INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 10 No. 1, May 2020, pp. 1-13

Available online at:
https://ejournal.upi.edu/index.php/IJAL/article/view/24956

https://doi.org/10.17509/ijal.v10i1.24956

The effect of raters fatigue on scoring EFL writing tasks

Amir Mahshanian* and Mohammadtaghi Shahnazari
Department of English Language and Literature, Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran

ABSTRACT

Given the importance of testing, in general, and scoring writing tasks in particular, the negative effect of fatigue on human raters is important to investigate. This study aimed to (1) explore the relationship between fatigue and scoring composition tasks written by upper-intermediate EFL learners; and (2) to investigate the discrepancy of the frequency of comments among EFL raters while scoring composition tasks. Four raters were selected, and each given 28 composition tasks to score and comment on. The data were analyzed through SPSS software by running ANOVA, Pearson correlation coefficient, and post-hoc tests. Results suggested that the scores assigned to the first 16 tasks were significantly lower than those assigned to the last 12 tasks and that the last four tasks were scored highest. Based on the results obtained from the questionnaire, the observed diversity is argued to be rooted in raters' fatigue and result in test bias. Furthermore, findings indicated that the frequency of comments given by the raters on the first 12 essays was significantly higher than those on the last 16 essays (the highest and the lowest frequency of comments were observed in the first four, and the last four scored essays, respectively).

Keywords: Assessing writing; EFL writing; fatigue; rater consistency; reliability; scoring composition tasks

First Received:	**Revised:**	**Accepted:**
20 January 2020	24 March 2020	28 April 2020
Final Proof Received:		**Published:**
26 May 2020		31 May 2020

How to cite (in APA style):
Mahshanian, A. & Shahnazari, M. (2020). The effect of raters fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics, 10*(1), 1-13. https://doi.org/10.17509/ijal.v10i1.24956

INTRODUCTION

Assessing writing

One of the core elements of any curriculum is evaluation, in general, and scoring language tests of different types, in particular. The quality of scoring is an essential part of evaluation in that it can affect the validity and fairness of language tests (Ling et al., 2014). Language test scores are used to infer learners' language ability and under different circumstances, effects due to sequencing, timing, and fatigue may introduce inconsistency into the way the rating criteria are applied (Bachman, 2004). In scoring essays, for example, a rater may start paying little attention to grammar, focusing mainly on cohesion, organization, and content; however, if raters encounter essays with numerous grammatical errors, they may unconsciously begin paying more attention to those errors (Bachman, 1990).

The most widely accepted method for scoring EFL writing in composition tests is conducted by a process of analytic/holistic scoring by at least one rater. Ghalib and Al-Hattami (2015) define holistic assessment in EFL writing as assigning an overall score to the entire constructed response, and in line with Cumming (1990), Hamp-Lyons (1995), and Reid (1993), believe that its "cost-effectiveness" (p. 226) makes it an appropriate method for assessing performance in large-scale writing tests (e.g., TOEFL, GRE, GMAT, etc.), and thus very practical. Also, in describing the positives of applying holistic rubrics in scoring EFL writing, Ghalib and Al-Hattami (2015) refer us to Weigle (2002) who asserts that holistic assessment is "short, do[es] not include detailed criteria of evaluation, and make[s] possible the evaluation of an essay by assigning one score to it after only one reading" (Ghalib & Al-Hattami,

* Corresponding Author
Email: mshn_amir@yahoo.com

2015, p. 227). Furthermore, Nakamura (2004), in an attempt to compare holistic and analytic assessment of EFL writing considers the former as more cost-benefit and a much faster procedure.

On the other hand, analytic scoring -assigning different scores to different aspects of writing- provides an in-depth examination of the writer's performance on EFL writing tasks in that it includes issues related to "the test taker's lexical, syntactic, discourse, and rhetorical competence" (Ghalib & Al-Hattami, 2015, p. 227). As regards applying analytic scoring rubrics, Hamp-Lyons (1995) asserts that these scoring rubrics present EFL teachers with comprehensive feedback and help them capitalize on the learners' writing strengths and/or weaknesses. Confirming this, Becker (2011) also adds that analytic scores determine where to add more instruction in EFL writing courses (Ghalib & Al-Hattami, 2015).

In general, there is census among researchers that with careful monitoring and training of raters, scoring procedure of these types can lead to results which are to some extent reliable (McNamara, 1996; Weigle, 2002). However, these rating procedures have been attacked for simplifying the constructs they are to demonstrate. For example, Cumming et al. (2002) maintained:

> "Holistic rating scales can conflate many of the complex traits and variables that human judges of students' written composition perceive (such as fine points of discourse coherence, grammar, lexical usage, or presentation of ideas) into a few simple scale points, rendering the meaning or significance of the judges' assessments in a form that many feel is either superficial or difficult to interpret" (p. 68).

Since these scoring procedures encompass a huge area of testing, it is essential to investigate values in decision making and behaviors of raters since they have at their heart the scoring criteria when rating composition tests, and to explore how decision making values are conducive to the test construct definition (Hall & Sheyholislami, 2013).

As McNamara (1996) maintained, assessing performance "necessarily involves subjective judgments" (p. 117), which lead to variability, or issues regarding inter-rater reliability of the test scores. Raters might be different in scoring for different reasons, such as different styles of rating (Charney, 1984), severity or overall rater leniency bias against/towards (a) particular group(s) of participants or the type of tasks, types of scoring procedures and scoring criteria (Barkaoui, 2007; Schoonen, 2005), variety in the interpretation of rating criteria whether raters' comments are focused on (a) specific part(s) of the text (Huot, 1993), the absence or existence of training and the effects of different training types (Harsch & Rupp, 2011; Huot, 1993; McNamara, 1996; Vaughn et al., 1993; Weigle, 1994).

Studies on rating process and the way raters exploit scoring criteria suggested that assessing essays is a repetitive operation (Freedman & Calfee, 1983) which involves self-monitoring (Cumming et al., 2002). Raters, too, are extensively engaged in a problem-solving process when it comes to decision making about the scores comparing with the time when they simply match rating criteria to the related aspects of tests (Cumming, 1990; DeRemer, 1998). In general, they focus on the discriminating features of a text, consider task requirements and the text audience (Freedman & Calfee, 1983), then attribute more points to different features of the text (Eckes, 2008; Vaughn et al., 1993). The nature of these judgements, decisions and self-feedback, in addition to the significance of this self-monitoring have not been brought into attention and are usually manifested in written comments and in the scoring outcomes of raters (Hall & Sheyholislami, 2013).

Weigle (1998) maintained "it is not enough to be able to assign a more accurate number to examinee performances unless we can be sure that the number represents a more accurate definition of the ability being tested" (p. 281). Therefore, it is necessary in test validation to collect evidence supporting an aptly-defined construct. The raters of a test make a great contribution to the definition of its construct in the sense that they elucidate criteria of rating, that for some tests are taken as the straightest and clearest definition of that construct. Raters' comments as judges, and authorities, their comprehension of the language, and their various biases are conducive to the reinforcement of the values determined in a test (Hall & Sheyholislami, 2013).

Raters' fatigue

In the literature, various definitions could be found for the concept of fatigue. Anastasi (1979), for example, defined fatigue as feelings of tiredness as well as qualitative and quantitative output reduction which, according to Ling et al. (2014), leads to increase in time of response and in the frequency of errors. That is, fatigue can cause raters to invest a lot of time and energy on rating and to make more errors in the process of rating. Cummings (1954), in another perspective, defined fatigue in terms of mental or physical signs (e.g. tension of muscles, tiredness feelings, drowsiness, sleepiness, lack of concentration, etc.). Ling et al., (2014), in line with Cummings (1954), maintain that these signs are more subtle than the output indicators in that they provide researchers with more space for error recognition.

In the past few decades, a large body of research has discussed the effect of fatigue on both test takers and testers (e.g., Bendig, 1955; Constable & Andrich, 1984; Cumming, 1990; Cumming et al., 2002; Drave, 2011; Ling et al., 2014; Lumley & McNamara, 1995; Mahshanian, et al., 2017; Massey, 1977; Schumm & Vaughn, 1991), among which very conflicting findings were observed.

Some scholars (e.g., Bendig, 1955; Cummings, 1954; Drave, 2011; Liu, et al., 2004; Massey, 1977; Tucker, 1948; Wohlhueter, 1966) believe that fatigue does not affect test-takers' scores and/or test-givers' judgment significantly. For example, Bendig (1955), in an attempt to investigate judgmental fatigue among test-takers, divided his subjects into 6 groups and asked them about their preferences according to "a nine-point scale" (p. 453). Results of his study indicated that "judgmental fatigue does not affect rater reliability" (p. 453). The major shortcoming, however, as Ling et al. (2014) rightly pointed out, was that tasks in his study were so simple (i.e. requiring a low level of cognitive ability and minimum level of attention) to be a concise representation of the effects of fatigue. Also, they lasted for only 20 minutes which is a short time during which the real effects of fatigue would neither be observable nor measurable. Another issue was that Bendig (1955) refers to college students (not EFL raters, markers, or testers) as raters. Thus, by concluding that fatigue cannot affect raters' judgements, he does not, by any means, hypothesize that it cannot affect raters of any type (e.g., EFL writing raters), or cannot affect them on different conditions (e.g., with different tasks, time limitations, etc.).

In another study on more than 80 advanced test-takers, Massey (1977) investigated how fatigue can affect test-takers' performance on GCE objective tests and argued that there is "no evidence of differences in the proportions of students choosing the correct response which might be attributed to the effects of candidate fatigue" (p. 203). Although admitting that "performance towards the end of the test may be inhibited by the onset of fatigue" (p. 203) in subjective as well as objective tests, Massey (1977) could not find any evidence to show a decline in performance of his participants at the end of the tests. While capitalizing on the effect of fatigue, Massey's argument (1977) was not conclusive in that it had its focus only on objective tests, on test-takers (but not raters), and in the sense that subjects were tested in a 1-hour period, which is, as for Bendig (1955), relatively short for the effects of fatigue to be significant.

Among very few studies examining the effect of fatigue on raters, Drave (2011) found that despite the fact that raters "indeed suffer from fatigue [based on the results obtained from the questionnaire survey], there is no evidence in [the] data [indicating] fatigue effects" (p.7). In his study, raters were asked to give a score of between 1 and 5 (5 being the highest score) to some 400-word essays using "onscreen marking (OSM), a system in which marking is done on computers" (Drave, 2011, p.1). Although Drave's (2011) study investigated the ratings of 3 raters over 4 hours, which was considered a long period compared to studies conducted by Cummings (1954), Massey (1977), and Bendig (1955), the fact that raters used (OSM), and that the only task of raters was to assign numbers to the essays, adds limitations of its generalizability to other types of scoring (e.g., paper-based assessment of essays) which include more demanding tasks (e.g., raters' comments on each scored essay).

Contrary to the above, some other studies have shown that the reliability and consistency of language tests can negatively be affected by fatigue (e.g., Erguvan & Aksu Dunya, 2020; Goodall, 2011; Hiramatsu, 2000; Ling et al., 2014; Mahshanian, et al., 2017; Sprouse, 2007; Wohlhueter, 1966). Among them, some capitalized on the effects of "judgement fatigue" (Ling et al., 2014, p. 481), "linguistic disease" or "syntactic satiation" (Snyder, 2000, p. 575), a phenomenon through which "some unacceptable sentences begin to sound more acceptable after days or weeks of repeatedly judging their acceptability" (Sprouse, 2007, p. 329). Snyder (2000), for example, asserted that although fatigue affects grammaticality judgement, it does not affect all types of sentences in the same fashion (Snyder, 2000). It should be added, despite the fact that subjects in his study were requested to judge the grammaticality of a series of 58 sentences (a highly cognitively demanding task), they were asked to do so by only providing a simple yes-or-no response (a relatively simple productive task). Also, the results of his study, along with those replicated by Hiramatsu (2000), Sprouse (2007), and Goodall (2011), are not analogous to rating/scoring essays in EFL contexts in that judgment in EFL rating/scoring is not limited to only that of grammaticality, nor is as simple as yes-or-no comments.

Among studies asserting that fatigue negatively affects raters' judgement, very few can be found with the focus on EFL contexts. For example, Ling et al., (2014) compared the quality of raters' scoring TOEFL iBT speaking tasks under different shift conditions and found varying levels of "rating accuracy and consistency across shift conditions" (p. 479) due to the negative effect of fatigue. They argued that the raters who suffer from the effect of fatigue (those scoring the responses in longer shifts) have lower "rating productivity, accuracy, and consistency" (p. 479). It should be pointed out, however, that since the raters in their study were to assign numbers to 14,000 audio responses to four TOEFL iBT speaking tasks in 2/4-hour shifts, and as they were concerned with recorded constructed responses, not written ones, their findings fail to address and/or be generalizable to paper-based assessment of essays (i.e., those including raters' comments).

More recently, Mahshanian et al. (2017) investigated how raters' comments on EFL writing tasks can be affected by fatigue and concluded that "fatigue brings about changes in the way raters comment on essays from the first to the last few ones" (p.310). Their study did not take into account the

scores assigned to the tasks and only focused on the comments given by the raters.

In a most recent study, Erguvan and Aksu Dunya (2020) analyzed rater severity among EFL raters while assigning scores to compositions written by freshmen Kuwaiti learners. Employing many facet Rasch model, they concluded that despite being consistent regarding applying rubrics, raters varied in terms of leniency and severity and the major reason for the observed inconsistencies in ratings were attributed to 1) misunderstanding the scale categories and more importantly 2) "fatigue toward the end of performance" (Erguvan & Aksu Dunya, 2020, p. 11). Despite reporting on inconsistencies among raters, this study did not put forth a standard time interval for assessing each written composition.

To recapitulate, scoring writing tasks places unavoidable burdens on human raters' cognitive processing ability and concentration and hence results in fatigue which can endanger judgment in general, and consistency and accuracy of the scoring process, in particular (Ling et al., 2014). While some studies investigated the impact of fatigue on tests-takers performance on language tests (e.g., Bendig, 1955; Cummings, 1954; Massey, 1977; Tucker, 1948), few can be found to examine such an effect on raters' quality of scoring constructed responses, such as writing or speaking (e.g., Drave, 2011; Ling et al., 2014). It should be pointed out, however, that very incompatible results came out of these studies in that some highlighted the effect of fatigue on human judgment in language tests (e.g., Bendig, 1955; Cummings, 1954; Drave, 2011; Liu et al., 2004; Massey, 1977; Tucker, 1948; Wohlhueter, 1966), while others deemphasized such an impact, asserting that fatigue does not affect test-takers', or raters' judgement significantly (e.g., Goodall 2011; Hiramatsu, 2000; Sprouse, 2007; Wohlhueter, 1966).

Unlike studies basing their methods on simple tasks (e.g., asking students about food preferences, or requesting them to provide yes-or-no responses to judge grammaticality, as in Bendig, 1955, and Snyder, 2000, respectively), and those in which the effect of fatigue was examined in relatively short periods of time, (e.g., Cummings, 1954; Bendig, 1955), 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. It should be added that, contrary to very few studies being conducted on fatigue's effects on raters while scoring constructed responses such as writing (e.g., Drave 2011) with its main stress on on-screen-marking (OSM), and those on the constructed responses such as speaking (e.g., Ling et al., 2014), the current study has its major focus on paper-based (not OSM) EFL constructed responses, i.e., writing tasks.

Thus, this study, in an attempt to fill the existing lack of sufficient research on the effect of fatigue on raters scoring quality, and due to very conflicting findings in this area, aims at inspecting such negative impacts on raters when they score EFL writing tasks. More specifically, the following research questions are investigated in this study:

1. Does fatigue affect scoring EFL writing tasks significantly?
2. Does the frequency of comments in various scored essays change due to the raters' fatigue?

METHOD

Design

The present study employs an ex-post-facto design and aims at exploring how raters' fatigue relates to the frequency of their comments on composition tasks written by upper-intermediate EFL learners, and whether it makes any difference in writers' scores given by the raters. There was a total of 28 essays to be scored and commented on by four Iranian EFL raters.

Participants

The participants of this study consisted of four raters who were selected from among EFL instructors in two language institutes. Raters were selected from among the most experienced instructors with more than 8 years of EFL teaching and rating experience. The raters' groups were shown in Table 1.

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

Before the scoring procedure, 28 upper intermediate EFL students were given the task of writing a five-paragraph opinion essay on a given topic. Thus, the total number of EFL learners contributed to this study was 28, and the total number of EFL raters doing so was 4.

Instruments

Instructional material

IELTS Advantage Writing Skills (Brown & Richards, 2011), containing 10 units, was used as the source to teach learners how to write an opinion essay. In the interest of time, and for the purpose of this study, only one unit of this book (the 3rd unit)

was taught to the participants of this study in a three-hour session before they were asked to write the final essays. The book teaches learners the most important issues regarding how to develop paragraphs and organize opinions. Also, it presents learners with samples of each type of essay. It should be added that the instructors were trained to teach the learners the most relevant issues (i.e., those in keeping with the research requirements). With respect to training raters, a scoring booklet elaborating on the scoring process and presenting sample scored compositions (adopted from Brown, 1991) was given and later in briefing sessions (see Appendix 1) explained to the raters. Based on these scoring procedures, while scoring the essays and commenting on them, raters were asked to take into a consideration six writing features as suggested by Brown (1991) (i.e., cohesion, content, mechanics, organization, syntax, and vocabulary).

Testing material

A random IELTS topic as a writing task was given to the participants (students) to complete in a 1-hour period. This contained a topic, and a task according to which learners were supposed to write a five-paragraph opinion essay regarding the differences between homeschooling and going to school.

Procedure

To make sure that the possible differences in scores and the frequency of comments given by raters are due to fatigue and not to other factors such as differences in actual writing quality, learners needed to be homogenized. In so doing, before conducting the study, 60 EFL learners from two language institutes, were selected from upper intermediate classes to write an essay on a certain topic. 6 raters, afterwards, were requested to score and comment on the essays. After scoring 60 essays, all raters, based on learners' overall writing proficiency, agreed that EFL learners in these institutes were not of the same or even approximate writing proficiency level in that some are, to some extent, qualified writers whereas some have not yet learnt the fundamentals of writing. Thus, to make sure the learners are of the approximately same writing proficiency, 40 learners, who were scored almost the same were selected and asked to write a five-paragraph essay on a different topic for the second time.

Notwithstanding the careful selection, still huge discrepancies among learners regarding their writing proficiency were observed. Accordingly, the pilot study was repeated a third time in which 12 learners were excluded from the study since, according to learners' scores and raters' judgments, they were not suitable for the purpose of this study due to their level of writing proficiency. Finally, 28 EFL learners from among 60 learners, quite selectively, were chosen and each given a new topic to write an essay based on (i.e., an opinion essay on the differences between "homeschooling and going to school"). It should also be highlighted that in an act of motivating learners to take the tasks seriously, instructors of the courses in the mentioned institutes were asked to assign the tasks as the complementary part of the course without which learners would lose marks, and probably fail.

In addition to the learners, raters needed to be homogeneous. To take homogeneity of the raters into consideration, four (from among six) raters were selected to take part in this study. In the pilot studies, the means of scores given to the tasks by the selected raters (all with more than eight years of EFL teaching and rating experience and all male), and the means of frequency of their comments on scored essays were almost the same (i.e. they were not significantly different among raters).

After being homogenized, in briefing sessions before scoring the essays, raters were asked to take certain rubrics (i.e., those suggested by Brown, 1991) into consideration while scoring the essays and were presented with sample scored essays (those including raters' comments) to have an overall understanding of the scoring process. As mentioned earlier, and as regards raters' comments on essays, six features (such as cohesion, content, mechanics, organization, syntax, and vocabulary) were into focus. While reading through the essays, raters were asked to underline any part they thought needed positive/negative feedback (feedback on the weaknesses and strengths of the writing) and provide the writer with their opinion on that part clarifying why they believed it was/was not appropriate regarding the mentioned features.

Also, the selected raters were asked to holistically score and comment on the essays without any break intervals during the scoring procedure which approximately took three hours. The comments, both on the content of the essay and on linguistic issues, intended to provide feedback to the writers, and to justify the holistic scores that were assigned.

Furthermore, to make sure that the possible differences in scores and the frequency of comments are due to fatigue and not to other factors such as the order that the particular essays happened to be presented in, raters were requested to score essays in an opposite order from one another. That is, rater1 scored essays from no.1 to no 28, whereas rater 2 scored essays from no.28 to no.1 (in an opposite order). This was also the case for raters 3 and 4. It is worth mentioning that all raters were observed during the scoring procedure. Moreover, to give raters a motivation to score the essays accurately, they were promised to have a raise in their payment, an option for choosing the classes to teach for the next term, and a gift card, in case of accurate scoring.

Finally, in retrospective interviews (based on the questionnaire developed by Drave, 2011), raters were asked whether they had experienced fatigue and how its effects were manifested (see section 3.1.2 for

the results). Further, during the interviews (lasting for 20-30 minutes), detailed notes were taken from the raters' responses and all four interviews were audio-recorded, transcribed and then reviewed by the authors. It should be noted that confidentiality of the interviews and anonymity of the interviewees (i.e., raters) were promised before the interviews. Each interviewee (rater) was interviewed separately in his mother tongue and the following open-ended and yes-no questions (in addition to some related follow-up questions), were posed.

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 judgement during scoring the essays was affected by fatigue?
9. How fast did you assign scores to the first few essays you scored?
10. How fast did you assign scores to the last few essays you scored?
11. Do you think, due to the effect of fatigue, your scoring became slower by the passage of time?
12. Do you think, due to the effect of fatigue, and by the passage of time, you became more lenient as far as assigning scores were concerned?

FINDINGS

Data analysis

Does fatigue affect scoring EFL writing tasks significantly?

Table 2 shows the mean and standard deviation of the total scores and the frequency of total comments. Scores are given from 0 to 9.

Table 2
Descriptive Statistics for the Mean of Scores and Total Frequency of Comments

	Mean	Std. Deviation	N
Score	5.5804	1.37473	112
Frequency of Total Comments	17.4821	2.97431	112

Also, to have a precise statistical view over the descriptive analysis of the scores, Table 3 is presented. In Table 3, as in other tables in this article, 28 papers were divided to groups of four for analysis. Thus, group 1 represents scores assigned to the first four essays (number 1 to 4), group 2, scores in the second four essays (number 5 to 8), and group 7, the last four essays (number 25 to 28). Since there were four raters as subjects of this study and in each group, they scored four essays, the total number of the essays to be scored in one group is 16 and the total number of all essays in all groups to be compared is 112.

Table 3
Descriptive Statistics for Scores

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1.00	16	4.6875	1.68201	.42050	3.7912	5.5838	1.00	7.50
2.00	16	5.1875	.57373	.14343	4.8818	5.4932	4.00	6.00
3.00	16	4.8438	.78991	.19748	4.4228	5.2647	3.50	6.50
4.00	16	4.9688	.69447	.17362	4.5987	5.3388	4.00	6.50
5.00	16	6.7188	1.04831	.26208	6.1601	7.2774	4.50	8.00
6.00	16	5.6563	1.35054	.33764	4.9366	6.3759	3.00	8.00
7.00	16	7.0000	1.12546	.28137	6.4003	7.5997	5.00	8.50
Total	112	5.5804	1.37473	.12990	5.3230	5.8378	1.00	8.50

To investigate the relationship between the scores assigned to 28 essays, an ANOVA was run and the results are presented in Table 4. As is obvious in Table 4, the p-value is estimated at (0.000). Thus, there is a significant relationship among groups regarding the scores assigned to the essays.

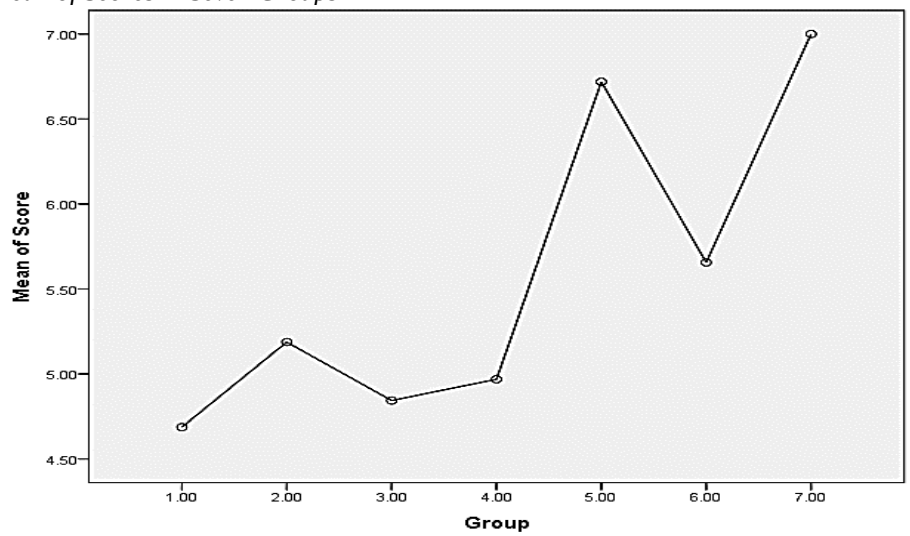
Table 4
ANOVA for Scores

	Sum of Squares	DF	Mean Square	F	Sig.
Between Groups	82.964	6	13.827	11.449	.000
Within Groups	126.813	105	1.208		
Total	209.777	111			

Multiple comparisons of the scores

To compare each of these seven groups with one another (scores assigned to 28 essays), and investigate the relationship between them, a post-hoc LSD test is used based on which some conclusions can be made. First, group 1 does not have a significant relationship with group 2, 3, and 4 (i.e., scores assigned to the first 16 essays were not significantly different). Second, group 1, however, has a significant relationship with group 5, 6, and 7 (i.e., scores assigned to the last 12 essays were significantly different from those assigned to the first 16). This clearly indicates that the effect of fatigue becomes significant after scoring 16 essays or raters' judgment regarding assigning scores to the essays would be affected by fatigue mainly after scoring 16 papers. Figure 1 clearly depicts such an effect.

Figure 1
Mean of Scores in Seven Groups



In Figure 1, the rise in scores assigned by the raters is clearly depicted. The highest and lowest scores are respectively assigned to the 7th and the 1st groups. That is, the first four essays were scored significantly lower than the last four essays. To sum up, the more essays the raters score, fatigue affects them more significantly, and as a result, they assign higher scores to the essays which are scored later.

The relationship between the scores and the frequency of comments

To explore the relationship between the assigned scores to each essay and the frequency of total comments on each essay, the Pearson correlation coefficient has been used. In Table 5, the correlation between the two variables is estimated at (-0.687).

Table 5
Frequency of Total Comments

		Score	Frequency of total comments
Score	Pearson Correlation	1	-.687**
	Sig. (2-tailed)		.000
	N	112	112
Frequency of Total Comments	Pearson Correlation	-.687**	1
	Sig. (2-tailed)	.000	
	N	112	112

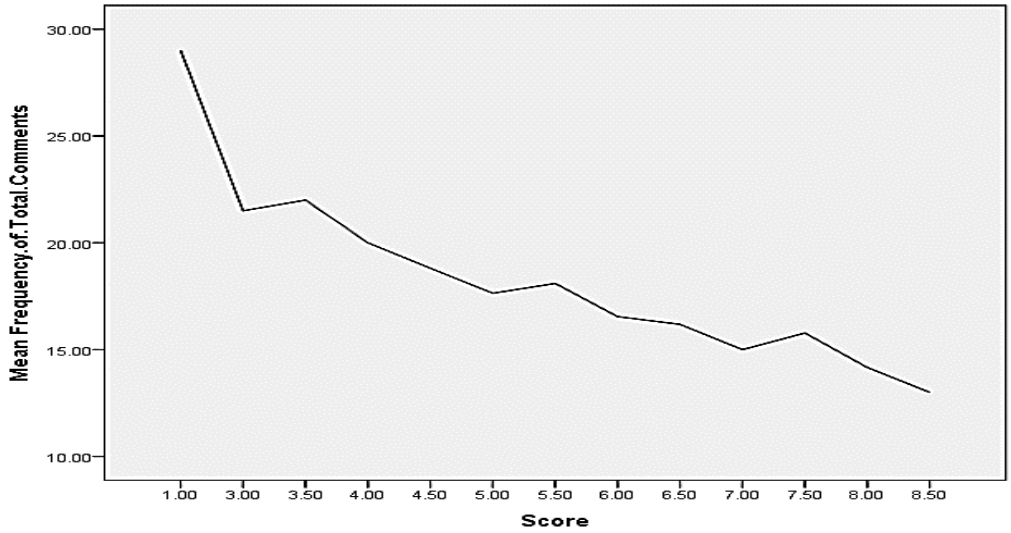
This indicates that there is a negative relationship between the variables. In other words, the more the frequency of the comments, the lower the scores are. With respect to the p-value estimates ($p = 0.000$), this is a meaningful relationship and the correlation indicates that obviously.

As illustrated in Table 5, there is a negative relationship between the frequency of comments and the scores given to the papers by the raters. The

relationship between comments and scores suggests that the raters only gave negative comments, and the scores were mainly based on the number of negative comments. Simply put, the more essays the raters score, the more fatigued they became, and as a result they will have fewer comments on the papers. One might argue, however that there is no theoretical reason why raters became more lenient with fatigue. As discussed earlier introspective interviews with raters, and controlling for other intervening variables, was to make sure that the observed discrepancy of scores and frequency of comments have their roots in fatigue. Thus, as is depicted in Figure 2 below, with fewer comments on the essays, as an effect of fatigue, raters tend to assign higher scores to the last few essays they score.

Figure 2

Negative Relationship between the Frequency of Comments and Scores



Does the frequency of comments change, due to the raters' fatigue?

In Table 6, the descriptive statistics for the total frequency of comments is shown. In seven groups, the mean, standard deviation, standard error of measurement, within 95% confident interval, minimum, and maximum of the data is given.

An ANOVA was also conducted to explore the relationship between the dependent variable, i.e., frequency of total comments, and fatigue. Results are shown in Table 7.

Table 6

Descriptive Statistics for the Total Frequency of Comments

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1.00	16	21.6250	3.64920	.91230	19.6805	23.5695	17.00	29.00
2.00	16	18.2500	1.69312	.42328	17.3478	19.1522	16.00	21.00
3.00	16	18.6250	1.40831	.35208	17.8746	19.3754	16.00	20.00
4.00	16	17.8750	.95743	.23936	17.3648	18.3852	17.00	20.00
5.00	16	16.1250	2.09364	.52341	15.0094	17.2406	12.00	19.00
6.00	16	15.5625	1.50416	.37604	14.7610	16.3640	13.00	18.00
7.00	16	14.3125	1.62147	.40537	13.4485	15.1765	12.00	17.00
Total	112	17.4821	2.97431	.28105	16.9252	18.0391	12.00	29.00

Table 7

ANOVA for the Frequency of Total Comments

	Sum of Squares	DF	Mean Square	F	Sig.
Between Groups	556.589	6	92.765	22.898	.000
Within Groups	425.375	105	4.051		
Total	981.964	111			

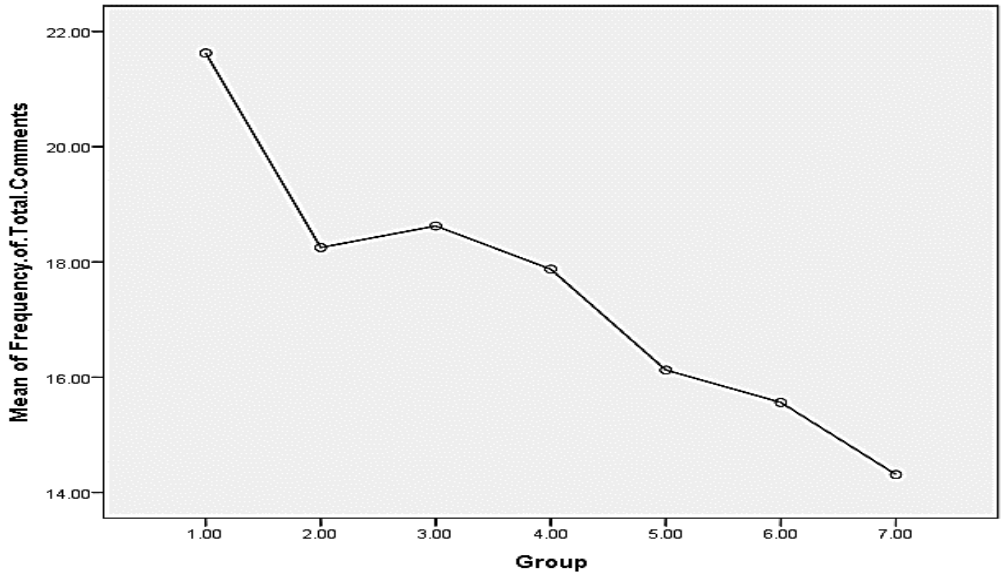
As is clear in Table 7, the amount of p-value is estimated at (0.000), indicating that there is a significant relationship between the frequency of comments on 28 scored essays. As Table 7 clearly shows, the frequency of total comments is highest in the 1st group (first four scored essays), and lowest in the 6th group (essay number 21 to 24).

Multiple comparisons of the frequency of comments

To compare 28 scored essays regarding the frequency of total comments, and find a relationship between them, a post-hoc LSD test was used and the results indicated that there is a significant relationship between the frequencies of total comments among seven groups (28 scored essays). Also, results revealed that there is not any significant relationship between group 1, and group 2 and 3. That is, the frequency of comments on the first 12 essays were not significantly different. However, group 1, has a significant relationship with group 4, 5, 6, and 7 (i.e. the frequency of comments on the last 16 essays were significantly different from those on the first 12). That is, raters' frequency of comment on essays are mainly affected by their fatigue after scoring 12 essays. This clearly indicates that the best time for raters' break, suggested based on these findings, is after scoring 12 essays. Figure 3 below depicts the drop in the frequency of total comments on 28 scored essays.

Figure 3

Means Plot of Frequency of Total Comments



Interviews

As earlier noted, there was a total of four interviews which were all recorded, transcribed, reviewed multiple times by the authors, and finally analyzed using, an emergent, constant-comparative method of grounded interpretation, (adopted from Cumming, 2001). The summary of raters' responses to 12 questions (mentioned in section 2.5) is presented in Table 8.

In the interviews, all four raters declared that they had suffered from fatigue while scoring the tasks. All raters also believed that their tired eyes, hands and neck, their lack of concentration, sleepiness, dizziness, pain in the muscles, unwillingness to write more (give more comments), were among the manifestations of fatigue and attributed these to scoring essays for long hours (3 hours of scoring with no break intervals) with no breaks. Moreover, raters claimed that fatigue had caused them to assign a score to the essays more quickly than normal. It should also be noted that three raters admitted that they were more willing to give higher scores to the last few essays due to fatigue, and one said that he was not sure whether or not fatigue made him more lenient in scoring the last few essays.

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
9	4	0
10	4	0
11	4	0
12	3	1

DISCUSSION

The aim of this study was to investigate the effect of fatigue on human raters. Overall, the study revealed that (a) fatigue negatively affects raters' judgment with regard to marking EFL writing tasks, mainly after scoring 16 essays, and (b) fatigue negatively affects raters' frequency of comments, mainly after scoring 12 EFL writing tasks (essays).

To answer the research questions, a summary of results around each of them is in order. For the first research question about the effect of fatigue on raters while scoring 28 essays, the mean of scores assigned to every four essays were compared in seven groups (i.e. scores assigned to the first four essays, second four essays, and so on). Based on multiple comparisons of the means, it was found that the scores assigned to the first 16 essays were significantly lower than those of the last 12 essays and that the last four essays were scored highest. By controlling for other variables, such findings could be attributed to the negative effect of fatigue.

Furthermore, to add empirical rationale to suggest that raters become more lenient as a result of fatigue, they were interviewed and asked about the reasons for the observed discrepancy among the scores. In the interviews, all four raters admitted that they had suffered from fatigue while scoring the tasks, three stated that they had been more lenient in scoring the last few essays due to the effect of fatigue, and only one rater had doubts about such an effect on his willingness to assign higher scores to the last few essays. Accordingly, findings suggested that scoring more than 16 essays causes fatigue and that fatigue makes raters more lenient in assigning scores to the last few essays. This is in line with the study conducted by Ling et al. (2014). Ling and colleagues argue that suggest time-related variations may end in discrepancies in scoring; hence some shifts and some conditions are more appropriate for raters while scoring.

It should be added that findings about the first research question are in contrast with a few previous studies (e.g., Cummings, 1954; Drave, 2011; Liu, et al., 2004; Massey, 1977; Tucker, 1948). Drave (2011), for example, reported no evidence of the impact of fatigue on human raters. This is different, however, from the present study in that in Drave's (2011) study, raters used (OSM) for assigning scores. Also, as mentioned before, there were no rubrics for assigning scores, nor were there any comments on the scored tasks.

For the second research question about whether the frequency of comments given by the raters change due to the effect of fatigue, it was found that there is a significant relationship between the frequency of comments in the 28 scored essays in that the frequency of comments given by the raters in the first 12 essays was significantly higher than those of the last 16 essays. Also, the highest and the lowest frequency of comments were observed in the first four and the last four scored essays, respectively. As for the first research question, in the interviews, raters were asked about the observed discrepancy in the frequency of their comments. All raters attributed their unwillingness to give more comments (and assigning scores faster than normal) to fatigue and non-stop scoring for long hours. Thus, findings suggested that the frequency of raters' comments on essays decreases by the passage of time and as a result of fatigue. This is in line with Mahshanian and colleagues' (2017) study which holds that fatigue affects raters' frequency of comments on grammar, choice of words, and organization.

Another possible explanation for the observed discrepancy of scores and the frequency comments rests on the relationship between them. Interestingly, findings also showed that there is a negative relationship between the frequency of comments and the scores. That is, the more the frequency of the comments, the lower the scores were. This implies that the raters mostly gave negative comments (i.e., the feedback on how to improve the writing or errors observed rather than positive feedback on the strengths of the writing), and the scores were mainly based on the frequency of the negative comments. That is to say, the lower scores assigned to the tasks resulted from the total number of comments. In other words, by the passage of time and after scoring 12 essays, fatigue significantly affected the raters and caused them to provide fewer comments on the essays and as a result become more lenient and assign higher scores to the last few essays. The fact which was also admitted by the raters in the interviews.

It should be noted, however, that the analysis presented in this study does not show us precisely the way a textual feature influence scoring judgements. Although certain rubrics (see Appendix 1) were followed, the complete context (e.g., cognitive processes raters experienced while scoring the tasks, their attitudes towards rating and scoring, their conditions before and after the scoring session, etc.) where raters provided the essay writers with their comments and the way the comments were exploited to assign a score are not known for certain. Thus, interviews, as elaborated on earlier, were included in order to shed more light on the issue of raters' judgments and decisions. Moreover, the analysis presented in this study, overlooks the importance of some factors which are conducive to scoring judgments. It is possible, for example, that some uncontrolled variables (e.g., perceived authority, raters' personality, etc.) affected the scoring procedure.

One goal of exploring such factors which are conducive to consistency among raters is to increase the level of test fairness. It is completely advantageous for raters to employ the criteria of rating constantly and similar to each other. Another goal is to investigate the way the test construct is being understood and inferred by the selected raters, and in so doing, define construct validity more meticulously. Nevertheless, as Constable and Andrich (cited in Lumley & McNamara, 1995) maintain, the rise reliability can paradoxically cause decrease in validity of the test construct by restricting the definition through using what Cumming et al. (2002) and Charney (1984) construe as improvised criteria which are only meaningful to the special

discourse community of a group of trained raters. Determining the areas of inconsistency among raters and/or criteria raters use which are not mentioned in the scoring rubrics and rating instructions, may supply test developers with more opportunities to reassess, refine, and develop the construct through rating criteria. In this respect, the factor of inconsistency in test rating is considered as a positive and practical function in the proceeding process of test validation.

CONCLUSION

Given the discussion above, some concluding remarks could be drawn. As results of the current study indicated, fatigue can seriously affect EFL raters' judgments and consequently add construct irrelevant factors to test results and interpretations. Test bias could be triggered by various factors including, but not limited to, test method facet, raters' background, test-takers' background, test-takers' fatigue, and raters' fatigue, among many others. In broad terms, the present study suggested that raters' fatigue could result in test bias and that fatigue can have a major impact on the scores given by the raters. Scoring a great number of writing tasks is a demanding task in its own right which causes fatigue, and as a result a sudden drop in the frequency of raters' comments. With fewer comments on the writing tasks, due to fatigue, raters become more lenient and tend to assign higher scores to the tasks.

ACKNOWLEDGEMENTS

Special thanks to Dr. Hossein Barati, Associate Professor in Applied Linguistics at Isfahan University, Iran, and Dr. Mohammad Javad Ahmadian, Assistant Professor in Applied Linguistics at University of Leeds, U.K., for their helpful hints and suggestions to improve the study. They have always been a source of inspiration for researchers in the field of language teaching and testing.

REFERENCES

Anastasi, A. (1979). *Fields of applied psychology.* Mcgraw-Hill College.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. http://dx.doi.org/10.1016/j.asw.2007.07.001

Becker, A. (2011). Examining rubrics used to measure writing performance in U.S. intensive English programs. *CATESOL Journal*, *22*(1), 113-130. https://files.eric.ed.gov/fulltext/EJ1112029.pdf

Bendig, A. W. (1955). Rater reliability and judgmental fatigue. *Journal of Applied Psychology*, *39*(60), 451-454. https://doi.org/10.1037/h0046015

Brown, J. D. (1991). 1990 Manoa writing examination. *Manoa writing project technical report no. 5*. University of Hawaii at Manoa.

Brown, R., & Richards, L. (2011). *IELTS advantage writing skill: A step-by-step guide to a high IELTS score.* DELTA Publishing.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*(1), 65-81. https://www.jstor.org/stable/40170979

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. https://doi.org/10.1177/026553229000700104

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing*, *18*(2), 207–224. https://doi.org/10.1177/026553220101800206

Cumming, A., Kantor, R., & Powers, E. D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Cummings, T. S. (1954). The clinician as judge: Judgments of adjustment from Rorschach single card performance. *Journal of Consulting Psychology*, *18*(4), 243–247. https://doi.org/10.1037/h0054679

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*(1), 7-29. https://doi.org/10.1016/S1075-2935(99)80003-8

Drave, N. (2011). Marker 'fatigue' and marking reliability in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE). *IAEA.*

Erguvan, I. D & Aksu Dunya, B. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia*, *10*(1), 1-20. https://doi.org/10.1186/s40468-020-0098-3

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185. https://doi.org/10.1177/0265532207086780

Freedman, S. W., & Calfree, R. C. (1983). Holistic assessment of writing: Experimental design

and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.),. *Research on writing: Principles and methods* (pp. 75-98). Longman

Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching*, *8*(7), 225-236. https://doi.org/10.5539/elt.v8n7p225

Goodall, G. (2011). Syntactic satiation and the inversion effect in English and Spanish wh questions. *Syntax*, *14*(1), 29–47. https://doi.org/10.1111/j.1467-9612.2010.00148.x

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), 1-17. http://www.journalofwritingassessment.org/article.php?article=66

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring . *TESOL Quarterly*, 24(9), 759-762. https://doi.org/10.2307/3588173

Harsch, C., & Rupp, A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*(1), 1-33. https://doi.org/10.1080/15434303.2010.535575

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

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating students essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment*, (pp. 206-236). Hampton Press.

Ling, G., 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. https://doi.org/10.1177/0265532214530699

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H.-J., & Burton, N. W. (2004). *A study of fatigue effects from the new SAT.* ETS Research Report.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177%2F026553229501200104

Mahshanian, A., Eslami Rasekh, A., & Ketabi, S. (2017). Raters' fatigue and their comments during scoring writing essays. *Indonesian Journal of Applied Linguistics*, *7*(2),302-314. https://doi.org/10.17509/ijal.v7i2.8347

Massey, A. J. (1977). Candidate fatigue and performance on GCE objective tests. *British Journal of Educational Psychology*, *47*(2), 203–208. https://doi.org/10.1111/j.2044-8279.1977.tb02348.x

McNamara, T. (1996). *Measuring second language performance.* Addison Wesley.

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing.* In M. Chapman, M. Fujioka, Y. Ishida, & T. Newfields (Eds.), *The 3rd annual JALT Pan-SIG Conference* (pp. 45-52). Tokyo Keizai University. http://hosted.jalt.org/pansig/2004/HTML/Nakamura.htm

Reid, J. M. (1993). *Teaching ESL writing*. Prentice Hall Regents.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling . *Language Testing*, *22*(1), 1-30. https://doi.org/10.1191%2F0265532205lt295oa

Schumm, J. S., & Vaughn, S. (1991). Making adaptations for mainstreamed students: General classroom teachers' perspectives. *Remedial and Special Education*, *12*(4), 18-27. https://doi.org/10.1177/074193259101200404

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, *31*(3), 575–582. https://doi.org/*10.1162/002438900554479*

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

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*.

Vaughn, S., Schumm, J. S., Niarhos, F. J., & Daugherty, T. (1993). What do students think when teachers make adaptations? *Teaching and Teacher Education*, *9*(1)107-118. https://doi.org/10.1016/0742-051X(93)90018-C

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223. https://doi.org/10.1177/026553229401100206

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

Weigle, S. C. (1998). Using FACETS to model rater training effects, *Language Testing, 15*(2), 263-287 https://doi.org/10.1177/026553229801500205

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

Appendix 1
Rubrics for Assigning Holistic Scores to Writing Tasks
Adopted from Brown (1991)

"0"- The essay is not the wanted response to the given task, or there is no response on the paper to the task.

"0.5-1.5"-The essay suffers from general incoherence and has no discernible pattern of organization. It displays a high frequency of error in the regular features of standard written English. Lapses in punctuation, spelling, and grammar often frustrate the rater. The effort does not respond to the question as posed, or it seems not to be a serious response to the question.

"2-3"- The essay begins with a response to the topic but does not develop that response. The response suggests that the writer misread or misunderstood the topic. Ideas are repeated frequently, or are presented randomly, or both. Words are often misused, and vocabulary is limited. Syntax is often tangled and is not sufficiently stable to ensure reasonable clarity of expression. Errors in grammar, punctuation, and spelling occur often.

"3.5-4.5"- The essay provides a response to the topic but generally has no overall pattern of organization. Vocabulary often is limited. The writer generally does not signal relationships between and within paragraphs. Syntax is often rudimentary and lacking in variety. The essay has recurrent grammatical problems or because of an extremely narrow range of syntactical choices, only occasional grammatical problems appear. Sentence fragments and run-on sentence appear; the writer does not always recognize sentence boundaries. The writer occasionally misspells common words.

"5-6"- The essay shows a basic understanding of the topic, as well as the demands of essay organization. The development of ideas is sometimes incomplete or rudimentary, but a basic focus and logical structure can be discerned. Vocabulary generally is appropriate for the essay topic but at times is oversimplified. Sentences reflect a sufficient command of standard written English to ensure reasonable clarity of expression. Common forms of agreement and grammatical inflection are usually, although not always, correct. The writer's use of punctuation suggests an understanding of the boundaries of the sentence. The writer spells common words, expect perhaps so-called "demons", with a reasonable degree of accuracy.

"6.5-7.5"- The essay provides an organized response to the topic. The response is built around a central focus and is expressed in clear language most of the time. It is clear the reader has understood the passage. The writer develops ideas logically and coherently. These ideas are presented in fairly well developed paragraphs and are supported with examples. The writer generally signals relationships within and between paragraphs. The vocabulary is varied and appropriate for the essay topic and avoids oversimplifications or distortions. Sentences generally are correct grammatically, although some errors may be present when structures are particularly complex. With few exceptions, grammar, punctuation, and spelling are correct.

"8-9"- The essay reveals that the writer has understood the topic completely. It provides a full and well organized response to the topic. It has a clear thesis or focus, and the writer demonstrates control from the start. The ideas are expressed in appropriate language. They reflect an element of originality and are presented in a thoughtful and confident voice. A sense of pattern of development reflected in well-developed paragraphs, is present from beginning to end. The writer supports assertions with explanation or illustration, and the vocabulary is well suited to the context. Sentences reflect a command of syntax within the ordinary range of standard written English. Grammar, punctuation, and spelling are almost always correct.