Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, 2023, 12(2), 87-106 (Previously Published under the Title: Maritime English Journal)

"Others have scored the same as me; do you want to change it"? Exploring rater dominance in negotiation rating sessions of EFL writing

[1]Leila Hajiabdorrasouli*
[2]Alireza Ahmadi

Research Paper IJEAP- 2308-1990 DOR: 20.1001.1.24763187.2023.12.3.6.8
Received: 2023-07-23 Accepted: 2023-09-20 Published: 2023-09-30

Abstract: Rater negotiation is a score-resolution method through which raters review and discuss performance samples to resolve rating discrepancies. The success of this method depends on raters getting equally engaged in negotiations. This study explored whether novice raters remain equally engaged in negotiations or rater dominance occurs. Eleven English teachers attended eight negotiation sessions. They scored ten writing samples independently using the IELTS rubric and then discussed rating discrepancies in groups. It has employed a mixed-methods approach to see whether any traces of rater dominance are observed or raters are equally engaged in negotiations. The chi-square test results for score changes indicated that only in Task Response category, raters were inequitably engaged. No dominance was observed for other dimensions. However, qualitative analysis of the negotiations revealed various patterns of rater dominance. Furthermore, the analysis of rater interactions in negotiation sessions indicated that rater dominance is a nonlinear construct demonstrated in interactions of raters during negotiation rating sessions. The findings illuminated that while some raters attempted to scaffold each other to form a unified understanding of scoring rubric by sharing the floor in discussion sessions, some tried to dominate other raters. The findings highlight the utility of negotiation, not just as a resolution method but a procedure with training effects for performance assessment in EFL contexts where access to expert raters is usually limited.

Keywords: Negotiation, Novice Rater, Rater Dominance, Resolution Methods, Writing Assessment

Introduction

Discussions of score validity and reliability in writing assessment prevail in L2 assessment research. Actually, in performance assessment, which usually requires multiple raters, raters' subjectivity can engender variability in test scores, and consequently, errors of measurement will be introduced (Yan, 2014). To reduce rater subjectivity and increase consistency in scoring, researchers have employed rater training (e.g., Davis, 2016; Papajohn, 2002; Sweedler-Brown, 1985; Weigle, 1994), benchmarks (Popp et al., 2003), and rating rubrics, preferably analytic rubrics (Johnson et al., 2000, 2001; Jonsson & Svingby, 2007). However, rater subjectivity and discrepancies in scoring are still major concerns in writing assessments. Thus, resolution methods are practiced to resolve discrepant ratings. Johnson et al. (2000) categorized resolution methods into four models: tertium quid, expert judgment, parity, and discussion. In the tertium quid model, to moderate any extreme scores, a third rater (adjudicator) conducts a blind review of the samples and forms the operational score by selecting one of the original scores and averaging it with their

[1] Member of the Faculty, l_rasouli@iauba.ac.ir; Department of English Language, Bandar Abbas Branch, Islamic Azad University, Bandar Abbas, Iran.

[2] Professor of Teaching English as a Second Language, arahmadi@shirazu.ac.ir; Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran.

awarded score. In the expert judgment model, an expert who has more expertise in scoring will help resolve the discrepancy. Unlike the tertium model, the original scores are replaced by the expert's score. In the parity model, the judgments of all the raters, including the third rater, are equally weighted. Thus, the original scores assigned by the novice raters and the expert's score are used to form the operational score. Finally, in the discussion method, raters come together to discuss and resolve discrepancies in scoring. No expert rater is involved in this method.

The discussion or negotiation method was first employed in contexts where access to trained raters was limited. In fact, due to practical constraints, it might be a challenge to implement rater training in some assessment contexts, then to aid raters to resolve discrepant scores, the use of negotiation among raters is suggested. In the absence of expert raters, this technique allows novice raters to validate their scores against each other. Thus, they can develop professionally (Lindhardsen, 2018), especially in the EFL contexts where access to expert raters is limited (Ahmadi, 2019, 2020; Smolik, 2008; Trace et al., 2017). In the other resolution methods explained above, the presence of an expert to resolve discrepancies is pivotal. In the negotiation method, in the absence of an expert rater, when the raters' initial scores differ, they engage constructively in exchanging ideas, reviewing the essays, and scoring rubric to reach the consensual score (Broad, 1997; Johnson et al., 2005; Moss, 1996). This method has been viewed as a paradigm shift in writing assessment, attributed to the hermeneutical paradigm where the raters can be influenced by each other. In this hermeneutical approach, differences are appreciated as an existing reality, but raters are not left to themselves and expected to stick to the standards recontextualized in the local assessment context.

Consequently, the ratings would benefit from the synergy produced by many raters working together (Lindhardsen, 2018). To have a successful assessment through negotiation, raters must remain equally engaged in the process of negotiation (Moss, 1996). While some studies report the efficiency of this method in minimizing score variance (Johnson et al., 2001, 2003, 2005), the underlying assumption that raters must negotiate equally and critically without relying on the assertive and dominating voice (Johnson et al., 2005; Moss, 1996) requires investigation.

Like in any dialectic context, negotiation as a rhetorical and political process may undeniably be influenced by conversational power relations shaped by cultural, racial, social, and gender differences, resulting in more dominating and assertive voices while deferring other voices. Some voices are democratic, while others tend to be more autocratic (Broad, 1997; Moss & Schutz, 2001). If raters engage equally during negotiations, no dominance effect would likely occur; meanwhile, if a majority of negotiation scores agree with the original scores of one of the raters, the trace of score dominance would be identified, which negates the primary assumption of the negotiation method. On the other hand, in the case of not having equal expertise, the raters tend to defer or dominate (Johnson et al., 2005); thus, the problem arises when other raters tend to score the same as the original score of the dominator rater, which could result in low accuracy and validity of the score resolved through negotiation (Johnson et al., 2005). As such, the present study aimed at investigating rater dominance in assessing writing.

Literature Review

To resolve score variance and increase the reliability and validity of the assessment, the literature recommends that different score resolution methods be used. In their study, Johnson et al. (2001) explored the effect of five resolution methods on the reliability of the final operational score. The methods included 1- reporting a single score by summing the discrepant ratings; 2-reporting the score awarded by the expert; 3- reporting the score awarded by the expert combined with the original scores from the raters; 4- using discussion among raters to reach a consensus score; 5- reporting the score awarded by the expert combined with the closest score of one of the raters. They found that the score assigned by the expert combined with the original scores from the raters produced the most reliable score.

In the same vein, Johnson et al. (2005) suggested that despite many score resolution methods being available to assessment users, the reliability and validity of the scores depend on the choice of the resolution method. The findings of their study revealed that negotiation improves the accuracy and reliability of scoring, but they suggested that assessment professionals must take steps to make sure that raters understand and observe the process of negotiation. The study found that rater dominance existed when raters employed a holistic rubric but not when they used an analytic rubric, a finding that was repeated in a recent study by Trace et al. (2017). They investigated the distribution of scores in the negotiation method while raters used an analytic rubric. The results of chi-square tests demonstrated no evidence of rater dominance.

In Trace et al.'s (2017) study, they made an effort to indicate how negotiated scores were affected by one rater; therefore, they computed a series of chi-square tests for each pair of raters on five categories of an analytic rubric and adopted the procedure used by Johnson et al. (2005). The chi-square tests were used to compare the distributions of how raters 1- keep their original scores 2- keep the score of another rater 3- assigned a mediated score. The results of chi-square tests demonstrated no evidence of rater dominance.

Previously, Moss et al. (1998) had found traces of rater dominance when negotiation was employed for scoring portfolios. In a qualitative study, they tried to document the negotiation quality between a pair of raters. One of the aspects of their research was to identify whether the raters participated equally in negotiations. Their observations demonstrated that some participants' reading and writing roles resulted in their dominance, which undermined participating coequally in the process of interpreting candidates' performances. Lindharsen (2018) studied rater dominance by investigating whether raters contributed equally in the negotiation session. She explored rater dominance in terms of the number of words and the decision-making behaviors of raters. The results indicated that the differences ranged from 5.4% to 38.6% regarding the number of words. Also, concerning decision-making differences, the average differences ranged from 3% to 30.9%. Lindharsen concluded that there was not much difference between verbosity and the number of decision-making behaviors of raters in their pair collaborative scorings. This indicates that there was no rater dominance in negotiation scoring sessions.

Also, Lindhardsen (2018) compared the relationship between score dominance and conversational dominance. In her study, the score dominator was a rater whose original score was chosen as the final score, and a deferent rater was a rater whose score was further away from the final score. The score dominator produced 204.9 words and 9.9 decision-making behaviors, whereas the deferent produced 216 words and 11.2 decision-making behaviors. The results showed a minimal difference between the number of words and the number of decision-making behaviors in score dominance and deferent. The results showed that score dominance is not the result of conversational dominance; thus, the raters are equally engaged in this study. She postulated that when one rater dominates, that is not the result of conversational dominance but different perspectives. One of the few studies on rater dominance in negotiation explored the phenomenon of rater dominance in the speaking assessment context. Using negotiation in the EFL context as a resolution method, Ahmadi (2020) investigated rater dominance by analyzing the raters' turn-takings and discourse. Findings indicated that although rater dominance obviously existed in negotiations, it could not be easily traced in turn-takings, amount of speech, and changes in scoring.

The literature suffers from a paucity of research on resolution methods, particularly negotiation. While a few studies conducted in this regard have focused explicitly on the efficiency of negotiation in improving consistency of scores, the issue of rater dominance which is against the underlying assumption behind negotiation and can ruin the intended consequences of negotiation, is only peripherally investigated in these studies. Rater dominance has been a minor objective of such studies. Furthermore, it has been studied quantitatively regarding change scores or counting turn-takings. The complexity of the rater dominance construct and its latent nature must be deciphered from the raters' interactions in negotiation sessions. In addition, rater dominance has been studied in pairs when two novice raters negotiate to resolve

discrepancies. Finally, few studies have focused on how dominance may emerge and function while novice raters interact with each other.

The present study has focused on rater dominance in the negotiation method as the main objective to fill such gaps in the literature. It has employed a mixed-methods approach to see whether any traces of rater dominance are observed or raters are equally engaged in negotiations. Moreover, the study has focused on groups of raters in negotiation to see how different raters with different perspectives and individual differences behave in this interactive and pluralistic assessment approach and how they may benefit from the synergy produced in groups. Finally, the study is conducted in an EFL context where novice raters usually have limited access to expert raters and training programs. The following research questions were accordingly targeted in this study:

Research Question One: Do novice raters participate equally in the process of negotiation?

Research Question Two: What interactional patterns are found in the novice rater negotiations?

Methodology

Design of the Study

The present study adopted a concurrent triangulation mixed-method design (QUAN + QUAL), in which the data are collected quantitatively and qualitatively concurrently. Then two sets of data are compared to identify convergence, differences, or combination of both if any, usually mixing is conducted in the interpretation or discussion section (Creswell, 2014). In the present research, the quantitative analysis was conducted via counting the indices of changing or maintaining the original scores, which helped to recognize the rater dominance in each category of the analytic rubric. In contrast, in the qualitative component, the interactions of raters in scoring sessions were analyzed to provide patterns of raters' engagement.

Participants

Novice Raters

The participants were 11 MA students of Teaching English as a Foreign Language, including two males and nine females, aged 25 to 39. They were already familiar with the theoretical aspects of language teaching and testing as they had passed the MA program's relevant courses. So, they had similar educational backgrounds. Moreover, they had similar teaching experiences, teaching English in language institutes for 2 to 5 years. None had received any rater training or had experience in rating. Usually, teachers do not receive formal training on rating language skills in the Iranian EFL context. Hence, they usually resort to their knowledge rather than standardized rubrics for the rating (Ahmadi & Sadeghi, 2016). Thus, the participants served as novice raters in this study. Finally, the participants voluntarily participated in the study and were paid for their participation.

Expert Rater

An expert rater was asked to instruct the novice raters on scoring the writing samples based on the IELTS rubric in the norming session. He was a Ph.D. candidate in TEFL and had eight years of experience rating the IELTS test and its mock version.

Instruments

Writing Samples

10 writing samples derived from Cambridge IELTS 10 (2015) and Cambridge IELTS 11 (2016) were used for negotiation rating sessions. These samples were IELTS candidates' essays written based on the Academic Writing Task 2. In negotiation scoring sessions, to train the novice raters to match a writing performance to one of the band scales of the IELTS and enable the raters to differentiate the larger number of performance levels, writing samples (all argumentative) were intentionally selected from different performance levels.

Analytic Rubric

The official IELTS Writing Assessment rubric for task 2 was used for assessing writing performances. The IELTS scoring scale ranges 0 to 9 indicating the lowest performance level to the highest. This analytic rubric includes four assessment criteria: Task Response, Cohesion and Coherence, Lexical Resource, Grammatical Range, and Accuracy.

Data Collection Procedure

Step One: Rater Grouping

The performance of raters and the level of raters' dominance and deference may be influenced by group structures. Additionally, to have an equal chance of exchanging ideas in scoring sessions and rule out preexisting differences, the raters were randomly assigned to each group of A or B. Group A consisted of 6 raters, and group B included 5 raters.

Step Two: Norming Session

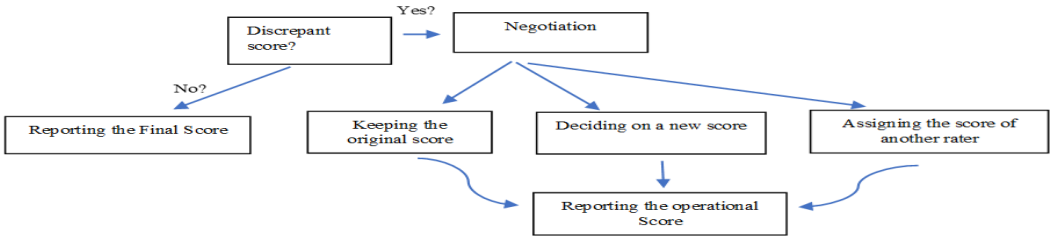
After the random grouping of the raters, a short norming session was conducted**.** This session aimed to familiarize raters with the rating procedure and assessment rubric. The writing preliminaries, essay features, and rubric categories were discussed in this session. Furthermore, an expert rater instructed them on how to score the essays and resolve discrepant scores by reviewing and then discussing the rubric descriptors and features of the sample essays.

Step Three: Negotiation Phase

After the norming session, the groups attended their weekly sessions to rate the essays independently. Both groups received the same samples. The analytic rubric required the raters to report a separate score for each rubric category. To initiate the negotiation process, they shared and reported their independent scores for each essay. If the assigned scores were different, the participants tried to justify their scores and challenge others by rereading the scripts, highlighting different writing features, and referring to the rubric descriptors for each dimension. To keep the autonomy of the raters, not acting as a scoring machine, and to appreciate multiple perspectives in scoring, the raters were free to disagree on the negotiated score and keep their original scores even after negotiation (see Linacre, 2010; McNamara, 1996; Trace et al., 2017). Alternatively, they could revise their original scores and assign the negotiated scores (see fig. 1.). The negotiated score was the original score of another rater or a new consensus score. To value the differences and disagreements among the raters, the accuracies of the scores were not determined. Therefore, the raters themselves determined the success score. In this study, the success score is a negotiated score, determined by reducing scoring variance and the raters' positive views of the rating process. The negotiations continued for eight sessions. Overall, 440 scores were assigned by 11 raters rating 10 essays. All the negotiation sessions and raters' interactions were audio-recorded and transcribed for further analysis.

Figure 1

The Schematic Representation of Rater Negotiations



Data Analysis

In the quantitative analysis, the frequencies of negotiated scores reflecting one of the sources (the original score of the rater, the new score, the score of another rater) for each rubric category were counted. Similar to the studies of Johnson et al. (2005) and Trace et al. (2017), we counted how often the other raters selected the original score of each rater or the new score after negotiation; subsequently, a series of chi-square goodness of fit was run to compare the distribution of scores over different raters and the consensus new score on four categories of analytic rubric in each group.

To examine any effect of rater dominance, the expected null hypothesis was that the scores are all equally distributed across raters and the new scores on each category of the analytic rubric. The alpha significance for the chi-square goodness of fit index was set to a level of 0.05. Concerning four categories of the analytic rubric, it was subsequently reduced to 0.01. To indicate how significant these sources were to the computed value of the chi-square, standard residuals for the raters and new scores on each domain were computed. The formula to determine standardized residual for each source was standard residual = (observed count – expected count) / √expected count (Hinkle et al., 2003).

All the interactions were qualitatively analyzed to get insights into the process of rating and patterns of raters' dominance in negotiations. Open, axial, and selective coding procedures suggested by Strauss and Corbin (1998) were used. Initially, the interactions were read and reread carefully, and the data were coded and labeled. Then the coded data were reread and organized into meaningful themes and subthemes. The relations were checked, and the final adjustments were made. Data triangulation, peer review, and member checks were utilized to establish the credibility of the findings. An ELT expert in qualitative research checked the emerged codes, categories, and subcategories for peer review. To check the reliability of the coding system, 10% of the raters' interaction transcripts were coded. The Kappa coefficient for the inter-rater agreement was 0.7. The coders discussed the discrepancies to reach an agreement on codes, categories, and subcategories.

Results

Quantitative Study

Table 1 presents the descriptive statistics about the raters' score dominance for the four categories of analytic rubric in group A. Table 1 exhibits that some raters tended to dominate other raters in negotiation sessions. Considering the dominant scores, the first column of each category represents the number of times other raters selected the original scores of other raters or the new scores after negotiation. The second column indicates the percentage of times the scores were selected. In the Task Response category, the negotiated change scores most frequently agreed with Rater A's original scores 61.9% of the time (N =13).

Concerning the categories of Cohesion and Coherence and Grammatical Range and Accuracy, the negotiated change scores most frequently reflected Rater A's original scores 22% of the time (N= 9) and 59% of the time (N=10), respectively.

Table 1

Descriptive Analysis of Score Dominance of Group A

Categories	*TA*		*CC*		*LR*		*GRA*	
Sources	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Rater A	13	61.9	7	22	7	26	10	59
Rater B	4	19	4	2.5	4	15	3	17.7
Rater C	5	24	0	0	0	0	3	17.7
Rater D	3	14.5	2	6.5	2	7.5	6	35.5
Rater E	4	19	5	16	5	18.5	3	17.7
Rater F	0	0	2	6.5	2	7.5	0	0
Average scores	2	9.5	5	16	10	37	0	0

As depicted in Table 1, in the Lexical Resource category, the negotiated change scores most frequently agreed with the new score, 37% of the time (N=10). Rater C' scores were not assigned by other raters in domains of Lexical Resource and Cohesion and Coherence (N = 0), while for the Task Response and Grammatical Range and Accuracy categories, the negotiated change scores did not reflect the scores of Rater F in any instances (N = 0).

Table 2 shows the results of the chi-square goodness of fit and standard residuals for Group A. To determine which sources, contribute the most to the chi-square value, the standard residuals of all sources were computed (fifth column). Table 2 reveals that the computed chi-square goodness of fit for the Task Response category is 15.2, which exceeds the critical value of chi-square (x^2 =15.086, *df* =5, α=0.01). Thus, the null hypothesis that the scores are distributed equally across all sources in the domain of Task Response is rejected. As Table 2 indicates for rater A, the computed standard residual is 3, the observed score=13, and the expected score= 5.2. If the value of the standard residual of a source is greater than 2, we expect that its observed frequency is significantly more than the expected frequency. The computed value of standard residual for Rate A indicates that based on the direction of change, the negotiated change scores agreed most frequently with the original scores of Rater A in the Task Response category. Hence, this source significantly contributes to the computed value of chi-square for the Task Response in group A. For Rater B, Rater C, Rater D, Rater E, Rater F, and the new score, the computed values of the standard residual are -0.29, 0.55, -0.29, -1.97, 0.42, 1.13 respectively, all less than 3, hence the score dominant of Group A is Rater A in the Task Response category. However, as shown in Table 2, the computed chi-square did not exceed the critical value of chi-square in other domains. Thus, we retained the hypothesis that statistically, no score dominance was detected in the given domains. Thus, the scores are equally distributed across the raters and new scores on the domains of Cohesion and Coherence, Lexical Resource and Grammatical Range and Accuracy.

Table 2

The Chi-square Goodness of Fit and Standard Residual Results for Group A

Categories	Sources	OBS score	EXP Score	SR	x^2	*df*	*P*
Task Response	Rater A	13	5.2	3	15.2	5	0.009
	Rater B	4	5.2	-0.29			
	Rater C	5	5.2	0.55			
	Rater D	3	5.2	-0.29			

	Rater E	4	5.2	-1.97			
	Rater F	0	5.2	0.42			
	average	2	5.2	1.13			
Cohesion and coherence	Rater A	7	4.3	1.3	4.92	5	0.42
	Rater B	4	4.3	-0.14			
	Rater C	0	4.3	0			
	Rater D	2	4.3	-1.11			
	Rater E	5	4.3	0.3			
	Rater F	2	4.3	-1.11			
	average	5	4.3	0.82			
Lexical Resource	Rater A	7	5	0.89	9.6	5	0.87
	Rater B	4	5	-0.44			
	Rater C	0	5	0			
	Rater D	2	5	-1.34			
	Rater E	5	5	0			
	Rater F	2	5	-1.34			
	average	10	5	2.24			
Grammatical Range and Accuracy	Rater A	10	5	2.24	7.6	4	0.107
	Rater B	3	5	-0.89			
	Rater C	3	5	-0.89			
	Rater D	6	5	0.44			
	Rater E	3	5	-0.89			
	Rater F	0	5	0			
	average	0	5	0			

SR: Standard Residual OBS Score: Observed score EXP Score: Expected Score

In Group B, the descriptive statistics indicate that the rater dominator is Rater I, who dominated 52.17, 37.5, 47.05, and 54.54 percent of the time in negotiated change scores in Task Response, Cohesion and Coherence, Lexical Resource and Grammatical Range and Accuracy, respectively (Table 3). In no case did the negotiated change scores reflect Rater J's scores for the three domains of Task Response, Lexical Resource, and Grammatical Range and Accuracy.

Table 3

Descriptive Analysis of Score Dominance of Group B

Categories	*TA*		*CC*		*LR*		*GRA*	
Sources	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Rater G	9	39.13	3	18.75	4	23.52	5	45.45
Rater H	2	8.69	1	6.25	1	5.88	2	18.18
Rater I	12	52.17	6	37.5	8	47.05	6	54.54
Rater J	0	0	1	6.25	0	0	0	0
Rater K	3	13.04	2	12.5	2	11.76	3	27.27
Average scores	5	21.73	4	25	6	35.29	2	18.18

As Table 4 reveals, the null hypothesis is rejected for Task Response in Group B, interestingly similar to Group A. Thus, the awarded negotiated change scores reflected the scores of one of these sources ($p < 0.1$, $x^2(5) = 19$). The computed values of standard residuals indicate that the negotiated change scores assigned by raters agreed more frequently with the scores of RI. Hence RI (with the standard residual +3) is a rater dominator in the Task Response category. The null hypothesis is retained in the other three domains, indicating no dominance.

Table 4

The Chi-square Goodness of fit and Standard Residual Results for Group B

Categories	Sources	OBS Score	EXP Score	Standard Residuals	x^2	*df*	*p*
Task Response	Rater G	9	5	1.79	19	5	0.002
	Rater H	2	5	-1.34			
	Rater I	12	5	3.13			
	Rater J	2	5	-1.34			
	Rater K	2	5	-1.34			
	average	5	5	0			
Cohesion and coherence	Rater G	3	2.8	0.2	6.64	5	0.24
	Rater H	1	2.8	-1.8			
	Rater I	6	2.8	3.2			
	Rater J	1	2.8.	-1.8			
	Rater K	2	2.8	-0.8			
	average	4	2.8	1.2			
Lexical Resource	Rater G	4	4.2	-0.2	7.81	4	0.099
	Rater H	1	4.2	-3.2			
	Rater I	8	4.2	3.8			
	Rater J	0	4.2	0			
	Rater K	2	4.2	-2.2			
	average	6	4.2	1.8			
Grammatical Range and Accuracy	Rater G	5	3.6	1.4	3.66	4	0.45
	Rater H	2	3.6	2.4			
	Rater I	6	3.6	2.4			
	Rater J	0	4	0			
	Rater K	3	3.6	-0.6			
	average	2	3.6	-1.6			

As Table 4 reveals, the null hypothesis is rejected for Task Response in Group B, interestingly similar to Group A. Thus, the awarded negotiated change scores reflected the scores of one of these sources ($p < 0.1$, $x^2(5) = 19$). The computed values of standard residuals indicate that the negotiated change scores assigned by raters agreed more frequently with the scores of RI. Hence RI (with the standard residual +3) is a rater dominator in the Task Response category. The null hypothesis is retained in the other three domains, indicating no dominance.

Qualitative Study

Through qualitative analysis, four patterns of raters' engagement were observed in interactions:

Sharing The Floor: Collaborative Scaffolding of Fellow Raters

Iwasaki (1997) defines the floor as the conversation unit; he argues that a conversation would be an open floor if all participants collaboratively developed it. In this study, this pattern of engagement was identified as sharing the floor. The salient feature of such collaborative interactions frequently observed in this study is scaffolding fellow raters. All the six members of group A had ample chance to defend their scores and comment on the other scores; they were willing to demonstrate flexibility and adaptability to confirm the scores of other teammates and reach the consensus score. Donato (1994) used the term "collective scaffolding" in groups where the members cooperate, rely on their resources, and scaffold each other to

resolve. Scaffolding occurs in the context of collaboration; thus, in this study, the term collaborative scaffolding describes where the participants are willing to engage with each other's ideas and complete each other's utterances; they "engage critically but constructively" (Storch, 2002, p. 130). There was no identifiable expert in this negotiation scoring group; hence, when scoring collaboratively, the raters pooled their resources to resolve ambiguities. Many requests and provision of information were observed in this pattern of talks. Seemingly this flow of giving and taking created challenging negotiation, which aided co-raters to co-construct their understanding of the rubric and components of writing.

One of the negotiation team members, a self-assigned leader (Rater C), tried to keep the discussion moving by providing logical arguments, asking other members to present proof and contribute to the discussion. While in most cases being submissive to the scores assigned by other raters, Rater C could affect other raters' scores in disagreement instances. Being depicted as the score dominator in the Task Response category based on score change indices, Rater A demonstrated the ability to convince others by providing logical arguments, although merely attempting to compete for the floor. He talked as much as other raters did but also aided the discussion development by providing smart hints, raising intelligent questions, and seeking answers, which created the tendency in most of his co-raters to change their scores to square with those of him.

The following excerpt from the negotiation of raters on scoring sample 2 on the Lexical Resource category illuminates how negotiation aids raters to reconstruct meaning out of the rubric jointly. While some raters cannot score the sample confidently, Rater C, with the aid of Rater A, actively strives to scaffold their co-raters to overcome their doubts and score the sample. Rater A and C have scored differently from other raters; they have scored 4, whereas others scored 5 and 6. To score the Lexical Resource category of this sample, all raters focus on the number of lexical errors and severity levels of the errors and whether the errors cause *some difficulty* or *strain* for the reader (7-21). Rater B, who has initially scored 5, reads the second descriptor of score 5 as her justification (2). To justify their scores, raters A and C review the lexical errors, believing that spelling and word formation errors obscured the meaning. In collaborative negotiation, the raters scaffold each other by completing each other's utterances (Storch, 2002) which is evident in lines 3 and 4; to complete the justification raised by Rater C, Rater A pinpoints some errors that are likely to impede the message. Rater B, who tentatively has assigned a score of 5, focuses on the level of difficulty which the errors may cause (2 and 8). In the same vein, to justify score 5, Rater D includes bad handwriting as one of the sources of errors (8). The dispute is over the difference between the words *strain* and *some difficulty* used in the second descriptors of bands 4 and 5 of the rubric, respectively (8- 11). To aid his co-rater in distinguishing the meaning of *strain* and *difficulty*, Rater A relies on some examples to clarify the notion of *causing strain* for the score of 4 (10). Rater B, who is still hesitant to decide between scores 4 and 5 from the beginning, ponders over the errors again (11). After that, Rater C provides a similar explanation for why score 4 is more appropriate and attributes the level of severity of errors to the amount of *strain* or *some difficulty* they produce for the reader, which eventually leads to the score change by Rater D. She prefers to score 4.5 since the rubric allows the total score, she regresses her score to 4 (12-15). Subsequently, triggering doubt about the accuracy of the assigned score, Rater B reminds herself how the errors distorted the message and then provides some examples of word choice errors. These talks resemble Vygotsky's self-speech (1962), having a more psychological function that scaffolds individuals cognitively, but she is still reluctant to change her score and expresses her doubt about scoring 4 (19). To scaffold rater B, Rater C reviews the lexical errors made by the writer and explains the severity of those errors (20). Rater B, who is doubtful from the beginning, tentatively accepts that the errors cause strain for the reader and changes her score (21). One can observe the high amount of equality which is the main feature of collaborative interactions in raters' negotiation. Van Lier (1996) defines equality as equal control over the distribution of the task, not just equal distribution of turns. As evident in the subsequent negotiation, the raters have equal control over the task where the alternate views are discussed and offered; therefore, according to Storch (2002), such collaborative negotiations lead to resolutions that seem acceptable to participants.

Excerpt 1

1 RC: But between 5 and 4, I think 4 is more appropriate.

2 RB: Ok, he used [reading aloud the 2nd descriptor of band 5] did you count the errors?

3 RC: No, there are so many of them, I can hardly figure out the meanings.

4 RA: Here, *in to do something to control*? What's this? *To form a house*!

5 RC: It doesn't make sense; dictations and punctuations are awful.

6 RA: Is this *consumption*?

7 RB: I think, 5 is good,… the errors caused *some difficulties* for the reader to get the message

8 RD: Because of his handwriting, I scored 5, It caused sort of *difficulty for the reader*.

9 RB: There are many errors, but not too many *to distort the message*, so I think 5 is appropriate

10 RA: It is more than *some difficulty*. Sometimes we hardly get the meaning, these errors *distorted the message.*

11 RB: For example, in this paragraph, I got the message, but I didn't get this word.

12 RC: For score 4, the rubric says that *the errors cause strain for the reader*, so there must be many errors blocking the message, but for score 5 the errors cause *some difficulty* for the reader. The word *strain* is much more intensifying than the word *difficulty*.

13 RD: I changed my mind; I wish I could give him 4.5.

14 RA: We don't have such a thing in the rubric.

15 RD: So, I score him 4.

16 RA: Ok, do you want to change?

17 RB: mmm, I think, 5... the rubric says having *noticeable errors in spelling*. It is difficult. Isn't it?

18 RD: Yes, especially his handwriting.

19 RB: He used *a limited range of vocabulary*……. like this part…. I'm not sure.

20 RC: I don't get the message. Does it make *some difficulty* or *strain*?

21 B: Yes, like this part…aha... Ok, I think 4 is appropriate.

The quantitative analysis did not depict rater dominance in some rubric categories, like the Lexical Resource category. It demonstrated that raters scaffolded each other in deciphering the rubric descriptors to assign scores in this category. As Hajiabdorrasouli and Ahmadi (2020) explored, to assign a score in this category, novice raters' knowledge and inner criteria were their primary and leading resources; relying on these resources, they scaffolded each other to co-construct meanings from the rubric. The equal collaboration of the raters in scoring this category reflects that the raters' background can potentially remove traces of dominance and deference in the negotiation sessions.

The Power of Cooperative Rater Coalitions

Because several raters participated in negotiation sessions rather than only a pair of raters, some instances of coalition formation were observed among the raters assigned the same score. Thus, the power of the coalition was undeniable in assigning the resolution score. The fellow members of the coalition tried to convince other individuals to change their scores. The dominating power increased if Rater A or sometimes

Rater C in Group A or Rater I or sometimes Rater G in Group B were fellow coalition members. The coalition members contributed jointly to engage with each other's utterances and completed each other's suggestions by forming a unity to convince another party. In this form of negotiation, there is no individual dominant rater tying to compete for the floor. Instead, the pro-raters create a synergy collaboratively to direct the negotiation flow.

Excerpt 2 illustrates this pattern where the raters of Group A could convince the leader of the group collaboratively to change her score. In an attempt to resolve discrepancies in the Task Response category, Rater A, B, and D have scored the sample similarly (23-25), then they make a coalition and, by providing different pieces of evidence, try to convince Rater C. Rater C, the self-assigned leader of the group has assigned a lower score to this sample. In a joint attempt to convince Rater C, Rater A pinpoints one of the descriptors of score 6 as the justification (23). Rater B provides a similar justification for scoring 6, focusing on the positive points of writing (22). Rater D, another coalition member, confirms her justification (25). From the analysis of the raters' interaction, it is evident that the coalition among pro-raters is formed at this point of negotiation. Rater C, who has scored differently, negates their justifications and elaborates on why this sample deserves 4 (27). Rater A asks for the evidence of irrelevancy, rater C (the opposing rater) refers to the *cost of cigarettes* as one of the clues claiming that it has nothing to do with the *traffic* (29), whereas Rater B and D believe that it is just one of the solutions raised to the problem. This does not satisfy Rater C, insisting that the writer has stated the repetitive ideas and could use more related examples (28-32). Rater C focuses on keywords in the first and second paragraphs (33). Rater B clarifies the point of confusion (34). To aid their ally, two other coalition members, Rater A and D, complete Rater B's justification by stating that these are two different solutions brought in two different paragraphs. Rater C accepts this justification (34-43). The discussion goes on. Eventually, Rater C is convinced that the examples are relevant and changes her score, not to theirs but 5, insisting on being repetitive. This is welcomed by other raters (44-45). The following excerpt illuminates the discussed issue.

Excerpt 2

22 RB: I think he was successful in the Task Response category, He could manage the ideas and put them into different categories, but the conclusion is weak.

23 RA: Yeah, the conclusion is very short, but I think he *addressed all parts of the task*, so I scored him 6.

24 RB: Yeah, it's 6.

25 RD: I agree with you, 6.

26 RB: The introduction is perfect, we get the main ideas like *cars, traffic, population.*

27 RC: I started from band 0, then band 4, I score him 4 because …. *these are difficult to identify and maybe repetitive …*

28 RA: Why is it *repetitive*?

29 RC: Look at these keywords, like *the cost of cigarettes*! it has no relationship with *traffic. H*e could use some relevant examples

30 RD: It's just an example.

31 RB: Yes, he said if the costs of cars and cigarettes increase, people will buy less.

32 RC: He could use some more relevant examples. Why *cigarettes*? Ridiculous!

33 RC: He used *buses, government vehicles*. There are some repetitive words in both paragraphs.

34 RB: That's a solution, something that government should do.

35 RC: In the second and third paragraphs, we have *transportations.*

36 RB: He says what that government should do.

37 RC: The writer talked about one solution in two paragraphs, that's what I mean by repetitive ideas.

38 RB: Yeah!

39 RA: In the first paragraph, he's talking about *banning*, in the second paragraph, about *being free of charge*, two different reasons.

40 RC: To resolve traffic, people can use public transportations, then the government should provide people with public transportations, it is repetitive!

41 RD: Something that government should do, two different things.

42 RA: And what people can do, using public transportations.

43 RC: Aha, got it, but I think he *presents some ideas, but they are limited and not sufficiently developed*.

[later in discussion]

44 RA: Ok, they are *relevant but limited*, look at the 3rd and 4th paragraphs; we see many repetitive ideas, well! In my idea 4, mmm! I can change it to 5 but I'm sure it's not 6.

45 RA: Fine, 5.

Dominator Rater: Holding the Floor

"When one participant is found to be developing and controlling a topic in a given floor, this person is the floor holder" (Iwasaki 1997, p. 664). The researchers took notice of asymmetric patterns in raters' negotiation, which was highly dependent on the level of engagement and control of raters and their roles. The in-depth analysis of raters' interactions in scoring sessions revealed that Rater I exerted assertive power on her fellow raters to change their scores and employed different strategies to exercise her power. At the same time, her co-raters positively or negatively responded. The rater's assertive dominance strategies to exert pressure on other raters' scoring are subdivided as follows:

Taking the Lead to Initiate the Negotiation. Rater, I initiated more interactions in rating courses while other raters tended to be more recipients of communications. She employed different strategies in initiating and maintaining the discussions and, in the so-called "bossy way," exercised different strategies in inviting other raters' contributions, discussing and discussing the rubric and essay elements. Taking the "in charge" approach in initiating and managing the flow of discussion required the given rater to be talkative, giving others little chance to talk.

Excerpt 3 shows that Rater I casts as an initiator of the discussion, followed by other raters' contributions to announce their scores. Following up on their contributions, here Rater I dominates the interaction, comments on the difficulty of scoring in the Task Response category, stresses the level of her fellow rater, and then provides further comments on why she assigned a score of 4 (49-51). As demonstrated in quantitative analysis, this rater was a rater dominator in the Task Response category. This monologue resembles self-talk, more directed to herself rather than other raters. Other co-raters' roles seem more limited and passive. One of the salient traits of asymmetric patterns observed through negotiation sessions is that the dominant rater rarely asked for assistance.

Excerpt 3

46 RH: What's your score?

47 RI: For TA, I gave 4, yours?

48 RH: You are mean! It was good writing; mine is 6.

49 RI: I think giving a score in Task Response is difficult. You speak very well, why are you stressed?

50 RJ: Oh, no, I'm not stressed.

51 RI: Mine is 4, because the prompt is about *transportation*. I have a problem with the word "transport," why not transportation! By the end of the essay, I didn't see any words related to transportation; just he used cars, rivers [a long monologue].

Taking the Lead in Convincing Other Fellow Raters. Interestingly throughout the scoring sessions, Rater I takes the responsibility to convince other raters and actively challenges their justifications. In the case of any discrepancies, when one of the raters assigned a different score, other fellow raters waited for Rater I's taking the initiative to provide justification and explanation, and there was little attempt by other raters to contribute. One of the characteristics of this kind of interaction is the high level of involvement of one of the raters and the inability or unwillingness of other raters to do which distinguishes this pattern of interaction from cooperative coalitions where the groups of raters with similar scores constructively and equally contributed to negotiations to reach consensus. The quantitative analysis indicated that in the Task Response category, the raters of group B mostly assigned the original scores of this rater. Rater I played an influential role in convincing her co-raters to change their scores to hers.

The following excerpt is illustrative of the theme mentioned above. Rater I dominates this course of rating by taking the lead to convince two other raters to change their scores to the score assigned by her and Rater H. Rater I commences discussion (52), assuming their assigned score is valid Rater I tries to provide justifications. At the same time, her co-rater seems content with delegating the authority to Rater I to convince the opposing raters. Without making any effort to initiate the negotiation voluntarily, Rater H contributed only if she was directly asked for her involvement by the self-assigned leader of the group. For example, in line 60, Rater I exhibits her authority over her fellow rater by assigning her a role to read aloud the band scale. Apparently, instead of group interaction, there is a pair dyadic interaction, in which Rater I demonstrates her authoritative role over the opposing rater by imposing her ideas and leading the opposing rater to award the same score as hers. For example, she leads Rater G to change his score to 3 by comparing bands 3 and 4, even if Rater I does not let him finish his sentence (60-63).

On the other hand, the opposing rater (Rater G) acting as a defensive rater submissively defends his original score rather than challenging the score awarded by Rater I (53, 55, 57, 59, and 62). He eventually gives up (65) and changes his score tentatively, evidenced by his frequent use of phatic utterances (59 and 64). Surprisingly during their entire dyadic interaction, other raters keep silent.

Excerpt 4

52 RI: Why did you score 4?

53 RG: Because band 3 of Task Response says [reading aloud the 1st descriptor of band 3]

54 RI: Yes, exactly!

55 RG: No, I think they are relevant.

56 RI: No, you know, these are the IELTS writings.

57 RG: He tried to develop the ideas but was not successful.

58 RI: Then the ideas are not developed.

59 RG: No …mmm …I chose 4.

60 RI: Let's compare 4 and 3, are you listening? Could you please read band 3?

61 RH: [reading aloud band 3]

62 RG: *The ideas are clear but in a minimal way……*

63 RI: When you don't get the meaning, surely, they are vague. Four of us agree with 3, so? I think you'd better change your score. Because band 4 has some other descriptors not describing this essay.

64 RG: mmm I think for example in this scale… mmm…

[later in discussion]

65 RG: All right

Casting as a Tutor in the Group. Implicitly declared by some of her fellow members, the authority of Rater I was legitimated by other raters of the group; accordingly, the floor was given to her to instruct them how to rate the essays, clarify unclear and ambiguous descriptors of the rubric band scales, and sometimes to assist them to read the illegible handwriting. In the following excerpt, Rater I rereads the prompt (66), Rater H and Rater K struggle with the handwriting, seeking aid from Rater I. Then Rater I as a tutor reads the illegible parts and provides an explanation about one of the words (68, 70 and 74) afterward she instructs them how to score (74).

66 RI: So, finished? [reading the prompt]

67 RH: Could you read this part?

68 RI: *Factual*

69 RK: Read this part, please

70 RI: *Illusion*

71 RH: Is it *doltish*?

72 RI: Yes

73 RK: Is it true?

74 RI: Keywords! Reread them, if you finish, we can discuss. Don't forget to write your names.

Urging Other Raters to Change Their Scores and Negating Others' Arguments. The following excerpt exemplifies the dominant pattern of interaction in which the dominant rater does not readily accept most of the suggestions made by the raters. It is worth mentioning that the dominant rater has control of more turns than others, and hence, she challenges other raters rather than being challenged by them. All the raters scored 6 except for Rater I, who assigned a score of 4. Through the negotiation, she could convince two raters to change their scores. Although all the raters contribute to negotiation, the negotiation is not constructive. Instead, the dominant rater demonstrates her authoritative role over the opposing rater by imposing her ideas, frequent disagreements, and sharp objections (82 and 90), subsequently leading other raters to award the same score as hers. To have the raters in line with herself and make them skeptical of the score they awarded, she sticks to one of the descriptors of Band 6 to negate their arguments (76 and 80); when she finds the raters as the hard nuts to crack, resorts to the use of the modal expressions with a certain coercive impact (82) and comparing the bands and their descriptors (84 and 92) to impose her score. Rater K, who is initially not content to award score 4, and even comments on the severity of Rater I (79), becomes skeptical about the accuracy of the score she initially assigned (81) and eventually is stuck under the pressure of Rater I to do so. Believing that assigning a score of 4 is unfair, Rater H agrees to change her score, not to 4 but 5.

Excerpt 5

75 RJ: It *addressed the task in a minimal way*, and the ideas are not mentioned clearly in the thesis statement.

76 RI: You scored him 6 because of these reasons? Did you check band 3 or 5? Or did you check other bands?

77 RJ: No, I started from 2, based on my experience from previous sessions.

78 RI: Then you decided to score him 6! what's your reason? It is so high.

79 RK: I think you are severe, for Task Response, based on the rubric, I gave him 6.

80 RI: [reading aloud the first descriptor of band 6]? Did you use this descriptor as evidence?

81 RK: Can 't it be score 6?

82 RI: No, I think it can't be even 5. It must be 4. You know, that's my ideal, but it's up to you. Did you score him 6? (addressing RH)

83 RH: [reading aloud the 2nd descriptor of band 6] ... Look at this sentence!

84 RI: Is it related to this? What about these descriptors…..? Did you score based on one of them? What about the 1st and 3rd descriptors of band 6? ha?

85 RH: This descriptor [the 1st] describes this essay.

86 RK: It can't be scored 4, because band 4 says *it presents some main ideas but is difficult to identify*, Is it difficult?

87 RI: No, I don't agree

88 RK: It is not difficult because it has major supports and a good body.

89 RH: Yes.

90 RI: No, it's not.

91 RJ: He tried to use some major supports, but the supports are superficially written.

92 RI: Let's check one by one. Are you with us? Ok, let's start from band 6 [reading aloud the 1st descriptor of band 6], but I think this sentence doesn't describe this writing [reading aloud band 5 and then comparing it with band 4] do you want to change it?

93 RK: yes, I do

[later]

94 RH: I think 5 is good.

Deferent Rater

Likely resulting from her self-confidence, Rater I resisted implementing the change in her original scores despite negotiating over disputed ones. While Rater I attempted to control the floor and take the authoritative role in scoring sessions, some of the fellow raters in Group B did not show any sign of participating in the discussions. Not following the flow of raters' interactions, nor raising questions or defending one's assigned score were taken as the signs that depicted a rater as a deferent and passive rater.

In Excerpt 6, in an instance of scoring the Task Response, while the raters of group B concede Rater I's score, rater K who initially refuses to change her score without arguing her reasons for assigning the original score, asks Rater I to convince her (96-102), but at the end, Rater K passively changes her score.

Excerpt 6

95 RI: Others have scored the same as me, don't you want to change?

96 RK: Ok, try to convince me!

97 RI: Step by step, I have to convince you to change it to 5, then 6, and 7. But I start from 7, look, it says [reading aloud the descriptor],

the 3rd descriptor could be identified here, I mean this! that's why I chose 7.

98 RK: Well, I scored it 4.

99 RI: Ok, she said 4, and now 4 says it *responds to the task only in a minimal way.*

100 RK: Yes, in a minimal way.

101 RI: But it is not in a minimal way, …. [a long monologue]

[Later]

102 RK: Ok, I change it to 5.

As indicated in the quantitative analysis, the deferent was the rater who readily changed his scores in favor of others. However, the qualitative analysis identified some typical rating behaviors of the deferent rater who was subservient in the negotiation process by not exchanging ideas. Thus, score changing cannot determine deference or dominance. Also, Rater J, identified as a score deferent through the quantitative analysis, actively participated in the negotiation and had a meaningful role in co-constructing meanings from the rubric with her co-raters. Hence, she seemingly does not fit under the category of rater deference in the qualitative analysis.

Discussion

It is important to cite that because of the limited number of raters (11), the conclusion of this study should be interpreted cautiously. Any conclusion drawn from the quantitative results are attributed to the raters who scored together in this study; thus, we cannot generalize the effect of negotiation on rater dominance across all groups of raters. This study made an attempt to enrich the understanding of rater dominance in negotiation as a resolution method and the definition of the construct of rater dominance in performance assessment contexts. The quantitative findings suggested that in the Task Response category in Group 1, Rater A, and Group 2, Rater I had potentially excessive influences on other raters' scores in the negotiation sessions. However, in other domains, the chi-square results revealed no score dominance. The findings of the domain of Task Response support the findings of Johnson et al. (2005) and Moss (1996), indicating that the raters were engaged in an inequitable process in negotiations.

On the other hand, the qualitative results revealed that rater dominance is a complex construct that could not be only identified by frequency analysis of score changes (Ahmadi, 2020). Hence, some rater interactions patterns were identified that were not disclosed in frequency analysis. In group A, the raters sought to form a unified understanding of the scoring dimensions and collaboratively scaffold each other to construct meaning out of the rubric by sharing the floor. More or less, all of the group members contributed to constructing an understanding of what the different categories of rubric signify and examining different parts of writings to arrive at an operational score. Previous studies have indicated the significance of negotiation in the co-construction of meaning underlying the assessed construct (Ahmadi, 2019, 2020; Johnson et al., 2005; Hajiabdorrasouli & Ahmadi, 2020; Lindhardsen, 2018; Trace et al., 2016, 2017). This concept is captured in the Vygotskian notion of the zone of proximal development too, wherein novices can display skills in the context of socially organized activities that they otherwise would not be able to accomplish on their own.

It is negotiable to what extent the raters have contributed to meaning construction. For example, in group A, Rater C had a significant role in forming an understanding of the rubric and writing samples but displayed the flexibility and adaptation to change her score to a new score or the original scores of other

raters. In contrast, Rater A, depicted in the quantitative analysis as the score dominator in Task Response in group A, talked as much as other raters, but most raters tended to change their scores to his original scores. Attempting to establish her authority as the leader of group B, Rater I kept her original scores in most of the score changes and tended to reject the opposing ideas and regularly countered the arguments raised (if any) by her fellow raters while most of her co-raters surrendered to the power of this self-assigned leader of the group.

A fascinating finding was the formation of the dominant coalitions among the raters as a newly observed pattern of dominant scoring behavior. In the absence of expert training, the novice raters who had assigned a similar score made a coalition and tried to construct their shared understanding of the rubric and convince others to change their scores. Being depicted as a rater dominator in Group B in the quantitative analysis, Rater I exhibited different patterns of scoring behavior in the negotiation session. The quantitative analysis indicated that she could establish her dominating role in the Task Response category when her original scores were reported as final on many occasions. She demonstrated the dominating scoring behavior in other rubric categories. She played a role in authoritatively leading her co-raters, but her dominance did not lead to the score changes because, more or less, her co-raters were more confident on how to score those categories by relying on their inner criteria and personal knowledge. On the other hand, Rater A was depicted as a score dominator in the Task Response category; the raters of Group B tended to legitimate his authority by changing their scores to his. Rater A played a pivotal role in aiding his co-raters in meaning construction by scaffolding them individually or making cooperative coalitions without exhibiting dominating and authoritative scoring behaviors.

The Task Response category encompassing ideational and rhetorical features is grueling for novice raters to assign scores (Hajiabdorrasouli & Ahmadi, 2020). In the absence of an expert rater and being a novice in rating, the raters in this study had difficulty understanding the descriptors of Task Response and matching them to the samples of writing performance (Hajiabdorrasouli & Ahmadi, 2020). They had to refer to other raters to fill this knowledge gap, so they became unequally engaged in interactions and changed their scores. On the other hand, in the qualitative analysis, more equitable rater involvements were observed in scoring the other rubric categories confirming the results of the quantitative analysis. For example, in scoring Lexical Resource and Accuracy, raters showed the tendency to share the floor and scaffold each other.

Equitable and inequitable patterns of raters' involvement in scoring different rubric categories suggest that raters' background and task complexity may likely lead novice raters to form equal or unequal engagement patterns in the negotiation process, share the floor, form cooperative coalitions, and be the individual dominant or deferent rater. Relying on their personal knowledge and discernable aspects of features (Barkaoui, 2010, Cumming, 1990; Hajiabdorrasouli & Ahmadi, 2020; May, 2009) to scaffold each other in making meanings from the rubric or making a dominant rater as a reference point in scoring the complex categories are issues that undeniably affect patterns of equability or inequality of raters' engagements in negotiation scoring sessions. The process of rater training facilitates understanding and assimilating the scale levels (Shaw & Weir, 2007; Hajiabdorrasouli & Ahmadi, 2020).

In addition to the complexity of the analytic rubric, rater dominance is likely affected by a host of other contributing factors such as the personality of raters, cognitive factors, and lack of expertise in scoring. As demonstrated in the qualitative analysis, the differences between Rater A and B in dominating behavior could be attributed to the psychological variables. In other words, psychological variables such as the raters' personalities and preexisting cognitive frameworks can play a role in rater dominance, threatening the accuracy of the scores (Ahmadi, 2019). An argument concerning these potential variables affecting the rater dominance should be accompanied by more pieces of evidence which is beyond the scope of this study. Indeed, our goal was not to solve the rater dominance in negotiation scoring sessions but to illuminate the complex nature of raters' interactions in negotiation sessions and explore this context-bound construct.

Conclusion and Implications

This study revealed that rater dominance is a multifaceted and context-dependent construct manifested in interactions of raters in negotiation. While rater dominance can threaten the accuracy of the scores resolved through negotiation (Johnson et al., 2005), the study found that different manifestations of dominance revealed in the qualitative findings were only partially transferred into the negotiation change scores. Only the Task Response category was flagged with rater dominance from the four assessment criteria. This indicates that rater dominance does not necessarily translate into score changing, and therefore, the accuracy of the resolved scores may not be severely affected by rater dominance, an issue that demands future research for illumination.

Furthermore, the qualitative findings indicated that when novice raters got engaged in sustained negotiations, besides demonstrating dominating behaviors, they co-constructed a shared understanding of the rating process and criteria. This means that the utility of negotiation, not only as a resolution method but also as a helpful procedure with training effects for performance assessment in EFL contexts where professional rater training is usually missing, should be seriously considered. Considering the challenges imposed by employing human raters in performance assessment contexts, the findings of this study are used to support a broader discussion of the need to develop a better understanding of the process of scoring essays, mainly when there is a movement toward the automated scoring of essays. This study employed one task type (Task 2 of IELTS, argumentative essay) for negotiation sessions; further research could investigate rater dominance in other writing task types. Besides, this study focuses on the analytic rubric for scoring the essays; further researches can incorporate a holistic rubric to explore rater dominance in negotiation sessions.

Acknowledgement

We would like to thank all those who helped us in conducting this research.

Declaration of Conflicting Interests

We have no conflict of interests to disclose.

Funding Details

This research did not receive any funding from any agency.

References

Ahmadi, A. (2020). Rater dominance in discussion as a resolution method. *Taiwan Journal of TESOL, 17*(1), 141-165.

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*(4), 341-358.

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.

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

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

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

Creswell, J. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). SAGE Publications.

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

Donato, R. (1994). Collective scaffolding in second language learning. In J. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33-56). Ablex.

Hajiabdorrasouli, L., & Ahmadi, A. (2020). Exploring Novice Raters' Textual Considerations in Independent and Negotiated Ratings. *Journal of Teaching Language Skills*, *39*(2), 43-87.

Hinkle, D., Wiersma, W., & Jurs, S. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Houghton Mifflin.

Iwasaki, S. (1997). The Northridge earthquake conversations: The floor structure and the 'loop 'sequence in Japanese conversation. *Journal of Pragmatics*, *28*(6), 661-693.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*(2), 121-138.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, *18*(2), 229-249.

Johnson, R. L., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education*, *16*(4), 299-322.

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.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement*, *11*(1), 1-10.

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

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

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

Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist*, *19*, 59–74.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher*, *25*(1), 20-29.

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

Moss, P. A., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal*, *38*(1), 37-70.

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

Popp, S. E. O., Ryan, J. M., Thompson, M. S., & Behrens, J. T. (2003, April). *Operationalizing the rubric: The effect of benchmark selection on the assessed quality of writing*. Paper presented at the State and Regional Educational Research Association Annual Meeting, Chicago, IL: ERIC.

Shaw, S. D., & Weir, C. J. (2007). Examining writing: Research and practice in assessing second language writing. *Studies in Language Testing* (Vol. 26). Cambridge University Press.

Smolik, M. (2008, September). *Does using discussion as a score-resolution method in a speaking test improve the quality of operational scores?* Presented at the International Association for Educational Assessment, Cambridge, UK

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*(1), 119-158.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. SAGE.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal*, *74*(5), 49-55.

Thunholm, P. (2004). Decision-making style: Habit, style, or both? *Personality and Individual Differences*, *36*, 931–944.

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

Van Lier, L. (1996). *Interaction in the language curriculum: Awareness, autonomy and authenticity.* Longman

Vygotsky, L. S. (1962). *Thought and language*. MIT Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501-527.