step-by-step guide to a high IELTS score. *Surrey: DELTA Publishing.*

Caban, H. L. (2003). Rater bias in the speaking assessment of four L1 Japanese ESL. *Second Language Studies*, *21*(2), 1–44.

Celce-Murcia, M., Brinton, M., Snow, M. (2014). Teaching English as a second or foreign language. In M. CelceMurcia, D. M. Brinton, & M. A. Snow (Eds.), Teaching English as a second or foreign language (pp. 222–237). Boston, MA: National Geographic.

Constable, E., & Andrich, D. (1984). Inter-Judge Reliability: Is Complete Agreement among Judges the Ideal? *ERIC.*

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Cumming, A. (2011). ESL/EFL instructors' practices for writing assessment: Specific purposes or general purposes? *Language Testing, 18*(2), 207–224.

Cumming, A., Kantor, R., & Powers, E. D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*, 67–96.

Cummings, S. T. (1954). The clinician as judge: Judgments of adjustment from Rorschach single card performance. *Journal of Consulting Psychology*, 18, 243–247.

Drave, N. (2011). *Marker 'fatigue' and marking reliability in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE)*. Paper presented at IAEA 2011. Retrieved from *http://www.iaea.info/documents/paper_30171b739.pdf*

Goodall, G. (2011). Syntactic satiation and the inversion effect in English and Spanish wh-questions. *Syntax, 14*(1), 29–47.

Guangming, L., Mollaun, P. & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499

Geacintov, T., & Peavler, W. S. (1974). Pupillography in industrial fatigue assessment. *Journal of Applied Psychology*, *59*, 213–216.

Hall, C., & Sheyholislami, J. (2013). Using appraisal theory to understand rater values: An examination of rater comments on ESL test essays. *The Journal of Writing Assessment, 6*(1), *pp.* Accessed 7th September 2016 from: http://journalofwritingassessment.org/article.php?article=66

Hiramatsu, K. (2000). *Accessing linguistic competence: Evidence from children's and adults' acceptability judgments.* Connecticut: University of Connecticut.

Johnson, R., Penny, J., & Gordon, B. (2001). Score resolution and the inter-rater reliability of holistic scores in rating essays. *Written Communication,* 18, 229–249.

Ling, G, Mollaun, P, & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking response. *Language Testing, Sage Publications*.

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H.-J., & Burton, N. (2004). *A study of fatigue effects from the New SAT* (ETS Research Report No. RR-04–46). New York: College Board.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 54-71.

Massey, A. J. (1977). Candidate fatigue and performance on GCE objective tests. *British Journal of Educational Psychology*, *47*(2), 203–208.

McCormick, E. J. (1970). *Human factors engineering* (3rd Ed.). New York: McGraw-Hill

McNamara, T. F. & Wesley, A. (1996). *Measuring second language performance*. London, UK: Longman Publications

Schumm, J. S., & Vaughn, S. (1991). Making Adaptations for Mainstreamed Students: General Classroom Teachers' Perspectives. *Remedial and Special Education, 12*(4), 18-27.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*, 27–33.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, 575–582.

Sprouse, J. (2007). Revisiting satiation. (Unpublished manuscript). *Retrieved from www.socsci.uci.*

Theander, K. (2007). Fatigue, functional status, health and pulmonary rehabilitation in patients with chronic obstructive pulmonary disease (*Linköping University Medical Dissertations No. 980). Retrieved from http://www.diva-portal.org/smash/record.jsf;jsessionid=22e91b72ab7c 6616422ab29729b8?parentRecord=diva2:139577&pid=diva2:23117*

Tucker, L. R. (1948). Memorandum concerning study of effects of fatigue on afternoon achievement scores due to Scholastic Aptitude Test being taken in the morning. *ETS Research Memorandum No. RM-48–2*.

Weigle, SC (1994). Effects of training on raters of ESL compositions. *Language Testing, 1994 - ltj.sagepub.com*

Wohlhueter, J. F. (1966). Fatigue in testing and other mental tasks: A literature survey. *ETS Research Memorandum.*

**Appendix 1. Descriptive Statistics for the Frequency of Comments on Grammar**

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 16 | 12.6250 | 3.70360 | .92590 | 10.6515 | 14.5985 | 8.00 | 19.00 |
| 2 | 16 | 13.5000 | 3.07679 | .76920 | 11.8605 | 15.1395 | 6.00 | 18.00 |
| 3 | 16 | 10.6875 | 2.77414 | .69353 | 9.2093 | 12.1657 | 6.00 | 16.00 |
| 4 | 16 | 10.6875 | 2.38659 | .59665 | 9.4158 | 11.9592 | 6.00 | 15.00 |
| 5 | 16 | 11.3750 | 4.12916 | 1.03229 | 9.1747 | 13.5753 | 8.00 | 25.00 |
| 6 | 16 | 9.5000 | 2.03306 | .50827 | 8.4167 | 10.5833 | 6.00 | 13.00 |
| 7 | 16 | 10.1875 | 1.04682 | .26171 | 9.6297 | 10.7453 | 9.00 | 13.00 |
| Total | 112 | 11.2232 | 3.10684 | .29357 | 10.6415 | 11.8049 | 6.00 | 25.00 |

**Appendix 2. Descriptive Statistics for Comments on Choice of Words**

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 16 | 1.9375 | 1.65202 | .41300 | 1.0572 | 2.8178 | .00 | 5.00 |
| 2 | 16 | 1.3125 | .94648 | .23662 | .8082 | 1.8168 | .00 | 3.00 |
| 3 | 16 | 1.5000 | 1.21106 | .30277 | .8547 | 2.1453 | .00 | 3.00 |
| 4 | 16 | 1.3125 | .94648 | .23662 | .8082 | 1.8168 | .00 | 3.00 |
| 5 | 16 | .8750 | .88506 | .22127 | .4034 | 1.3466 | .00 | 3.00 |
| 6 | 16 | 1.0625 | 1.18145 | .29536 | .4329 | 1.6921 | .00 | 4.00 |
| 7 | 16 | .6875 | .70415 | .17604 | .3123 | 1.0627 | .00 | 2.00 |
| Total | 112 | 1.2411 | 1.14880 | .10855 | 1.0260 | 1.4562 | .00 | 5.00 |

**Appendix 3. Descriptive Statistics for Comments on Punctuation**

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 16 | 1.6875 | 1.30224 | .32556 | .9936 | 2.3814 | .00 | 4.00 |
| 2 | 16 | .9375 | 1.06262 | .26566 | .3713 | 1.5037 | .00 | 3.00 |
| 3 | 16 | 1.6875 | 1.30224 | .32556 | .9936 | 2.3814 | .00 | 5.00 |
| 4 | 16 | 1.1875 | 1.10868 | .27717 | .5967 | 1.7783 | .00 | 4.00 |
| 5 | 16 | 1.8750 | 1.02470 | .25617 | 1.3290 | 2.4210 | 1.00 | 4.00 |
| 6 | 16 | 2.0625 | 1.48183 | .37046 | 1.2729 | 2.8521 | .00 | 4.00 |
| 7 | 16 | .9375 | .85391 | .21348 | .4825 | 1.3925 | .00 | 2.00 |
| Total | 112 | 1.4821 | 1.22277 | .11554 | 1.2532 | 1.7111 | .00 | 5.00 |

**Appendix 4. Descriptive Statistics for Comments on Organization**

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 16 | 2.2500 | 1.34164 | .33541 | 1.5351 | 2.9649 | .00 | 5.00 |
| 2 | 16 | .7500 | .93095 | .23274 | .2539 | 1.2461 | .00 | 3.00 |
| 3 | 16 | 1.5000 | 1.21106 | .30277 | .8547 | 2.1453 | .00 | 3.00 |
| 4 | 16 | 1.6875 | 1.07819 | .26955 | 1.1130 | 2.2620 | .00 | 4.00 |
| 5 | 16 | .8750 | 1.25831 | .31458 | .2045 | 1.5455 | .00 | 3.00 |
| 6 | 16 | 1.4375 | 1.15289 | .28822 | .8232 | 2.0518 | .00 | 4.00 |
| 7 | 16 | .6250 | .88506 | .22127 | .1534 | 1.0966 | .00 | 3.00 |
| Total | 112 | 1.3036 | 1.22907 | .11614 | 1.0734 | 1.5337 | .00 | 5.00 |

**Appendix 5. Descriptive Statistics for Comments on Dictation**

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 16 | 1.5000 | 1.36626 | .34157 | .7720 | 2.2280 | .00 | 4.00 |
| 2 | 16 | .8125 | .91059 | .22765 | .3273 | 1.2977 | .00 | 3.00 |
| 3 | 16 | 1.5000 | .96609 | .24152 | .9852 | 2.0148 | .00 | 3.00 |
| 4 | 16 | 1.7500 | 1.00000 | .25000 | 1.2171 | 2.2829 | .00 | 3.00 |
| 5 | 16 | 1.1250 | .88506 | .22127 | .6534 | 1.5966 | .00 | 3.00 |
| 6 | 16 | .6250 | .61914 | .15478 | .2951 | .9549 | .00 | 2.00 |
| 7 | 16 | .8750 | .71880 | .17970 | .4920 | 1.2580 | .00 | 2.00 |
| Total | 112 | 1.1696 | 1.00349 | .09482 | .9817 | 1.3575 | .00 | 4.00 |

**Appendix 6. Descriptive Statistics for Comments on Capitalization**

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 16 | 1.5000 | 1.36626 | .34157 | .7720 | 2.2280 | .00 | 4.00 |
| 2 | 16 | .8125 | .91059 | .22765 | .3273 | 1.2977 | .00 | 3.00 |
| 3 | 16 | 1.5000 | .96609 | .24152 | .9852 | 2.0148 | .00 | 3.00 |
| 4 | 16 | 1.7500 | 1.00000 | .25000 | 1.2171 | 2.2829 | .00 | 3.00 |
| 5 | 16 | 1.1250 | .88506 | .22127 | .6534 | 1.5966 | .00 | 3.00 |
| 6 | 16 | .6250 | .61914 | .15478 | .2951 | .9549 | .00 | 2.00 |
| 7 | 16 | .8750 | .71880 | .17970 | .4920 | 1.2580 | .00 | 2.00 |
| Total | 112 | 1.1696 | 1.00349 | .09482 | .9817 | 1.3575 | .00 | 4.00 |

## Appendix 7. Interview Questions

1- How did you physically feel during scoring the essays?
2- Have you experienced fatigue during and/or after the scoring procedure? If yes, what were the symptoms? And when was it at its highest level (in the beginning, in the middle, or toward the end of the scoring procedure)?
3- Which one/any number of the following items are among the symptoms of fatigue? (lack of concentration, sleepiness, dizziness, pain, unwillingness to give more comments)
4- In which, if any, parts of the body did you feel pain?
5- What do you think the mentioned symptoms can be attributed to?
6- Do you think scoring essays for long hours can cause the mentioned symptoms?
7- Do you think having breaks during scoring would help improve your quality of scoring?
8- Do think your judgment during scoring the essays was affected by fatigue?