Journal of Language Horizons

Scientefic Quarterly Journal of Language Horizons, *Alzahra University*
Volume 7, Issue 1, Spring 2023 (Serial No. 15)

The Impact of Test Length on Raters' Mental Processes During Scoring Test-Takers' Writing Performance

Research Article
pp. 159-182

Fateme Nikmard[1]
Kobra Tavassoli*[2]

Received: 2021/08/19 Accepted: 2022/02/12

Abstract

Different factors such as the writing genre, writing prompt, and/or test length can influence the raters' mental processes while scoring writing tests. Accordingly, whether an increase or a decrease in test length has any impact on how raters evaluate test-takers' writing performance was the motive underlying this research. For this purpose, 12 EFL students who scored between 5.5 to 7.5 on the writing section of a mock IELTS test were selected based on availability sampling. The participants wrote three argumentative essays (the original, longer, and shorter versions). The three versions from each test-taker were then scored by three raters using IELTS task 2 writing band descriptors. Meanwhile, the raters provided verbal protocols explaining in detail the reasons underlying their scores to each test-taker's essay. Then, the verbal protocols were transcribed and content analyzed using Nvivo version 11 to extract the themes mentioned by the raters in scoring each writing test. The results showed that the raters paid more attention to certain factors in the band descriptors and ignored some other factors. However, there was a similar pattern among the raters in scoring the three writing tests. The results did not show any significant differences in the raters' mental processes while scoring each of the three writing tests. The conclusion was that test length is not a determining factor

* Corresponding Author
[1] PhD Candidate, Department of ELT, Faculty of Literature and Foreign Languages, Karaj Branch, Islamic Azad University, Karaj, Iran. fateme.nikmard@kiau.ac.ir
[2] Assistant Professor, Department of ELT, Faculty of Literature and Foreign Languages, Karaj Branch, Islamic Azad University, Karaj, Iran. kobra.tavassoli@kiau.ac.ir

DOI: 10.22051/lghor.2022.37340.1545
DOR: 20.1001.1.2588350.2023.7.1.7.5

influencing the mental processes of raters in writing tests. Therefore, raters and test developers do not need to worry about test length influencing the raters' scoring.

Keywords: mental processes, rater, test length, verbal protocol, writing test

Introduction

Evaluation is an integral part of almost all learning processes and it is defined as the systematic gathering of information to make appropriate decisions (Alderson & Banerjee, 2002). Appropriate decision-making and the results drawn from it can greatly influence the test-takers' lives. Thus, people involved in decision-making processes, such as raters, are considered important in any act of evaluation, especially in evaluating test-takers' performance on speaking and writing tests as the scores may depend on their subjective viewpoints. Of course, there are different solutions to overcome the problem of subjectivity, one of the most important of which is to have more than one rater in evaluating test-takers' performance (Alderson & Banerjee, 2002). However, rater subjectivity and bias may be influenced by various factors such as the length of the tests, and it is vital to do research on such factors to reduce subjectivity and bias as much as possible. For instance, a longer writing test might give raters the impression that the writer is well-informed about a subject and lead them to assign a different score to such a test, while a shorter test might convey the opposite impression. In fact, Ackerman and Kanfer (2009) believed that an important factor that may influence the raters' scoring, especially in writing tests, is test length since the examinees' performance may change when the number of words they should write differs. Another important factor influencing the raters' scoring is their characteristics, such as their mental processing, which is considered a prominent factor in Eckes's (2012) perspective.

A close look at the related literature shows that even though the factors influencing rater subjectivity and bias in assigning scores to the examinees' performance have been the subject of a plethora of studies (e.g., Barkaoui, 2019; Humphrey-Murto et al., 2021; Weigle, 2002; Wind, 2019), there is a gap in the literature in investigating the effect of test length on the raters' mental processes in assigning scores to writing tests and the difference it may cause in their assigned scores to the best of the researchers' knowledge. However, a similar study was done by Hoora (2019) on investigating the examiners' mental processes when evaluating

three versions of the IELTS speaking test with different lengths. Thus, the present study aimed to check the impact of test length on raters' evaluation of test-takers' writing performance through identifying their mental processes when scoring three similar essays. Test length in this study refers to the number of words test-takers were required to write when participating in the writing tests. We asked the participants to write argumentative essays on three topics and named them the original, longer, and shorter versions. The results of such a study can be useful for test designers since if it shows that raters' mental processes differ while they score different versions of a writing test, it is necessary to think about test length as a determining factor influencing the test-takers' scores and the decisions made on their future lives.

Literature Review

Writing and its Evaluation

Writing is a common activity in which writers construct the required response and it is frequently used in most of the standardized writing assessments such as GRE, TOEFL iBT, and IELTS, to name a few (Eckes, 2012). However, as Eckes stated, a piece of writing is not a product being produced by an individual in isolation, rather, writing is assumed to be socially and culturally produced. A piece of writing is created in a particular context, is aimed to achieve a specific goal, and is directed towards a special class of audience. To produce an acceptable piece of writing, which is an extremely complex task, writers need to have a good command of words, know how to write grammatically correct sentences, be familiar with the organization of paragraphs, be familiar with the mechanics of writing, pay attention to cohesion and coherence, and develop the content appropriately (Weigle, 2002). Sahragard and Mallahi (2014) further added that when writing, learners should set a goal for it, plan their writing carefully, think about its format and structure, and finally revise it carefully. According to Cohen (2003), how a test-taker does the task of writing is an essential factor in the successful use of the cognitive and metacognitive processes associated with the writing task. Writing in a second/foreign language is even more demanding considering any of these aspects, therefore, it is essential to teach these aspects to students, as well as to pay attention to them in evaluating their writing to ensure students have learned the essential

points successfully.

A crucial aspect of writing is its evaluation (Eckes, 2012). In evaluating writing, one of the three scales of primary trait, holistic, or analytic can be used. The analytic scoring scale contains different writing components, such as organization, register, coherence, cohesion, mechanics of writing, content, and accuracy of linguistic devices. The good point about this kind of rating scale is that each element can be marked independently of the other components (Sawaki, 2007). Moreover, it has more benefits over the other scales. One of its major advantages is that since it provides information on different aspects of a piece of writing, it is more useful to recognize the writer's specific writing abilities and weaknesses (Eckes, 2012). The analytic scale can also be used as a criterion when training raters since it breaks writing into different subcategories such as vocabulary, grammar, content, organization, and mechanics of writing, by assigning a separate score to each. In other words, each writing should be scored based on the different aspects of the scale (Brown, 2005). Moreover, since different aspects of a test-taker's ability are evaluated separately, it becomes clear in which domain(s) a test-taker is strong or weak. The last important advantage of analytic scoring is the high inter-rater reliability it creates among the raters (Weigle, 2002). Because of these advantages, analytic scoring was used in this study, too.

It is worth mentioning that because of its importance, writing has been the subject of an abundance of studies throughout the years. Different aspects of writing have already been investigated, for example, the writing resources second language learners use in their writing (Oh, 2020), the effect of planning time and different task conditions on EFL learners' writing complexity, accuracy, and fluency (Fazilatfar et al., 2020), the sources of inconsistency in L2 writing scores (Barkaoui, 2019), the degree of similarities and differences between the students' performance on writing tasks and actual academic writing tasks (Llosa & Malone, 2019), and the effect of task-repetition and elicitation on the learners' writing (Asadi Vahdat & Tavassoli, 2019), to name some. However, the raters' mental processes (i.e., the way they interpret and use the rating scale) when evaluating test-takers' performance in writings of different lengths have rarely been investigated.

Raters

There are many issues involved in the process of interpreting the results of language assessment, one of which is the issue of raters in speaking and writing tests, especially when there is more than one rater scoring the same task produced by the same test-takers (Marcoulides & Ing, 2014). Raters have an influential role in evaluating writing and speaking tests since their knowledge and experience might affect the scores they assign (Duijm et al., 2018). Additionally, the raters' background seems to be a factor determining the language features they pay attention to in a piece of writing (Bijani, 2018). Kang and Veitch (2017) also confirmed the effect of raters' backgrounds on the way they assessed ESL writing pieces. In addition, rater training is considered as another influential factor (Bijani, 2018). Further, Attali (2016) investigated rater training and found that training had a parallel effect to that of experience. In other words, the scores the newly-trained raters assigned were close to those of experienced raters. On the other hand, Duijm et al. (2018) tried to minimize background effects by training raters and giving them detailed instructions, but raters in their study tended to assign dissimilar scores to the same performance due to their different backgrounds. Wind (2019) also believed that the differences in the judgments provided by various raters are mainly due to their construct-irrelevant characteristics, and such differences can threaten the fairness of their scores. Thus, researching different characteristics of raters, such as their background, preceding experience, mother tongue, and tolerance for errors like word order and verb form (Huang & Foote, 2010) can be quite informative about how they score writing and speaking tests, and how the consistency of their scoring can be improved.

One more determining factor in the scores raters assign to the test-takers' oral and written performance is the rating scale used (Purpura, 2004). Purpura believed that the analytic rating scale is a better choice in comparison to other scales since it provides the raters with detailed explanations on each important component to be measured. However, raters may translate the explanations differently and assign different scores to the same performance which may lead to bias and unreliable test scores. Purpura (2004) also offered some techniques to reduce the effects of such matters. He recommended using as clear and detailed scales as possible, training raters regarding the rubric, utilizing sample performances in

training sessions so that raters get familiar with different points of the scale, asking a third rater to judge the scores in case there is a huge inconsistency in the scores assigned by the first two raters, and continuously observing the raters and providing them with feedback if necessary.

Variations in ratings can be attributed to differences in the raters' mental processes, too (Esfandiari & Noor, 2019). That is, the scores assigned to the examinees' performance are not merely the result of the quality of their performance. The rater's cognition is also involved since the scores they assign are affected by their ability to compare and contrast the mental representations of the examinees' responses with their own mental processes (Purpura, 2014). Raters' cognitive processes, which are related to the structure of the human information processing system, can affect the way raters assign scores and the strategies they use in the act of rating (Han, 2016).

One common technique to come up with the main cognitive and/or metacognitive processes involved in the raters' scoring is using and analyzing their verbal protocols (May, 2011) through which it is possible to find out the features often emphasized by the raters (Ducasse, 2010). Lumley (2005) introduced verbal protocol (also known as think-aloud) as a way of data gathering in which the participants are required to either think aloud as they are performing a task (called introspective) or after they finished it (called retrospective). In verbal protocols, participants are asked to pronounce what comes to their minds regarding their performance. Such information is useful since it provides the researcher with good insights into the participants' cognitive processes. Formal verbal protocols are first recorded, then transcribed, and finally analyzed (Lumley, 2005).

In recent years, the raters' cognitive processes are being studied by different researchers. For instance, Esfandiari and Noor (2019) utilized a 4-stage processing model proposed by Han (2016) to investigate the cognitive processes two groups of raters (novice and expert) followed when they rated the examinees' responses to a speaking task. They concluded that different degrees of expertise have significant effects on the decisions made on different aspects of responses since they interpreted the responses differently and paid attention to different aspects of the criteria to judge the responses.

In another study, Humphrey-Murto et al. (2021) investigated the effect the

prior familiarity of the raters with the examinees has on the raters' judgment. They found that familiarity makes the raters somehow biased either towards or away from the examinees; the raters' negative prior evaluation of the examinees had a greater negative influence. They also found that, although the raters' expertise or the training they receive cannot reduce such effects, their higher levels of accountability, following particular standards, and also decreasing their cognitive role can play a reducing role for such effects.

Test Length

Test length as one of the aspects of the expected response, itself a major test method facet (Bachman, 1990), can be an influential factor influencing not only the test-takers' performance but also the raters' scoring of that performance. In fact, test method facets, including the rubric, environment, input, expected response, and the relationship between input and expected response, are probable factors underlying the inconsistencies in test-takers' performance, raters' mental processes while scoring, test interpretations, and inferences made about test-takers' abilities (Weigle, 2002). The length of a response, which can vary from a word to an essay, may be effective on the test-takers' performance since the longer the output is, the more is the possibility of the effect of intervening factors such as the knowledge of vocabulary and grammar.

Furthermore, the length of a test seriously influences productivity measures whose aim is to assess lexical or grammatical diversity (Shirai & Vercellotti, 2014). Lexical diversity, for instance, can be measured by the ratio of type to token. That is, lexical diversity is usually calculated by dividing the number of various words used in a text (type) over the entire number of used words (token). In other words, longer texts tend to achieve higher scores due to the more space they provide for the writers to use more words of different types. On the other hand, to measure grammatical complexity, the number of dependent clauses is counted in each T-unit and then averaged across all the T-units used throughout the text (Biber et al., 2011). The assumption is that using more subordinate clauses is a sign of more grammatical complexity.

Regarding the writing ability of language learners or test-takers, length is considered as an indirect measure of development; that is, it may help the writer to

elaborate more on a topic with fluidity and therefore gain a higher score. Such flexibility is then one of the common indications of the writer's proficiency level. In other words, a writer with a higher proficiency level is able to respond longer and a writer with a lower proficiency level produces a shorter response (Plakans, 2014).

Test length in different types of tests was sporadically investigated before. Ackerman and Kanfer (2009) researched the relationship between test length and cognitive fatigue. Although the more time spent on the test caused a kind of fatigue for the test-takers, it improved the quality of their performance. In a more recent study, Sahin and Anil (2017) used three unidimensional dichotomous models of item response theory (IRT) to discover the possible impacts of the two factors of test length and sample size on item parameters. They concluded that the synthesis of the two factors of test length and the sample size is more prominent than each individual factor.

Thus, it seems important to do more research on test length to examine how it might influence the raters' mental processes in scoring test-takers' writing ability. The present research was an attempt in this regard. Accordingly, the following research question was posed:

Does an increase or a decrease in test length make any difference in the raters' mental processes when they score test-takers' writing performance?

Method

This study was carried out through quantitative content analysis, in which themes and subthemes were extracted from the collected data (Dörnyei, 2007). According to Coe and Scacco (2017), content analysis presents descriptions of the data, and then, it is possible to make generalizations about the available patterns within the data. This is the quantification of patterns or coding where there are instructions about the features to be derived from the data.

Participants

Two groups of participants took part in the present study. The first group consisted of 12 female EFL students with the age range of 28-36 years old who were selected based on availability sampling. In order to ensure the homogeneity of these participants, only those who scored between 5.5-7.5 based on IELTS band

descriptors of writing were selected.

The second group of participants consisted of three female raters who were experienced EFL teachers and raters using IELTS band descriptors in IELTS preparation classes. Their age range was 31-37 and their teaching and rating experience ranged from 7 to 13 years. We tried to control the raters' gender, age, and experience to minimize their potential effect on the findings of the study. In addition, to ensure consistency in scoring, the raters participated in a one-hour training session to provide them with instructions on how to use the scale in this study and how to provide verbal protocols on their rating.

Tables 1 and 2 show the demographic information of the two groups of participants. Both groups participated willingly in this study.

Table 1

Demographic Information of the Test-Takers

Student	Age	Gender	IELTS Score
1	28	Female	6.5
2	32	Female	6
3	35	Female	7.5
4	29	Female	6.5
5	31	Female	6.5
6	35	Female	6
7	36	Female	6.5
8	30	Female	5.5
9	35	Female	6
10	32	Female	5.5
11	29	Female	6.5
12	34	Female	6.5

Table 2

Demographic Information of the Raters

Rater	Age	Gender	Experience
1	37	Female	13 years
2	31	Female	7 years
3	35	Female	11 years

Materials and Instruments

We asked the EFL students to write three argumentative essays taken from the Cambridge English IELTS 10 (2015) and named the three writings as the original version, the longer version, and the shorter version.

In the original version of the writing test, the test-takers were required to write 250 words in 40 minutes, which is the case in the IELTS writing exam. On the other hand, in the longer version of the writing test, they were asked to write 300 words in 50 minutes, whereas in the shorter version, the word limit was reduced to 200 words and time was limited to 30 minutes. The time limit and word count in the longer and shorter versions of the writing test were calculated mathematically by dividing the word count by the time limit in the original version and rounding up the numbers. Such time allotment was also used in Ahmad's (2021) study in which he allocated 30 and 50 minutes as low-timing and long-timing conditions for test-takers to write two IELTS argumentative essays.

The topics for the three writing tests were as the following.

- *The original version*:

You should spend about 40 minutes on this task.
Write about the following topic. Write at least 250 words.

It is important for children to learn the differences between right and wrong at an early age. Punishment is necessary to help them learn this distinction. To what extent do you agree or disagree with this opinion? What sort of punishment should parents and teachers be allowed to use to teach good behavior to children?

- *The longer version*:

You should spend about 50 minutes on this task.
Write about the following topic. Write at least 300 words.

Some people think that all university students should study whatever they like. Others believe that they should only be allowed to study subjects that will be useful in the future, such as those related to science and technology. Discuss both these views and give your own opinion.

- *The shorter version*:

You should spend about 30 minutes on this task.
Write about the following topic. Write at least 200 words.

Every year several languages die out. Some people think that this is not

important because life will be easier if there are fewer languages in the world. To what extent do you agree or disagree with this opinion?

In addition, the IELTS task 2 writing band descriptors was used as the scoring rubric to score the three essays in this study. The scale consists of four main components of task achievement, coherence and cohesion, lexical resource, and grammatical range and accuracy, each of which is composed of detailed descriptions on how to be evaluated.

Also, the three raters were asked to provide a verbal protocol/report when scoring each of the three versions of the participants' writings by explaining in detail the reasons underlying their scores. These verbal protocols were recorded and later transcribed and content analyzed through the Nvivo software version 11.

Procedure

Twelve EFL female students were selected based on availability sampling and their willingness to participate in the study. Each test-taker was asked to write three argumentative essays with different time limits and word counts. At first, the original version of the writing test, derived from Cambridge English IELTS 10 (2015), was administered to the students. That is, they were asked to write an argumentative essay of 250 words in 40 minutes on a certain topic as it is the case in the IELTS exam. After two weeks, the participants were asked to write a longer argumentative essay of 300 words in 50 minutes on another topic. Finally, after another interval of two weeks, the participants wrote a shorter argumentative essay of 200 words in 30 minutes on another topic. The 12 participants were those who scored 5.5-7.5 based on IELTS task 2 writing band descriptors in the original version of the test. The reason for using argumentative topics in the three versions of the writing test was to keep the variation due to genre of writing to the minimum and the two-week interval was set to reduce the potential effect of genre on the test-takers' performance.

In the next phase of the study, the three raters, who passed a one-hour training session, scored the three writings of each test-taker based on the IELTS rubric and simultaneously provided a verbal protocol/report on how and why they scored each writing and what points in the scoring rubric they paid more attention to. The verbal protocols were then transcribed, saved in three separate files (the

original, longer, and shorter versions) in Nvivo version 11, and later content analyzed to answer the research question of the study.

Results

The Coding System

The IELTS band descriptors have four criteria including (1) task achievement, (2) coherence and cohesion, (3) lexical resource, and (4) grammatical range and accuracy, which were identified as the major themes in this study. In addition, some subthemes were identified for each of these themes based on the details in the IELTS band descriptors and the data extracted from the raters' verbal protocols. The subthemes other than those that existed in the IELTS scale, which were extracted from the verbal protocols, were also categorized under the four major themes of the IELTS scale. The process of identifying and classifying the subthemes was done by the researchers who worked collaboratively to do this. We cross-checked the subthemes and their classification under the four themes with another researcher familiar with this process and made the necessary modifications. The classification of themes and subthemes is represented in Table 3.

Table 3

The Themes and Subthemes Extracted from the IELTS Scoring Rubric and the Raters' Verbal Protocols

Task Achievement	Coherence and Cohesion	Lexical Resource	Grammatical Range and Accuracy
Addressing the task	Paragraphing	Range of vocabulary	Range of structures
Writer's position towards the topic	Sequencing information and ideas	Using less/uncommon lexical items	Error-free sentences
Extending and supporting ideas	Aspects of cohesion	Word features (e.g., spelling, word choice, word formation)	Using complex structures
Answering the question	Progression of ideas		Punctuation

Non-overgeneralization of ideas	Using cohesive devices	Avoiding grammatical errors
Conclusion	Clear presentation of the central topic	
Main ideas	Referencing	
Format		
Avoiding irrelevant details		

Investigation of the Research Question

At first, the transcriptions of all the verbal protocols were imported into three Nvivo files, one for each version of the writing test (the original, longer, and shorter versions). Then, the content of each file was analyzed word by word to find out the themes and subthemes each rater mentioned in scoring each writing along with their frequency of occurrence. Tables 4-7 present sample examples from the raters' verbal protocols for the subthemes of *Task Achievement*, *Coherence and Cohesion*, *Lexical Resource*, and *Grammatical Range and Accuracy*, respectively.

Table 4

Examples for the Subthemes of Task Achievement

Subthemes	Example
Addressing the task	She addressed all parts of the task.
Writer's position towards the topic	She presented a clear position about the topic.
Extending and supporting ideas	She presented, extended, and supported ideas well.
Answering the question	She presented well-developed responses to the questions with relevant extended and well-supported ideas.
Non-overgeneralization of ideas	There is a tendency to overgeneralize, and supporting ideas lack focus.
Conclusion	Although she presented the conclusion, it may be unclear or repetitive.
Main ideas	Some limited main ideas are presented.
Format	The format is inappropriate in some places of this writing.
Avoiding irrelevant details	There are some irrelevant details and ideas.

Table 5

Examples for the Subthemes of Coherence and Cohesion

Subthemes	Example
Paragraphing	She used paragraphing but not always logically.
Sequencing information and ideas	The writing was logically organized.
Aspects of cohesion	She managed all aspects of cohesion well.
Progression of ideas	There is a clear progression of ideas throughout the writing and the passage.
Using cohesive devices	She used a range of cohesive devices appropriately throughout the essay.
Clear presentation of the central topic	She presented a clear central topic within each paragraph.
Referencing	She did not always use referencing clearly and appropriately.

Table 6

Examples for the Subthemes of Lexical Resource

Subthemes	Example
Range of vocabulary	She used a sufficient range of vocabulary that allowed some flexibility and precision.
Using less/uncommon lexical items	She used less common lexical items with some awareness of style and collocation.
Word features	She made some errors in spelling but they do not impede communication.

Table 7

Examples for the Subthemes of Grammatical Range and Accuracy

Subthemes	Example
Range of structures	She used a wide range of structures.
Error-free sentences	The majority of the sentences were error-free, and she made only very occasional errors.
Using complex structures	She used a variety of complex structures.
Punctuation	She had good control of punctuation.
Avoiding grammatical errors	She avoided some grammatical errors.

Next, Tables 8-11 report the frequency of each subtheme of the four major themes in the original, longer, and shorter versions of the writing test, along with the significance values of chi-squares to check whether the differences in each subtheme and major theme along the three writing tests were significant or not.

Table 8

The Frequency of the Subthemes of Task Achievement

Task Achievement	Original Writing Test	Longer Writing Test	Shorter Writing Test	Significance Value of Chi-Square
Addressing the task	35	33	32	.93
Writer's position towards the topic	31	27	30	.86
Extending and supporting ideas	22	26	22	.79
Answering the question	1	3	3	.56
Non-overgeneralization of ideas	6	4	3	.58
Conclusion	11	7	8	.60
Main ideas	25	31	27	.71
Format	3	0	6	.31
Avoiding irrelevant details	0	3	2	.65
Total	134	134	133	.99

As shown in Table 8, *addressing the task*, *writer's position towards the topic*, and *main ideas* were the most referred subthemes in all the three versions of the test. This shows the importance of these three subthemes for the raters. On the other hand, *non-overgeneralization of ideas*, *answering the question*, *format*, and *avoiding irrelevant details* were the subthemes the raters mentioned the least again in the three versions of the test, which shows their lower significance for the raters. The other two subthemes under this major theme (i.e., *extending and supporting ideas*, and *conclusion*) fell between these two extremes. Looking at the significance

values of chi-square comparing the differences in frequencies of each subtheme across the three versions of the writing test, it was concluded that there was not a statistically significant difference in the frequencies of the subthemes of *Task Achievement* mentioned by the three raters in the three tests since all the significance values are higher than .05 probability level (α = .05; $p > \alpha$). Further, the total frequencies of the subthemes of *Task Achievement* were very close in the three versions of the test and the related significance value of chi-square was higher than the critical .05 level (p = .99; α = .05; $p > \alpha$). It means there was not a statistically significant difference in this major theme among the different versions of the test, either. Overall, the conclusion about the major theme of *Task Achievement* was that although there were subtle differences in the frequencies of its subthemes and the total frequency mentioned by the raters, test length did not make any considerable differences in the mental processes the raters mentioned regarding *Task Achievement* while scoring the three essays.

Table 9

The Frequency of the Subthemes of Coherence and Cohesion

Coherence and Cohesion	Original Writing Test	Longer Writing Test	Shorter Writing Test	Significance Value of Chi-Square
Paragraphing	30	28	31	.92
Sequencing information and ideas	28	26	29	.91
Aspects of cohesion	5	8	7	.70
Progression of ideas	20	17	22	.72
Using cohesive devices	29	29	28	.98
Clear presentation of the central topic	11	10	5	.30
Referencing	3	2	7	.17
Total	126	120	129	.84

Based on the information in Table 9, *paragraphing* was the most frequent and therefore the most important subtheme in the raters' ideas in the three versions

of the test, which was followed by *using cohesive devices*, *sequencing information and ideas*, and *progression of ideas*, whereas *referencing* was the least mentioned subtheme, followed by *aspects of cohesion*, and *clear presentation of the central topic*. Here again, the significance values of chi-square were all above the critical value of .05 ($\alpha = .05$; $p > \alpha$), meaning that there were no significant differences in the frequencies of each subtheme of *Coherence and Cohesion*. Once more, the total frequencies of the subthemes of *Coherence and Cohesion* were close to each other and the related significance value of chi-square was non-significant and above the critical value of .05 ($p = .84$; $\alpha = .05$; $p > \alpha$). Thus, there was not any substantial difference between the raters' viewpoints regarding the importance of the subthemes and the major theme of *Coherence and Cohesion* in the three versions of the test.

Table 10

The Frequency of the Subthemes of Lexical Resource

Lexical Resource	Original Writing Test	Longer Writing Test	Shorter Writing Test	Significance Value of Chi-Square
Range of vocabulary	33	33	32	.99
Using less/uncommon lexical items	18	16	13	.66
Word features	32	30	33	.92
Total	83	79	78	.91

As it can be seen in Table 10, *range of vocabulary* and *word features* (e.g., spelling, word choice, and word formation) were the most frequently mentioned subthemes of *Lexical Resource* by the raters in the three versions of the test, whereas the subtheme of *using less/uncommon lexical items* was the least frequently mentioned one, which shows its less significance in the raters' views. Once again, the results of the significance values of chi-square showed that there were no significant differences in the frequency of the subthemes of *Lexical Resource* as the related significance values of chi-square were all above the critical value of .05 ($\alpha = .05$; $p > \alpha$). In addition, the total frequencies of the subthemes of *Lexical Resource* were not much different from each other in the three versions of the test and the

related significance value of chi-square showed a non-significant difference since it was above the critical value of .05 (p = .91; α = .05; $p > \alpha$). To summarize, regarding the subthemes as well as the major theme of *Lexical Resource*, there was no significant difference between the raters' ideas in scoring the three versions of the writing test.

Table 11

The Frequency of the Subthemes of Grammatical Range and Accuracy

Grammatical Range and Accuracy	Original Writing Test	Longer Writing Test	Shorter Writing Test	Significance Value of Chi-Square
Range of structures	26	24	32	.56
Error-free sentences	10	10	5	.36
Using complex structures	13	21	16	.37
Punctuation	14	24	23	.22
Avoiding grammatical errors	28	29	25	.85
Total	91	108	101	.48

Finally, the information in Table 11 shows that *avoiding grammatical errors*, *range of structures*, and *punctuation* were the most-frequently mentioned subthemes of *Grammatical Range and Accuracy* by the raters, and therefore, the most important ones in the three versions of the test from the raters' perspectives. On the other hand, *error-free sentences* was the least mentioned subtheme by the raters, and therefore, the least important one again in the three versions of the test. The other subtheme under this major theme (i.e., *using complex structures*) fell between these two extremes. The same as the previous major themes, the significance values of chi-square for all the subthemes of *Grammatical Range and Accuracy* were above the critical .05 value (α = .05; $p > \alpha$), meaning that there were no statistically significant differences in the frequency of the mentioned subthemes. Further, the total frequencies of the subthemes of *Grammatical Range and Accuracy* were compared with each other. Here again, the related significance value for chi-square was non-significant and above the critical value of .05 (p = .48; α = .05; $p >$

α). Once more, the conclusion was that regarding the major theme of *Grammatical Range and Accuracy*, test length did not make any considerable differences in the mental processes the raters mentioned in scoring the three versions of the writing test.

To wrap up the findings and to answer the research question of the study, it can be said that content analysis of the raters' verbal protocols showed that neither increase nor decrease in test length made any noticeable differences in the raters' mental processes while scoring writing tests.

Discussion

To summarize the results of the present study, it can be said that the length of the writing test did not result in any significant differences in the mental processes through which raters assigned scores to the test-takers' writing performance. In other words, the raters scored the three versions of the writing test (the original, longer, and shorter versions) with a similar viewpoint towards the important elements of writing.

The results of this study are in line with other studies in the literature. In a similar study, Hoora (2019) investigated the effect of the duration of the IELTS speaking test on the examiners' evaluation of the candidates' performance. Similar to the results of this study, it was found that, although there were certain factors in the band descriptors of the IELTS speaking test the raters paid more attention to while ignoring some other factors in rating three versions of the IELTS speaking test (i.e., the original, longer, and shorter versions), their mental processes in scoring the three versions were not significantly different from each other. In addition, there was not a considerable difference in the scores they assigned to the three versions of the speaking test. The similar results of this study and Hoora's show that test length, whether in writing or speaking tests, does not make a significant difference in the raters' mental processes when scoring the test-takers' performance.

Examining the raters' preferences in scoring the test-takers' performance, Van Batenburg et al. (2018) researched the examiners' judgment of the speaking performance of learners in different task types where the raters scored the participants' performance both holistically and analytically using two different scoring rubrics. They found that there was a high correlation between the scores

given by the three raters on both holistic and analytic scales in measuring different task types. In other words, the scores assigned by the three raters to different task types were not significantly different from each other following either scale. These results are somehow in line with the results of the present study, showing the consistency in the raters' scoring in different situations.

Considering the differences in the raters' scoring of writing tests, Attali (2016) conducted a study in which the performance of newly-trained and experienced raters was compared. He came up with the conclusion that the scores provided by the newly-trained raters were very much similar to the ones by the experienced raters. In other words, the training turned out to be an influential factor in increasing the correlations between the raters' scores. In fact, training resulted in a beneficial decrease in the difference between the newly-trained and experienced raters' scoring of writing. Thus, training, in contrast to test length, made a difference in the rater's scoring of writing performance.

Overall, the comparison of the results of the present study with other studies showed that raters, their mental processes, and their characteristics (e.g., experience) can be considered as important factors when scoring speaking and/or writing performance. Regarding the raters' mental processes, which was the focus of the present study, the interesting finding was that the length of the writing test was not a determining factor for the raters and they paid more attention to the criteria in the writing scoring rubric regardless of the test length. The differences in the tasks or tests may not make much differences in the scores assigned by the raters or the mental processes they engage in as the results of this study showed. In other words, variations in task types, the word count (or test length), the time allotted to do the task, and related issues might not have a noteworthy effect on the scores assigned by the raters. However, the raters' individual characteristics, such as experience, training, and related factors might be probable influential factors.

Conclusion

The goal of this study was to investigate the effect of test length on the raters' mental processes while scoring candidates' writing performance where both groups of participants (i.e., test-takers and raters) were chosen based on availability sampling. In general, it was found that raters considered certain elements of writing

more important and disregarded some other elements when assigning scores to a piece of writing regardless of its length.

The main limitation of the study was the few number of test-takers and raters, which reduced the generalizability of the findings. Therefore, the results of the study should be interpreted cautiously. Further, the major delimitation of the study was that both test-takers and raters were female since we wanted to control the gender effect in this study. However, using a combination of male and female test-takers and raters makes comparisons across the different groups possible and may result in more valuable findings in the future.

The results of this study can be useful for raters since, to be a fairer rater who assigns more reliable scores, it is necessary to be aware of one's own characteristics, to go to routine training sessions, to gain more experience, to get familiar with different types of scoring rubrics, and so forth. The findings of the present study should make raters aware of different important points to take into consideration when scoring a piece of writing following a specific scoring rubric in addition to the common and typical points.

Rater trainers can also benefit from the results of the present study since it makes them aware of some important points raters do not usually pay attention to in scoring writing. Although *answering the question*, *format of the essay*, *avoiding irrelevant details*, *aspects of cohesion*, and *referencing* are some of the essential points that need more attention on the part of the raters, it was found in the present research that raters often ignore them. Therefore, rater trainers should emphasize such points more when training the raters.

Finally, interested researchers are invited to conduct follow-up research where the personality type of raters, their gender, their experience, and their training are considered to see if such differences result in the same or different mental processes raters engage in when assigning scores to different types of writing and/or speaking tasks or tests.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163-181. https://doi.org/10.1037/a0015719

Ahmad, B. A. D. (2021). Effect of time allotment on test scores for academic writing of Indonesian learners of English. *Multicultural Education, 7*(1), 134-141.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(2), 79-113. https://doi.org/10.1017/S0261444802001751

Asadi Vahdat, Z., & Tavassoli, K. (2019). A comparison of the effects of task repetition and elicitation techniques on EFL learners' expository and descriptive writing. *Language Horizons, 3*(1), 243-267. https://doi.org/10.22051/lghor.2019.27890.1173

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 1-17. https://doi.org/10.1177/0265532215582283

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

Barkaoui, K. (2019). Examining sources of variability in repeaters' L2 writing scores: The case of the PTE Academic writing section. *Language Testing, 36*(1), 3-25. https://doi.org/10.1177/0265532217750692

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5-35. https://doi.org/10.5054/tq.2011.244483

Bijani, H. (2018). The investigation of rater expertise in oral language proficiency assessment: A multifaceted Rasch analysis. *Language Horizons, 2*(2), 103-124. https://doi.org/10.22051/lghor.2019.26072.1123

Brown, J. D. (2005). *Testing in language programs* (2nd ed.). McGraw-Hill College.

Coe, K., & Scacco, J. M. (2017). Content analysis, quantitative. In C. S. Davis & R. F. Potter (Eds.), *The international encyclopedia of communication research methods* (pp. 1-11). John Wiley & Sons, Inc.

Cohen, A. D. (2003). Learner strategy training in the development of pragmatic ability. In A. Martinez Flor, E. Usó Juan, & A. Fernández Guerra (Eds.), *Pragmatic competence and foreign language teaching* (pp. 93-108). Publicacions de la Universitat Jaume I.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies.* Oxford University Press.

Ducasse, A. M. (2010). *Interaction in paired oral proficiency assessment in Spanish: Rater and candidate input into evidence-based scale development and construct definition.*

Peter Lang.

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing, 35*(4), 501-527. https://doi.org/10.1177/0265532217712553

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292. https://doi.org/10.1080/15434303.2011.649381

Esfandiari, R., & Noor, P. (2019). Iranian EFL raters' cognitive processes in rating IELTS speaking tasks: The effect of expertise. *Journal of Modern Research in English Language Studies, 5*(2), 41-76. https://doi.org/10.30479/jmrels.2019.9383.1248

Fazilatfar, A., Kasiri, F., & Nowbakht, M. (2020). The comparative effects of planning time and task conditions on the complexity, accuracy, and fluency of L2 writing by EFL learners. *Iranian Journal of Language Teaching Research, 8*(1), 93-110.

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 16*(1), 1-24. https://doi.org/10.7916/D82R53MF

Hoora, E. (2019). *The effect of the duration of IELTS speaking test on examiners' evaluation of candidates' performance* [Unpublished Master's thesis]. Karaj Islamic Azad University, Iran.

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*(3), 219-233. https://doi.org/10.1080/15434300903540894

Humphrey-Murto, S., Shaw, T., Touchie, C., Pugh, D., Cowley, L., & Wood, T.J. (2021). Are raters influenced by prior information about a learner? A review of assimilation and contrast effects in assessment. *Advances in Health Sciences Education, 25*(2), 4-24. https://doi.org/10.1007/s10459-021-10032-3

Kang, H. S., & Veitch, H. (2017). Mainstream teacher candidates' perspectives on ESL writing: The effects of writer identity and rater background. *TESOL Quarterly, 51*(2), 249-274. https://doi.org/10.1002/tesq.289

Llosa, L., & Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing, 36*(2), 235 263.

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

Marcoulides, G. A., & Ing, M. (2014). The use of generalizability theory in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment: Evaluation, methodology and interdisciplinary themes* (Vol. 3, pp. 124-141). John Wiley & Sons, Inc.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145. https://doi.org/10.1080/15434303.2011.565845

Oh, S. (2020). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly, 17*(1), 60-84. https://doi.org/10.1080/15434303.2019.1674854

Plakans, L. (2014). Written discourse. In A. J. Kunnan (Ed.), *The companion to language assessment: Evaluation, methodology and interdisciplinary themes* (Vol. 3, pp. 305317). John Wiley & Sons, Inc.

Purpura, J. E. (2004). *Assessing grammar.* Cambridge University Press.

Purpura, J. E. (2014). Cognition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp.1452-1476). John Wiley & Sons, Inc.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in 1452item response theory. *Educational Sciences: Theory & Practice, 17*(1), 321-335. https://doi.org/10.12738/estp.2017.1.0270

Sahragard, R., & Mallahi, O. (2014). Relationship between Iranian EFL learners' language learning styles, writing proficiency, and self-assessment. *Social and Behavioral Sciences, 98*(1), 1611-1620. https://doi.org/10.1016/j.sbspro.2014.03.585

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390. https://doi.org/10.1177/0265532207077205

Shirai, Y., & Vercellotti, M. L. (2014). Language acquisition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment: Evaluation, methodology and interdisciplinary themes* (Vol. 3, pp. 300-314). John Wiley & Sons, Inc.

University of Cambridge ESOL examinations. (2015). *Cambridge English IELTS 10 with answers: Authentic examination papers from Cambridge English language assessment*. Cambridge University Press.

Van Batenburg, E. S., Oostdam, R. J., Van Gelderen, A. J., & De Jong, N. H. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Language Testing, 35*(1), 75-100. https://doi.org/10.1177/0265532216679452

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

Wind, S. A. (2019). A nonparametric procedure for exploring differences in rating quality across test-taker subgroups in rater-mediated writing assessments. *Language Testing, 36*(4), 595-616. https://doi.org/10.1177/0265532219838014

©2020 Alzahra University, Tehran, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC- ND 4.0 license) (https://creativecommons.org/licenses/by-nc-nd/4.0/)