Journal of Teaching Language Skills
(JTLS)
39(2), Summer 2020, pp. 43-87
DOI: 10.22099/jtls.2020.37967.2858

Online ISSN: 2717-1604
Print ISSN: 2008-8191

Research Paper

Exploring Novice Raters' Textual Considerations in Independent and Negotiated Ratings

Leila Hajiabdorrasouli [*] **Alireza Ahmadi** [**]

Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran

Abstract

Educators often employ various training techniques to reduce raters' subjectivity. Negotiation is a technique which can assist novice raters to co-construct a shared understanding of the writing assessment when rating collaboratively. There is little research, however, on rating behaviors of novice raters while employing negotiation techniques and the effect of negotiation on their understanding of writing and rubric features. This study uses a qualitative method to keep track of 11 novice raters' scoring behaviors and examine their textual foci during three phases of scoring through an analytic rubric: pre-negotiation, negotiation, and post-negotiation. To ensure triangulation, multiple sources of data including raters' verbal protocols of independent scoring during the initial and final phases, audio-recorded interactions in the negotiation phase, and semi-structured interviews were gathered and analyzed. Results indicated that in their initial independent rating, raters mostly scored based on their understanding of the writing skill and the writing features that were important to them, but negotiation sessions aided them to refine their judgments and attend to a wider array of textual features more consistently and in line with the rubric, thereby expanding their understanding of the rubric categories. Post-negotiation ratings were also more similar to negotiation than prenegotiation ratings, meaning that the raters attended to more features of the rubric for scoring. The findings

Received: 03/08/2020 Accepted: 26/10/2020

[*] Ph.D. Candidate, Email: l_rassouli@iauba.ac.ir
[**] Associate Professor, Email: arahmadi@shirau.ac.ir, Corresponding Author

may have implications for rater training. In the absence of expert raters to train novice raters, negotiation can be considered a useful technique to improve raters' understanding of the rubric features.

***Keywords**:* Negotiation, Novice raters, Analytic rubric, Textual features, EFL Writing assessment

Writing assessment has always been a challenging issue for raters. Though rater training can help them develop competency in rating (Davis, 2016), studies have indicated that even after training, raters still show subjectivity in rating (see e.g., Eckes, 2005; Lumley, 2002, 2005; Papajohn, 2002) because in the complex process of rating, they are affected by various extraneous factors including rater experience and background (e.g., Attali, 2016; Davis, 2016; Kim, 2015), rating scale (e.g., Barkaoui, 2010; Davis, 2016; Eckes, 2005, 2012; Isaacs & Thomson, 2013), and task type (e.g., Ahmadi & Sadeghi, 2016; In'nami & Koizumi; 2016) among others. As such, exploring the rating process and the features which draw the raters' attention during rating has always been an intriguing topic for researchers.

On the other hand, score resolution methods are employed to reduce subjectivity in rating. Negotiation (discussion) is one of the resolution methods through which raters come together and engage constructively to resolve score discrepancies (Broad, 1997; Johnson, Penny, Gordon, Shumate, & Fisher; 2005; Moss, Schutz, & Collins, 1998). This method, which was "originally adopted due to resource constraints-in particular the lack of trained raters" (Trace, Janssen & Meier, 2017, p. 2) for example in EFL contexts, has been shown to be effective in minimizing score variance (Clauser, Clyman, & Swanson, 1999; Johnson et al., 2005) and in enhancing validity by providing opportunities to the raters to refine their judgments while scoring collaboratively, thereby making more credible judgments (Ahmadi, 2019; Lindhardsen, 2018; Trace et al., 2017). The literature suffers from a lack of research in the negotiation scoring method. It is not yet known what exactly happens in the negotiation scoring method. What features do raters focus on when using negotiation? Are these features different from the ones considered in individual ratings? How does negotiation affect raters in shifting their attention to different features? The current study was aimed at answering these questions by focusing on the rating process in an independent rating and negotiation rating method. The study intended to enlighten the understanding of textual features raters attend to while scoring independently or through negotiation. Following the relevant literature in rater orientation studies (e.g., Broad, 2003; Ducasse & Brown, 2009; May, 2009) textual features were defined in a general way to include any text-related features raters find salient in rating performance samples.

Review of Literature

Studies on rating behavior have mainly focused on the raters' rating behavior and specifically the textual features they attend to while scoring independently. Although some studies have also focused on exploring raters' scoring behaviors and their textual considerations in negotiation scoring sessions, these studies are sparse. Such studies have mainly focused on the rating behavior of experienced raters and the features salient to them during negotiation rating sessions. In what follows, at first, the studies investigating raters' rating behavior in independent ratings are presented. Then in the second section, the studies specifically focusing on the negotiation rating method are reviewed. In both sections, attempts are specifically made to refer to the textual features considered by raters while rating performance samples. It must, however, be noted that not all the studies of rating behavior have focused on features salient to raters as well.

Numerous studies have explored the features raters attended to in rating second language writing (e.g., Barkaoui, 2010; Broad, 2003; Cumming, 1990; Cumming et al., 2001; Eckes, 2008; Lumley, 2002, 2005; Lim, 2019; Lumley & McNamara, 1995; Vaughan, 1991). One of the earliest studies conducted on the textual aspects which raters focused on is Vaughan's (1991). Based on the essay features salient to the raters, Vaughan categorized the raters' rating into the single-focused approach, the first-impression-dominates approach, and the grammar-oriented approach. In a similar attempt and in a series of studies, Cumming (1990), and then Cumming, Kantor, and Powers, (2001, 2002) provided a descriptive taxonomy showing a detailed account of rater behaviors. This taxonomy is one of the most widely-cited taxonomies of raters' behaviors in the literature. As the raters in these studies were not provided with a rating scale, the researchers do not discuss the effect of a rating scale on the aspects of writings considered salient to the raters; thus one important point argued in their studies was the need for holistic and analytic scoring procedures.

It is obvious that the rating scale is one of the most important factors in the rating context because it mainly determines what raters attend to while assigning scores (Weigle, 2002). Thus, a large number of studies have investigated scoring behaviors of raters while adopting these rating scales. Some studies solely focused on whether the raters adopting a rating scale have a similar understanding of the scale criteria; treated the categories consistently or attended to other criteria not present in the scoring scale. These studies (e.g., Lumely, 2002, 2005; Sakyi, 2003; Vaughan, 1991) indicated that raters may have different understandings of the same rating criteria, use criteria not mentioned in the rating scale such as handwriting and length (e.g., Barkaoui, 2010, Lumely, 2005; Smith, 2000; Vaughan, 1991) or weigh some criteria over others (e.g. Barkaoui, 2010; Cumming et. al., 2002; Eckes, 2008, 2012).

For example, Smith (2000) reported that raters were not consistent in interpreting and applying the assessment criteria and adopted different approaches in their rating. He stated that although raters adopted the rating criteria to justify their scores, they had different interpretations of the rubric. For instance, the raters who adopted the first impression-dominates approach relied more on their internalized and personalized view about the quality of writing and did not adhere strictly to the guidelines of the criteria. The first-impression-dominates raters commented on greater number of features and were influenced by extraneous features particularly at the level of textual coherence which was absent in the rating rubric. Lumley's (2005) findings also could answer some of the relevant questions concerning the role of the rating scale in scoring and the aspects of writing which raters attended to. He postulated that although raters used the scale consistently in general, they weighed the components differently based on their personal interpretation. Moreover, Lumley reiterated the findings of Vaughan (1991) and Smith (2000) as when raters encountered difficulty, they took advantage of different resolution strategies like comparing the essays and using their own interpretation of the scale.

In terms of using analytic or holistic rubrics and their effect on writing features raters attend to, mixed results are reported in the literature. For example, Eckes (2008) found that raters do not distribute their attention evenly across all the categories of an analytic rubric. Based on raters' foci on different features of writing, Eckes categorized them into six types: the syntax type who focused strongly on vocabulary, syntax, argumentation, and completeness; the correctness type with a strong focus on correctness; the structure type with a strong focus on the global impression; the fluency type; the non-fluency type and the non-argumentation type who put less emphasis on argumentation and train of thought. This study provided a fuller description of raters' types based on the categories they attended to.

To explore the differences between different types of rubrics and the features salient to the raters, Barkaoui (2010) indicated that while scoring holistically, raters attended to the essay features more frequently, whereas scoring analytically they attended to the scale categories. Kim and Lee's (2015) findings were similar. They concluded that when using a holistic scale, raters focused on the quality of language and non-scale features, while when scoring analytically, they paid more attention to coherence and grammar. Lim (2019) reiterated the findings of previous studies that raters considered spelling errors, length, authorial voice, and syntactic voice but did not attend to negation density, conceptual cohesion, and noun phrase density. All of these studies conducted on the independent scoring process revealed that despite using scoring rubrics, raters generally did not score uniformly and tended to distribute their attention across different aspects of writing unevenly.

Studies on Rater Negotiation

Studies focusing on the writing features raters attend to while scoring collaboratively are still scarce. Among the earliest studies on rater negotiation, Moss, Schutz, and Collins (1998) employed rater negotiation for mathematic portfolio assessments where the experienced teachers serving as raters were engaged in collaborative and dialogic assessment. While in independent rating, the teachers demonstrated instances of having different interpretations of the rating criteria, negotiation helped them to be more consistent in following the assessment criteria and having a sound evaluation.

In another exploratory study, Jølle (2014) studied rating behaviors of novice raters over ten months while scoring in pairs and examined the extent their scoring practices changed over time. The raters' assessment practices were classified into two categories: rating practices with which the raters were already familiar and rating practices developed within discussion sessions. The first category comprised reference to text, citation of text, reference to the initial score, comparing texts and referring to text knowledge. The use of meta-discussion such as discussing the quality of rubric, plenary discussion, referring to rubric and expert was developed within the discussion scoring sessions. Results revealed that the majority of their rating practices were "referring to the texts" indicating that the traditional scoring practice to argue the quality of the text was the dominant scoring practice throughout the period. Also, Jølle found that as the novice raters gained expertise and proficiency over time, their use of assessment rubric to validate their judgments and metadiscussion significantly increased providing positive evidence for the role of collaborative assessment in developing rating expertise.

In a recent study, Lindhardsen (2018) explored the decision-making behaviors of raters in independent and negotiation rating sessions in the context of writing assessment. The researcher employed think-aloud procedures and retrospective reports from experienced raters to identify how the raters distribute their attention to the assessment criteria. In independent rating sessions, raters employed a complex rating practice and attended to a wide range of textual and contextual features, with specific attention given to language. In negotiation rating sessions, attention to language-related features was reduced to less than half and was devoted to other textual features like content, organization, style, format, and amount of text. The results indicated that as the raters moved from independent rating to negotiation rating, their attention to the features corresponding to the rating criteria became more balanced.

Although the literature is replete with studies investigating raters scoring behaviors and their textual foci in independent scoring sessions, little is known about the raters' textual foci and rating behaviors in negotiation scoring sessions. Those few studies which have examined the raters' textual foci were confined to the portfolio assessment (e.g., Jølle 2014; Moss, et al., 1998). The only study which compared independent and negotiation ratings in language-related fields was conducted by Lindhardsen (2018), explained above. This study was limited to the experienced raters and more importantly failed to follow raters' behavior after negotiation, to explore whether after attending negotiation sessions raters would return to their individual rating habits or would continue using the negotiation habits. So the current study is methodologically different from previous studies as it is the only qualitative study employing an inductive method (grounded theory method) to analyze and compare the textual features raters attended to in three phases of scoring: two rounds of independent scorings and negotiation scoring sessions. In other words, it is the only study that has focused on exploring raters' behavior before and after receiving negotiations. As such, it was aimed at filling the above gaps by first focusing on what features novice raters rather than experienced raters would attend to, and second, by comparing the features they would attend to in independent rating, in negotiation and collaboration with other raters, and in post-negotiation independent ratings. The study employed think-aloud protocols, raters' interactions, and interviews to identify different aspects of writing attracting raters' attention. Thus, the following research questions were put forward.

a) What features do novice raters consider crucial when rating writing samples individually?

b) What features do novice raters consider crucial when rating writing samples through group negotiations?

c) What features do novice raters consider crucial when rating writing samples individually after group negotiations?

Method

Design

To explore the textual features raters attended to, the researchers employed grounded theory based on the field data. Grounded theory as a qualitative search approach develops a theory of social phenomena. The theories emerge from the data; experience with the data generates insights, hypotheses, and questions, which researchers pursue with further data collection. This inductive qualitative approach explains a process, interaction, etc. In this approach, theories are developed through induction and verification techniques (Ary, Jacobs, Razavieh, & Sorenson, 20۱8).

Participants

Eleven raters including two males and nine females, who ranged in age from 25 to 39 with a mean of 31.7 participated in this study. To ensure the homogeneity of the participants, only those with similar backgrounds in terms of teaching and rating experiences and educational background were selected. They were all MA students of Teaching English as a Foreign Language and had passed courses on language teaching and testing. All had 2-5 years of experience teaching English in language institutes, but none had attended any rater training programs, and their rating experience was very limited, so in line with the literature (e.g., Kim, 2015) they were characterized as novice raters. They were paid for their participation in this study.

Materials

Writing Samples. For the independent scoring sessions, the raters were asked to score 20 scripts written by junior students majoring in English at a university in a southern city of Iran. The scripts were responses to IELTS writing task 2, in which the students had written an essay on a topic about two types of teaching. The students had already passed two writing courses, namely Basics of Writing and Advanced writing. Ten writing scripts were also selected from the Cambridge IELTS series (2015, 2016) for use in negotiation rating sessions.

Analytic Rubric. In both the independent and negotiation scoring sessions, the raters were asked to rate the scripts based on IELTS Writing rubric for task 2 including four rating criteria: Task achievement, Cohesion and Coherence, Lexical Resource, and Grammatical Range and Accuracy.

Verbal Protocol. The present study used a think-aloud protocol to identify the features the raters took notice of while rating the scripts in the independent scoring sessions. The raters were asked to report the stream of their thoughts and whatever came to their minds while scoring the essays following the procedures explained in the literature (Barkaoui, 2007, 2010; Cumming et al., 2002; Lumley, 2002). The verbal protocols were audio-recorded and then transcribed for analysis.

Interview. As think-aloud protocol may have limitations in portraying the real process of rating (Barkaoui, 2010), immediately after the introspective process, the semi-structured interviews were conducted to uncover the textual aspects raters attended to while rating independently. The interviews were conducted individually to obtain further clarification on rating. The semi-structured interviews covered a broad range of questions like the raters' perceptions about the elements of writing and features of the rubric and their descriptors, their perspectives about the use of analytic rubric during the rating process in independent and negotiation rating sessions, and the strategies they employed in independent ratings. All the questions were asked in English and audio-recorded. The interviews were transcribed by the first researcher for further analysis.

Data Collection Procedure

The data were collected in three phases (Table 1):

Phase 1: Pre-negotiation phase. The participants attended a briefing session about the objectives of the research and using the analytic rubric. In addition to instruction on the features of the rubric and scoring process in independent and collaborative sessions, the raters were trained on how to verbalize their thoughts while rating the essays independently. After this introductory session, the raters were asked to score 20 scripts individually while thinking aloud. Immediately after scoring, semi-structured interviews were conducted. All the think-aloud and interview sessions were audio-recorded for further analysis.

Phase 2: Negotiation phase. After the pre-negotiation phase, the participants attended negotiations that lasted for eight weeks (each week, one session of 60-90 minutes). The participants were randomly assigned to two groups of 6 and 5 members. Overall, 10 scripts were discussed by the raters in these sessions. Each session started by first having raters score one or two writing samples independently and then discussing their scores in groups to resolve discrepancies in rating. They reviewed the samples and the rating criteria together, challenged each other's' scores, provided reasons for scoring and tried to come to a consensus. All the interactions were audio-recorded.

Phase 3: Post-negotiation phase. In this phase, the participants rated the same scripts as those in the pre-negotiation phase independently. Like the initial phase, we used the think-aloud protocol to identify the features they attended. An interview was conducted at the end about the raters' perceptions of negotiation sessions, the process of rating and the features they attended. The think-aloud protocols and interviews were audio-recorded for further analysis.

Table 1.
Different Phases of the Study

Phase 1: week 1	Phase 2: weeks 2-9	Phase 3: week 10
Presentation phase	Independent rating of one or two scripts in each session (overall 10 scripts)	Independent rating of 20 samples by each rater (think aloud and interview)
Independent rating of 20 samples by each rater (think aloud and interview)	Negotiation on the discrepant scores	

Coding System, Reliability and Validity

The data were analyzed through the procedures suggested by Corbin and Strauss (2014). To make meaning from the raw data, open, axial, and selective coding procedures were conducted to induce the categories. First, for open coding, the qualitative data were reviewed several times to identify the codes and thematic categories, and then in the axial coding stage, the subcategories of each theme based on common axes were illuminated. Finally, selective coding was utilized to show how the emerged thematic categories are related, then themes were supported by the relevant quotations. The themes and their subcategories were extracted from a thorough evaluation and reevaluation of the transcripts and were finalized after many modifications and remodifications.

To enhance the credibility or trustworthiness of the data, triangulation of data, peer review and member check were conducted. As stated by Cresswell (2009), data triangulation is used to build coherent themes by converging several sources of data. In fact, in the triangulation of the data, the researcher investigates if the data collected by one source of data confirm the data collected by another instrument (Ary et al., 2018). In this case, to enhance the credibility of the data, the researchers triangulated three sources of the data:

(a) raters' interactions in negotiation scoring sessions, (b) raters 'verbal protocols in initial and final independent scoring sessions (c) interviews conducted in initial and final independent scoring sessions. Another invaluable procedure to establish the credibility of the data is peer review. To this end, an expert in ELT qualitative research was asked to check the codes, themes, and categories.

To check the reliability of the coding, initially, the coding schemes were discussed with a third person, other than those who conducted the study. She was an expert in TEFL with experience in rating. Then 10% of the transcripts were randomly selected and independently coded by her. The Kappa coefficient for the inter-rater agreement was 0.78. To reach consensus in controversial codes, themes and subthemes, the coders negotiated to resolve disagreements. The rest of the data were then coded by the first researcher.

Results

The examination of the raters' interactions, verbalizations, and interviews identified the major categories of textual features they attended to. Some features were common in different stages (although the raters' level of attention to and perception of such features varied qualitatively), and some were unique to negotiation and to some extent to the post-negotiation phase. The features are explained below. Although this study aimed to explore the textual features salient to raters in rating writing samples, some contextual features were also noticed and referred to by the raters. So, they were included in the findings of the study as well.

Common Features Across the Three Phases

Table 2 exhibits the features that were commonly noticed by the raters in all the phases. It lists the major categories and subcategories of features. The features that were noticed by the raters in the prenegotiation phase were mainly manifestations of raters' knowledge of EFL writing courses and to some extent features of the analytic rubric. The negotiation process provided the raters with this opportunity to discuss the rubric criteria and descriptions and increase their understanding of the features which were vague for them initially. In post-negotiation scoring, most of the raters attended to the features negotiated in the collaborative scoring sessions, although some subtle differences were observed.

Table 2.
Common Features in the Three Phases

General textual and contextual features	**Subfeatures**
Textual features	
Ideational and rhetorical features	Ideational features
	Organization and Coherence of ideas
Language-focused features	Grammatical Complexity
	Frequency and gravity of Grammatical Errors
	Lexical and word formation errors
	Lexical Diversity
	Spelling Errors
	Mechanics
Non- scale features	Authorial Voice Comprehensibility of ideas Handwriting
Contextual features	Comparing the essays Language proficiency

Ideational and Rhetorical Features

These features were noticed in all three stages, though not with the same focus. Comparing to other features, the raters spent more time discussing scoring the ideational and rhetorical features in negotiation and post-negotiation phases. But in the prenegotiation phase, the raters tended to span their attentions to the language-focused features more.

Ideational Features. One of the most prevalent features in all three phases was the ideational feature. However, raters' understanding of these features varied noticeably across the phases. Also, the allotted time to discuss this feature, and the quality of the evaluation varied across phases. In the prenegotiation phase, unexceptionably, all the raters attended to this feature. In the following excerpts, these raters in their verbal protocols clearly point to the idea development, completeness, and relevance of the ideas.

G: He couldn't cover all the ideas, for example in Band 5, it says it addresses all the tasks but in a minimal way, so his explanations are not enough….

H: I can't give him 2, because he could support the third reason…

K: The reason is clear, but he couldn't justify the reasons and explain the conclusion, then I have to decide whether I should give him a higher or lower score.

Comparing to the initial phase, in the negotiation sessions, the ideational features were discussed more meticulously, as the raters repeatedly referred to the scale descriptors displaying this feature to justify their assigned scores. In terms of time allotment, the raters spent more time discussing them, and comparing to the prenegotiation phase, a wider range of ideational features such as idea development, task fulfillment, the relevance of ideas, completeness and originality of ideas were discussed too, suggesting that negotiations made the raters attend to the scale more carefully. The raters' negotiation on this feature shows that they gained more control over using the scale and comprehending this feature of the essay. In the following excerpt, the raters are discussing the relevance of ideas while referring to the scale descriptors.

G: I gave her 6 because it presents the relevant ideas but some main ideas are inadequately developed.

H: I think that this essay addresses the task partially, its development is not clear.

I: The ideas are repetitive and irrelevant in some parts and don't have a clear position.

The verbalizations of raters in the post-negotiation phase showed that like the negotiation phase the raters had formed the tendency to refer to the scale to justify their scores more frequently. It is worth noting that most of the ideational features discussed in the negotiation phase were mentioned in the post-phase too.

Organization and Coherence of Ideas. The interview data conducted as confirmatory for the think-aloud data revealed that initially, the raters had different perceptions toward the construct of cohesion and coherence. For some raters scoring this category was easy; however, for others making sense of the rubric description for this criterion was difficult. For instance, Rater H stated that "*for me scoring cohesion and coherence was not difficult, first I found the main idea, then I looked whether paragraphing is used or not, and at last I checked whether the main idea is supported or not*", or another rater defined cohesion and coherence as "*when there is a logical relationship among the parts of the essays"* or "*writing the essay within the outline*". As it is evident, there was no consensus among the raters on how to approach this criterion in the prenegotiation phase. Although at each level the rubric has several descriptors for cohesion and coherence, the raters mainly scored it based on their own perceptions. The negotiation phase revealed that to discuss organization and coherence of ideas, the raters referred to the scale descriptors displaying these features and subsequently exemplified directly from the essays to justify their scores. The trend of negotiation reveals that in addition to discussing the concepts of overall progression and logical organization of the ideas mentioned in the rubric, most of their discussion was devoted to determining the scale score that classifies the writer's performance. For example, in the following excerpt, the raters are discussing the overall organization of an essay.

K: I gave it 5 because it presents relevant information with some organizations, but… I think it lacks the overall organization.

H: I guess it doesn't have a good overall progression ……

B: no, it has a sort of overall organization, to say it lacks overall organization is too much!

The raters' verbalizations in the post-negotiation phase revealed that in the absence of raters' negotiations, they mostly attended to the organization, progression and coherence of ideas to score the category of Cohesion and Coherence, so they clearly referred to the scale descriptors to justify their scores. Therefore, in contrast to the negotiation process in which the raters employed dual focus (script and scale) strategy, in the post-negotiation phase, they were only dependent on the scale. For example, rater I assigned 4 to Cohesion and Coherence because of the descriptor for this level.

I: I think 4 is good for cohesion and coherence, just because of 4.1. the ideas are not arranged coherently and there is no clear progression.

Language-focused Features

A group of raters was identified as the language-dominated raters, especially in independent ratings. In their verbalizations, they showed a tendency to attend to the language features more frequently. They mainly commented on lexico-grammatical features. The scoring approach they adopted was to get a visual inspection of the scripts through reading the scripts, discerning the lexico-grammatical errors and determining the level of severity of the errors. Then based on the errors, they judged the scripts and assigned scores. This trend was especially observable in prenegotiation sessions.

In negotiation sessions, the raters relied on the analytic rubric and treated the errors and lexico-grammatical features as the rubric dimension of Lexical Recourse and Grammatical Range and Accuracy. It was evident the raters could distribute their attention over all the rubric categories and essay features in negotiation sessions, the length of discussion over various categories differed significantly though. In the post-negotiation phase, like the prenegotiation phase, the raters tended to get a visual inspection of the script but besides commenting on the severity of the lexico-grammatical errors, they analyzed the content and ideational features of the essays, consequently to assign a score they were more faithful to the rubric.

Grammatical Complexity. This feature is one of the subcategories of the criterion of Grammatical Range and Accuracy in the rubric. The analysis of think-aloud data in the prenegotiation phase revealed that only some of the raters addressed grammatical complexity in their verbalizations. Though their perception of this feature seemed inaccurate initially, analysis of their interactions revealed that raters G and H had a major role in drawing other raters' attention to this feature. The trend of interactions showed that eventually, the raters considered it as one of the determining factors, besides grammatical errors, to score Grammatical Range and Accuracy. In the following excerpt, in her verbalization in the pre-negotiation phase, rater G analyzes the grammatical complexity of the essay and counts the syntactic varieties and instances used in the essay while scoring independently.

G: Alright, he has used complex structures in this essay, such as present perfect, past perfect and passive structures such as these two cases of passive structure or these three instances of present perfect.

Although it was attended by a few of the raters in the pre-negotiation phase, the most accurate manifestation of this language-related feature was observed in the negotiation phase which continued to the post-negotiation phase. Qualitative analysis revealed that superficial language-focused features were easily dealt with by the novice raters (e.g. grammatical, lexical and spelling errors) because of being more tangible and easily discernable compared to other writing features. Therefore, in the negotiation groups, after automatizing ideational and rhetorical features, the raters initiated to dwell on more complex language-related features in detail. Among those was the complexity of grammatical structure which was not directly tapped on by the raters in the early sessions of negotiation. The analysis of raters' interactions showed that the raters had a different and somehow inaccurate perception of this feature initially, but eventually, they tended to modify their perception and could approach the notion of complexity of grammatical structure more accurately.

In the following excerpt, the raters are discussing the notion of complexity of structure and whether the syntactic structures used in the essay are complex. This excerpt shows that at first the raters provided a superficial definition of grammatical complexity and just mentioned diversity of structures (verb tenses) as an example.

H: What is the definition of complexity of structures?

A: Using future tense, past perfect tense, present progressive tense,

But later, they revised their definitions and included multi-clausal sentences.

I: There is no complex sentence in this essay, all the sentences are simple.

H: yes, he used them, look, there is an "if clause" here.

I: There is just one conditional phrase and it's incorrect…. the second clause is missed, If I do exercise…what about the second clause?! There are a lot of basic sentences, just he used simple present tenses.

Frequency and Gravity of Grammatical Errors. The analysis of qualitative data revealed that some raters tended to give higher weight to the errors of grammar in their independent rating sessions. Hence, despite using an analytic rubric, for them, erroneous language-related features were more influential in assigning low scores than ideational and rhetorical-focused features. Thus, as language-dominated raters, they exhibited a typical scoring behavior while scoring independently; they tended to identify the grammatical errors and edit them and then assign scores on other domains based on the frequency and gravity of such errors. A large number of erroneous structures led to low scores in all the other categories. In the following excerpt, in the post-negotiation phase, Rater I is referring to the low scale level immediately after pointing to the grammatical errors.

I: This essay is full of errors, I can hardly understand this paragraph, so I start from the 2nd scale of rubric [reading the 2nd scale of Task Achievement category].

Raters' verbalizations showed that while rating independently, language-dominated raters tended to refer to the grammatical errors more frequently, so based on the number of errors, they judged the quality of essays and decided the scale levels in other categories accordingly. Initially, they commented on the errors and some tended to edit them; as a result, the gravity of errors determined the scores on other categories. Treating the categories of the rubric as a chain, one of the raters of this class went too far by stating in the interview that "*low accuracy affects directly the performance on other domains*".

Another scoring behavior employed by the language-dominated raters was scanning the whole script in the independent scoring sessions. In the pre-negotiation phase, Rater G stated in the interview that, after identifying and editing lexico-grammatical errors, she scanned the whole script to see the degree of severity of the errors, then assigned a score for grammar.

G: Ok, there were many grammatical errors…. The sentences don't have the correct structures, I tried to find the errors and corrected them. I scanned the paper to find how many corrections I had done then I assigned the score.

In the negotiation and post-negotiation phases, besides maneuvering over the grammatical errors, the raters analyzed the language of the scale descriptors to determine the range of severity of errors. In the following excerpt, Rater A stated that by counting the number of grammatical errors he could determine the level of severity of errors.

A: If we count the errors of grammar, a score of 3 is good because the grammatical errors predominated and distorted the message.

It is worth mentioning that in contrast to the initial independent scoring sessions in which the raters' only criterion to score the Grammatical Range and Accuracy category was the frequency and gravity of errors, in negotiation and post-negotiation phases, they treated grammatical errors as one of the subsets of the criterion, not the only one.

Lexical and Word Formation Errors. While scoring independently in the pre-negotiation phase, the raters used one heading to refer to both grammatical and lexical errors; therefore, they treated syntactic errors and lexical errors as one component although they were needed to be considered distinctly based on the rubric. The lexical errors are explicitly mentioned in the category of Lexical Resource in the analytic rubric, but in the prenegotiation phase, they disregarded this feature as one of the subsets of Lexical Resource criterion. However, while editing the essays, the raters corrected the lexical errors. On the other hand, in the negotiation and post-negotiation phases, they noticed the lexical errors and scored them under the heading of Lexical Resource based on the rubric.

Lexical Diversity. Qualitative data analyses revealed that the raters attended to lexical diversity in all phases, although most of them encountered difficulty in recognizing the range of diversity of the lexical items, especially in the independent scoring sessions. One of the descriptors of the Lexical Resource category emphasizes the ability to use a varied range of vocabulary items. To exemplify lexical diversity, the rubric uses some vague phrases (as stated by one of the raters in the interview) determining the range of lexical items e.g. very limited, extremely limited, basic. In the prenegotiation phase, in the absence of a well-defined criterion for the novice raters to evaluate and determine the range of diversity of lexical items, they showed the tendency to rely on their intuition, consequently treated this feature inconsistently and cautiously though. For example, in the interview Rater H defined the basic words as those which the elementary students understand easily; therefore, there is no need to look them up in a dictionary. Another rater stated that to evaluate this feature and determine the range of vocabulary commands of writers, he used to count the uncommon words.

H: He used a limited range of vocabulary, I think a score of 5 is appropriate, I'm glancing the paper to find the good words …. 3 to 4 words… the good words are those which are one level above the basic and common words…. then limited words …, I have to keep them in mind.

Thus, in the prenegotiation phase, the raters were struggling to make sense of the scoring rubric and the essays; they mostly paid attention to the errors as the most vivid realizations and representations of the textual features. But in the negotiation phase, they made an attempt to determine the level of diversity of the lexical items collaboratively. Thus, during collaborative scoring, they allotted more time to decipher the language used to describe lexical diversity in the rubric.

C: What about the range of vocabulary items? Lets' talk about it and then we can make our mind.

I: He used common vocabulary items.

A: They are enough but not well developed….

In another excerpt, in negotiation sessions, Rater I is arguing whether the writer used a wide range of vocabulary or not. Initially, she wants to assign 8 for the Lexical Resource category but by referring to the 1st descriptor of the given category explaining lexical diversity, changes her mind and assigns 7, because she believes that the writer did not use a wide range of vocabulary but sufficient range of vocabulary.

I: Look, he used a wide range of vocabulary, … It is not wide!!! fluently, flexibility to convey precise meanings, I don't agree with 8 (score), maybe 7 is good, he uses a sufficient range of vocabulary. Look at the lexical items of the script…. [referring to the vocabulary items].

In the post-negotiation, some of the raters declared that scoring this feature was still demanding, but comparing to the initial phase, they relied more on the criteria to justify their assigned scores.

Spelling Errors. In the independent ratings, analysis of the verbal protocol and interview data showed that to score a trait, the raters tended to weigh some of its features such as spelling errors differently. The rubric points to the spelling error as one of the descriptors of the Lexical Resource category, but the raters did not attend to this feature consistently. This trend was especially observed in the initial independent scoring session, for example, the following excerpt shows that the spelling errors were overlooked, under the shadow of ideational features.

B: This essay is one of the best ones, the writer used a good style, he could develop the ideas, the ideas are original and he could connect the ideas… he used the complex words. There are some spelling errors, but it's not fair to assign a lower score.

In the negotiation sessions, the raters treated the spelling errors inconsistently, while some pointed to the spelling errors to lower the score of lexicons, others considered spelling errors negligible. For example, in the following excerpt, the raters lowered the lexicon score and assigned 7 because of some spelling errors. They argued whether the spelling errors are rare or occasional (these descriptors were used in the rubric to describe the spelling errors).

B: 8 is high, there are some errors in spelling…, for the writing at this level [score 8], we don't expect to see spelling errors, I'll give him 7, the spelling errors are not rare, they are occasional errors! All the descriptors of score 7 are identified.

C: yes, there are at least 2 to 3 spelling errors in his text.

But in another group, the raters who scored the same script overlooked the spelling errors and scored the lexicon in terms of lexical diversity.

Mechanics. Capitalization, indentation, and punctuation were the features more or less considered by the raters in different phases. Among the features of mechanics in the analytic rubric, punctuation was only mentioned explicitly as one of the subsets of Grammatical Range and Accuracy. In the prenegotiation phase, the raters approached the feature of mechanics inconsistently. For example, in the interview, one of the raters stated that he did not lower the scores of those who did not observe indentation rules, but those who observed the rules received higher scores.

H: None of the writers capitalize the first letters, it seems that they were not familiar the indentation rules. Most of them didn't observe the rules but I considered the positive points for those who observed the indentation and marginalization.

In the negotiation sessions, the raters tended to rate mechanics as it was defined in the rubric. On the other hand, in the post-negotiation phase, most of the raters showed the tendency to point to the punctuation errors occasionally. Hence, they treated all the subsets of Grammatical Range and

Accuracy rather similarly. Seemingly, their doubts on how to score different subsets in this category were somehow resolved. For example, in the following except, rater B is verbalizing her scoring process, commenting on the punctuation problems besides other features. Only her comment on punctuation problems is excerpted.

B: … This word is not capitalized… [reading the essay] he didn't use punctuation marks …. I can't read this part…why full stop? It doesn't need full stop here, if so, the next word must be capitalized…oh no…

Non-scale features

These features were not mentioned in the scale but were observed in raters' verbal protocols, interactions, and interviews. The following features could mainly be considered a reflection of raters' prior knowledge about writing rather than the categories of the analytic rubric.

Handwriting. The illegibility of the essays was approached differently by the raters in the prenegotiation phase, while some raters considered it as a penalty and lowered the scores (there is no relevant corresponding descriptor showing this feature though), some others approached it leniently and struggled to make sense of illegible handwriting. In the following excerpt, Rater K insisted on lowering the score because of scribble handwriting.

K: What handwriting! It isn't beautiful at all, it's not my responsibility to read such messy handwriting, I'll lower his score.

The analysis of the raters' interactions in negotiation sessions demonstrated that the raters did not have any consensus to include bad handwriting as an error. Some attributed it to one of the descriptors of the lexical Resource category saying that the errors distorted the message and justified that the illegibility could be categorized as an error, causing difficulty for the raters and readers, but some others believed that, that is the raters' problem! The analysis of verbalizations in both pre and post-negotiation phases revealed that there was no consensus among the raters to include illegibility as an error. Some tried to read the words or asked for help, some left the illegible parts unread and lowered the score. The following excerpt shows that raters B and G categorized illegible words as spelling errors so lowered the score while Rater A (the last line) ruled it out and just relied on the rubric.

B: His handwriting really affects my judgment, I can't read some words, then, illegibility is a kind of error. Isn't it?

G: Yes, they are errors, I gave it 5.

B: Yes, I'll change the score to 5 too, because it makes some difficulty for the reader, then I think it's an error [referring to the 2nd descriptor of band 5 of Lexical Resource category].

A: no, we don't have such a thing in the rubric…. misspelling is different from bad handwriting.

Comprehensibility of Ideas. Another content-focused aspect of writing addressed by the raters in the initial independent scoring was the comprehensibility of ideas, although this was not mentioned clearly in the analytic rubric. In fact, comprehensibility stems from a lack of observance of other features mentioned in the rubric such as organization, and grammar and lexicon. Raters pointed to this feature in different occasions; for example, when they were scoring cohesion and coherence, one of the raters believed that lack of logical organization among the sentences distorted the overall message of the text; consequently, raters had difficulty understanding the text. In another negotiation session, the raters argued that the frequency of grammatical and lexical errors distorted the comprehensibility of the essay.

L: I read it again, I didn't get the overall message.

I: Wow, it has many grammatical errors, I don't understand what he is talking about!

Authorial Voice. Although the rubric does not have the relevant descriptor for this feature, some raters considered this feature of significance while rating independently. They emphasized that to write an argumentative essay, the writer needs to use a strong voice to convince the readers. This feature is an umbrella term encompassing different aspects of writing such as textual features that enable the reader to interpret, evaluate and organize the information (Crismore, Markkanen & Steffensen, 1993). Although there were some individual variations in attending this feature, those who addressed this feature in the prenegotiation phase had a significant role in shifting the direction of negotiation sessions to evaluating the quality of writing and richness of the language of the essays. These individual variations could be partially attributed to the lack of explicit mentioning of this feature in the analytic rubric.

In the post-negotiation phase, during the raters' verbalizations, this feature was detectable too. The following quotes exhibit that in the negotiation phase, these raters were tackling with the richness of language employed by the writers to convince the readers.

G: He couldn't state the good ideas with the strong language, just stated some of his teaching and personal experiences.

C: yeah, it's not well written, I don't like such style, he just told a story……

Contextual features

One of the most frequently observed idiosyncratic strategies was the essay comparison which had an undeniable effect on the scores assigned by the raters. In all the phases, they vividly pointed to this issue while rating independently or collaboratively. In the interview, one of the raters mentioned that they ought to compare the essays, to assign accurate scores. Another rater stated that she unconsciously compared the essays, although she was aware that she should not do so. This was significantly detectable in the independent scoring sessions (pre and post-negotiation phases). Consequently, in the absence of assistance from other raters or when they had doubts, they resorted to this strategy to resolve any ambiguity. For example, in the interview one of the raters stated that "when I scored a pile of essays, I looked back to check whether they were written in the same level, I see that most of them were in the same level, then I didn't change the scores". Using the same strategy was observable in negotiation and post-negotiation sessions. The following excerpt shows that in a negotiation session, the raters were discussing the ideational features of an essay. To justify her assigned score, Rater G resorted to essay comparison while her fellow rater confirmed it.

G: *Comparing to the previous essay, the writer provided more tangible reasons to support the main ideas, I think 5 is a good score*

H: *Yes, we assigned 4 for the previous script, this one is so much well-written.*

Another contextual factor, the raters attended to in all the three phases was the writers' language proficiency. In the following excerpts, raters are discussing how the language proficiency of the learners can likely affect the sore they assigned in one of the negotiation sessions.

H: I don't know the level of test taker's language proficiency?

K: is it important? We only need to consider whether the writer could fulfill the task

H: But if I know their levels, I could justify my expectations to their language proficiency levels

K: For IELTS the students must be upper intermediate students….

In the pre-phase, raters repeatedly commented on the language proficiency of writers. For example in the following excerpt, Rater G is commenting whether the writers are qualified to participate in IELTS.

L: His English is poor, as far as I know, IELTS is for those whose language proficiency levels are intermediate or above.

(Post)negotiation-specific Features

In the previous section, we explained the features attended by the raters in the three phases of the study; however, there existed features that were specifically addressed by the raters in just negotiation and more specifically post-negotiation phases. Table 3 exhibits these features as scale features, non-scale features, and contextual features.

Table 3.

Features in Negotiation and Post-negotiation Sessions

Textual and contextual features		Subfeatures
1-	Scale features	Cohesive devices
		Generic Features
2-	Non-scale features	Length
		Using 1st person pronoun anecdotes
		Accuracy of idea
3-	Contextual features	time limitations rater's attitudes

Scale features

These features included cohesive devices and generic features that were overlooked by the raters in the prenegotiation phase; however, raters were found to reflect on these features while scoring Cohesion and Coherence and Task Fulfillment Categories respectively in collaborative scoring sessions.

Cohesive devices. The rubric category of cohesion and coherence features using cohesive devices as one of its subsets. While raters discussed this category, they commented on cohesive devices as one of the contributing factors to assigning scores. For example, in the following excerpts, the raters are discussing whether the writer used cohesive devices accurately.

F: He used some basic cohesive devices; the 1st and 2nd paragraphs are not connected with each other. The 1st paragraph talks about an idea and the 2nd paragraph talks about something else, there is no link between them.

B: And also, we have after that … the road in Britain after that… it is not an appropriate conjunction to connect these sentences…. You should take the bus, blah blah, after that the roads should …

A: It's used for sequences.

G: No, that's a cause and effect statement… if X occurs, Y would happen…. this is what the writer meant, but this cohesive device is completely misused.

In the post-negotiation phase, the raters were relatively consistent in attending to this feature.

Generic features. The analytic rubric describes generic structures of the essays in the category of Task Achievement. In the negotiation sessions, the raters analyzed the generic structure by pointing to the descriptors of the given category, addressing the writer's stance toward the question they were posed to and task fulfillment. As the negotiation process provides the raters with a fertile ground to distribute their attention to various features, they focused their attention on the generic features of the essays in the collaborative scoring sessions. In fact, due to the difficulty raters encountered in understanding the categories and the subsets of the analytic rubric, in the prenegotiation phase, the raters just considered the easily discernable content-related aspects and ideational features of writings, but more complex rhetorical features such as generic features were overlooked. For example, in the following excerpt, in the negotiation phase, the raters are discussing whether the writer was successful to establish his position in this argumentative essay.

A: *This essay presents a clear position, two positions are argued; the negative and positive points are argued and in conclusion, he wrapped them up…*

B: *The writer is supposed to write about the pros and cons.*

F: *Yes, pros and cons! The first paragraph talks about pro, the 2nd paragraph must be for cons.*

A: *Yes, the problem is that the paragraph about the cons is missed.*

G: *So, a score of 4 is good.*

It must be noted that according to the verbal protocols, in the post-negotiation independent scoring, the raters' distribution of attention to the complex rhetorical feature was more systematic and consistent.

Non- scale features

The following features are not explicitly mentioned in the analytic rubric, but while scoring collaboratively, raters addressed these features. Most of these features were discussed by the raters in the negotiation phase but eventually, they convinced each other to attend to the scale-related features rather than those not mentioned in the rubric. The examination of raters' verbalizations in the post-negotiation session showed that more or fewer raters tended to attend to the features of the rubric discussed in the negotiation sessions.

Length. The length of the essay is not explicitly stated in the rubric, but in negotiation sessions, the raters referred to the task instruction which explains that the test takers are required to write at least 250 words. This feature caught the raters' attention, specifically when the essays were too short, on the other hand, the multiple page essays did not receive high scores.

In the following excerpt, the raters attended to the length of the essay which was less than expected and then decided to lower the score of Task Achievement category.

A: How many words does an essay have?

1: In task instruction, it is mentioned that an essay should contain about 250 words.

C: But the word number of this essay is less than 250.

I: Ok, the writer couldn't develop the ideas, so we should consider it in Task Achievement.

Similarly, in the post-negotiation phase, the raters attended to this feature while scoring Task Achievement category.

Using 1st person pronoun anecdotes. While scoring collaboratively, some raters attended to this feature and treated it as an error and lowered the scores. By stating that using 1st person anecdotes is rather informal and not appropriate for academic writing, the raters categorized it as an error and reduced points for it. The trend of negotiation showed that initially raters did not have a consensus on how to treat this feature, so there was a tense negotiation over this feature among the raters. But as the raters exhibited a shift toward attending more frequently to the subsets and elements of the analytic rubric, they reached an overall consensus to overlook this feature and stick to those aspects of writing specifically described in the rubric. In the following excerpt, Rater I dissented using this feature in the writing, arguing that it is not an academic style.

I: As far as I know, in essay writings, the writers can't use "I "first-person pronoun", instead they have to use "writer" or the passive structures, I think its style is not academic.

Accuracy of idea. As being unique to the negotiation scoring sessions, this feature did not come up in the verbal protocols in any of the independent scorings of pre or post-negotiation phases. In the following excerpts, the raters are examining whether the ideas brought by the writer are rational.

H: The 1st paragraph explains that people can take buses or taxis instead of driving their cars … this is a solution, isn't it?

G: Yes, so rational, the problem is raised in the introduction and the solution is brought up in the next paragraph, but could the problem of transportation be solved by this?

Contextual Features

In the negotiation phase, the raters assessed the essays by taking account of the contextual factors such as time limitation and rater's attitudes as well. Although they were not mentioned in the raters' protocol, but collectively, they partially influenced the score assigned by the raters. In the following excerpts, the raters are discussing that time limitation might affect the performance of test-takers.

H: I think we should have in mind that the test takers have only 40 minutes to write for the IELTS writing test…

G: Yes, under such time pressure we can't expect them to write a well-organized essay.

In these excerpts, raters are discussing whether they are allowed to consider the writer's personal attitude toward the prompt and their position in assigning scores.

K: I' m going to give him a higher score because what he wrote is what I believe. In my view, definitely teachers' personality is more effective than his knowledge….. but look at the prompt the language of the prompt implicitly directs the students to say the opposite.

H: This is what you believe! We are not supposed to involve our attitudes and personal beliefs in assigning the scores.

In the negotiation and post-negotiation phases, the raters' attendance to the rubric descriptors noticeably increased. The qualitative analysis showed that as the raters gained more control over the features of the rubric in the negotiation phase, contextual factors and raters' idiosyncratic behaviors had less effect on their assigned scores. This was mostly observed in the post-negotiation phase too.

Discussion

We kept track of raters' textual foci in three phases: independent rating sessions conducted prior negotiations, negotiation scoring sessions and independent rating sessions conducted after negotiations. The analysis of the qualitative data revealed that negotiation scoring sessions were influential to direct the raters' attention to a wider spectrum of textual aspects of writing corresponding to the analytic rubric; as a result, it might lead to higher validity and reliability in scoring.

In the pre-negotiation phase, because the raters were not familiar with the scale and wordings of the descriptors, they focused on the script and exemplified from the essays to justify their scores. Unlike Barkaoui's (2010) novice raters who tended to depend on rating criteria to score the scripts, the raters of this study did not attend to the features of the rubric, and instead relied on their perception of writing and the features seemed important for scoring writing. Their verbal protocols and interviews in the initial phase of independent scorings revealed that in the absence of any training and experience in rating, they had no way but to rely on their own perception of how rating should be done. This finding is in line with Cumming's (1990) study, in which novice raters tended to assess the essays by relying on their general reading abilities and their prior knowledge in editing. Similar logic is discussed by May (2009) in studying raters' rating behavior.

In terms of raters' attention to textual features, in the pre-negotiation phase, the raters attended to an array of textual features; however, they did not pay equal attention to all the textual features. The think-aloud data showed that the rubric did not affect much the textual and essay features the raters attended to. Thus, it appeared that the application of the analytic rubric did not result in attending to all the displayed features of the rubric. The raters' behavior in prenegotiation scoring sessions is discussed below:

First, the raters paid more attention to the gravity and frequency of the lexico-grammatical errors. They typically did not separate lexical errors from grammatical errors, treating them as one category while they were required to separate them as two distinct features under two different headings, based on the analytic protocol. This could be attributed to their lack of experience in rating second language writing. The literature has indicated that novice raters tend to focus on the local and discernable aspects of writings (Barkaoui, 2010; Cumming, 1990; Sakai, 2003).

Second, although they attended to the overall quality of the essays, they had imprecise and inconsistent perceptions about the features they addressed such as cohesion and coherence. Thus, they defined them based on their inner criteria (perception of rating criteria) rather than relying on the rubric descriptors.

Third, the raters tended to weigh some features over others. To score the subsets of the analytic rubric categories, they did not follow a consistent scoring approach. This is in line with previous findings conducted in independent scorings (e.g., Charney, 1984; Lumley, 2002; Sakyi, 2003; Vaughan, 1991). The qualitative results revealed that the raters took an either-or approach to treat the subsets of the category; that is, all of the subsets of a category were not given equal weight (e.g., Barkaoui, 2010, Lumely, 2005; Smith, 2000; Vaughan, 1991). It is clear that in the pre-negotiation phase, the features of the rubric salient to the raters when they awarded higher or lower scores to the essay were mainly reflections of their scoring perception; as a result, they treated the categories of the rubric inconsistently and imprecisely. For instance, in their verbal protocols, raters overlooked the errors in mechanics or somehow grammatical complexity as the subcategories of Grammatical Range and Accuracy criterion or treated spelling errors inconsistently. Even the features not included in the rubric were salient to the raters such as authorial voice, handwriting, and illegibility. Although the focal point of this study was to explore the textual features salient to the raters, we noticed some idiosyncratic behaviors adopted by raters such as comparing the essays when encountering difficulty and considering learners' language proficiencies. These findings lend support to the literature (Barkaoui, 2010; Lumely, 2005; May, 2009; Smith, 2000; Vaughan, 1991) indicating that the raters tend to refer to other criteria rather than those mentioned in the rubric. In addition to relying on their scoring perception, sometimes the raters referred to the analytic rubric functioning as the justification for the assigned scores, meaning that they found the supporting evidence for their ratings in the analytic rubric descriptors. In other words, at times they pointed to the rubric features to validate their assigned scores.

On the other hand, in negotiation and post-negotiation phases, analysis of the introspective comments and raters' interactions showed that they tried to have a logical and accurate scoring by exemplifying directly from the scripts and referring to the scale descriptors frequently. This finding is consistent with dual focus (focus on both text and scale) explained by Lumely (2002). Referring to scale descriptors is also stated as a typical rating behavior of expert raters (Barkaoui, 2010). In negotiation scoring, raters exemplified from the scripts to justify their scores and present concrete evidence for their scores. This finding confirms Lindharden's (2018) finding about the positive potentiality of negotiation through which raters can validate the scores they assign.

In terms of raters' attention to the textual features of writing, the overall trend of negotiation revealed that the raters gained more control over the essay features and the analytic rubric while scoring collaboratively. The features discussed and negotiated in negotiation sessions were shadowed in the post-negotiation phase too, although individual variations were detectable in terms of the textual aspects they attended to. The findings suggest that compared to the pre-negotiation phase, the raters distributed their attention more evenly over a wider spectrum of essay features. This reflects Lindhardsen's (2018) findings that through negotiation, with wide involvement of the members of the groups, the raters balanced their attention more evenly in negotiation scoring sessions. The rating behaviors of raters in negotiation and post-negotiation sessions observed in the current study resemble experienced raters' rating behaviors reported in previous studies like Cumming's (1990) and Barkaoui's (2010). As raters gained expertise while participating in the negotiation scoring process, they were able to balance their attention and shift their attention from lexico-grammatical features to various textual features.

As the raters worked collaboratively to discuss ideational and rhetorical features, their comprehension of these features and the corresponding categories in the analytic rubric increased. While the raters exchanged their ideas in negotiation sessions, they had plenty of time to cross-examine reasoning and evidence. It could be argued that in the prenegotiation phase they did not have a clear understanding of some textual features or somehow did not consider them important. But negotiation sessions provided them with opportunities to learn and get more experienced. Thus, comparing to the pre-negotiation phase, a wider range of ideational features such as idea development, task fulfillment, the relevance of ideas, completeness, originality and accuracy of ideas were discussed. This could be likely attributed to the increased awareness and knowledge of raters in evaluating writings and using the scale while rating collaboratively which was also observed in their rating behaviors even in post-negotiation sessions.

Compared to the prenegotiation phase, the trend of negotiations reveals that the raters in the absence of an expert rater, could scaffold their co-raters to decipher the language of the rubric, initially considered challenging, and gain a better understanding of the textual features. The negotiation process seemed to have drawn the raters' attention to the construct of coherence and cohesion, a writing aspect that they did not notice initially and mentioned inconsistently when rating independently in the prenegotiation phase. This finding was also reported in Barkaoui's (2010) study. Moreover, because the rubric contains specific descriptors for lexical diversity and grammatical complexity about which the raters had difficulty interpreting in the initial phase, the negotiation process could resolve the dilemma and facilitated their understanding. Thus, in the negotiation phase, raters paid more attention to these correspondent features and tried to decipher the language of the rubric describing such features.

In the negotiation sessions, most of the discussions were devoted to clarifying the key phrases describing writers' levels of performance on each category. The strategy, they employed was pointing to the essay features and then matching them with the relevant descriptors in the analytic rubric. The most negotiated items were the terms discriminating the scale levels related to each category such as "noticeable", "adequate", "limited", etc. Therefore, in the absence of professional training sessions, the negotiation sessions seemed to function as a training for the novice raters on how to use and interpret the analytic scale by analyzing the relevant descriptors and providing evidence from the essays. Moreover, they gained a better understanding of the critical textual features, thus negotiations helped them notice more features and consider them more logically. This can also confirm the positive potentiality of the negotiation scoring sessions in reaching consistency in attending to different aspects of writing.

There were some ideational and language-related features attended by the raters in negotiation and post-negotiation phases which were unique to these phases, such as cohesive devices, generic features and first pronoun anecdotes. This indicated that the negotiation process has the positive potentiality to draw the raters' attention to these rhetorical features of writing, with which the raters seem to be less familiar. Furthermore, as the raters gained more control over rating writing aspects, they attended to features not mentioned explicitly in the rubric such as length (as evidenced by Barkaoui, 2010) which could be attributed to the importance the raters place on fluency by treating short paragraphs as the writer's incompetency to deliver the intended message. In addition to the aforementioned features, they considered some contextual factors such as time limitation and writers' attitude toward the writing prompt. This finding lends support to Barkaoui's (2010) finding that the experienced raters showed the tendency to focus on non-scale features as well.

Conclusion and Implications

Having a vague understanding of the whole rubric, employing idiosyncratic strategies, ignoring some features of the writing skill, overweighing some textual features over others, and having an inaccurate and imprecise understanding of the features of the rubric are, as the findings of the present study showed, indices of novice raters' behaviors which could endanger the accuracy of scoring. For example, the raters in this study were found to ignore some of the rubric features or to overweigh some features over others while rating samples of writing.

The textual features which were considered by the novice raters in this study were not limited to any specific type. Rather they included features of different natures including organizational features, language-related features, voice, comprehensibility, and even handwriting. They were categorized in this study as ideational, rhetorical, language-related, non-scale, and contextual features. Negotiation scoring sessions aided the raters to attend to more features in line with the rubric categories.

It can be implied from the findings of the current research that through negotiation novice raters are able to attend to a wider spectrum of textual features that are usually overlooked or overweighed in individual ratings. They could obtain a more comprehensive perspective in evaluating EFL writing samples when they attend negotiation sessions.

Furthermore, negotiation can aid raters to reconcile the categories of the rubric with different aspects of writing samples while rating. This means they will have clear guidelines for rating which direct them to rate within the framework of the rating rubric, and as a result, score the writing samples more precisely.

The findings may have implications for rater training as well. The complexity of rating performance skills and the necessity for rater training has been well documented in the literature (e.g., Davis, 2016; Lumley, & McNamara, 1995; Papajohn, 2002). So, in the absence of expert raters to train novice raters, for example in EFL contexts, negotiation can be employed as an effective technique to improve raters' understanding of the rubric and their rating behaviors (Ahmadi, 2019; Trace et al., 2017). The present study showed that this can be achieved by raising awareness about the features of the rubric and the relevant descriptors. Kim and Lee (2015) also reiterate that rater negotiation is useful for preparing benchmark essays and materials for rater training sessions and revising testing materials and that this technique could be employed as a rater training technique per se.

Finally, this study was an exploratory study exploring the features novice raters attend to in rating before and after receiving negotiation. While the study lent support to the positive potentiality of negotiation in helping raters attend to various scoring features in line with a rubric, further studies are needed to explore whether the training effect created through negotiations would last for a long time or the raters would return to their previous rating behavior. Moreover, how the group composition can affect the quality of negotiations requires further investigation.

References

Ahmadi, A. (2019). A Study of Raters' Behavior in Scoring L2 Speaking Performance: Using Rater Discussion as a Training Tool. *Issues in Language Teaching*, *8*(1), 195-224.

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, 13, 341–358. https://doi.org/10.1080/15434303.2016.1236797

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience, *Language Assessment Quarterly*, *7*(1), 54-74.

Broad, B. (1997). Reciprocal authorities in communal writing assessment: Constructing textual value within a “New politics of inquiry”. *Assessing Writing*, *4*(2), 133-167.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah: Utah State University Press.

Cambridge University Press. (2015). *Cambridge IELTS 10: Authentic examination papers from Cambridge ESOL*. New York, NY: Cambridge University Press.

Cambridge University Press. (2016). *Cambridge IELTS 11: Authentic examination papers from Cambridge ESOL*. New York, NY: Cambridge University Press.

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.

Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement*, *36*(1), 29-45.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory (4th ed.)*. SAGE.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication*, *10*(1), 39-71.

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

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision* *making, and development of a preliminary analytic framework.* TOEFL Monograph Series, Report No. 22.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86* (1), 67–96.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken *language*. *Language Testing*, *33*(1), 117-135.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing*, *26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221

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

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

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341-366.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly: An International Journal*, *2*(2), 117-146.

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing*, *20*, 37–52.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239-261.

Kim, S., & Lee, H. K. (2015). Exploring rater behaviors during a writing assessment discussion. *English Teaching*, *70*(1).

Lim, J. (2019). An investigation of the text features of discrepantly-scored ESL essays: A mixed-methods study. *Assessing Writing*, *39*, 1-13.

Lindhardsen, V. (2018). From independent ratings to communal ratings: A study of CWA raters' decision-making behaviors. *Assessing Writing*, *35*, 12-25.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276.

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

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397-421.

Moss, P., Schutz, A., & Collins, K. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, *12*(2), 139–161.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, *36*(2), 219–233.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors* [Unpublished doctoral dissertation]. The University of Toronto.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. *Studies in immigrant English language assessment*, *1*, 159-189.

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing*, *34*(1), 3-22.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp Lyons (Ed.). *Assessing second language writing in academic contexts* (pp.111–125). Ablex.