Taiwan Journal of TESOL
Vol. 17.1, 141-165, 2020 DOI: 10.30397/TJTESOL.202004_17(1).0006

RATER DOMINANCE IN DISCUSSION AS A RESOLUTION METHOD

Alireza Ahmadi

ABSTRACT

Rater subjectivity has long been an intriguing topic. The use of discussion as a resolution method is a practical way to reduce this subjectivity. However, the efficacy of discussion depends on whether different raters get equally engaged in it or one rater tends to dominate others. This study investigated whether and how rater dominance occurs in discussion. To this end, three discussion sessions in which five Iranian raters negotiated to resolve discrepancies in rating were analyzed. Findings indicated raters were *unequally* engaged in discussions, so rater dominance obviously existed. However, it did not necessarily display itself in more turn-takings, a higher amount of speech or changes in scoring. The joint construction of discourse was found to affect raters' understanding of the rating criteria and scoring method. This in turn played a key role in how dominance was realized. The findings illuminate the complexity of rater dominance as a highly context-dependent issue.

Key words: rater dominance, discussion, resolution method, speaking assessment, EFL learners

INTRODUCTION

Performance assessment has always been a challenging issue for researchers. Of great concern has been the rater subjectivity in performance assessment. It has been studied in its relation to factors such as rater experience (Attali, 2016; Barkaoui, 2010; Davis, 2016; Isaacs & Thomson, 2013); rating scale (Barkaoui, 2010; Davis, 2016; Eckes, 2005, 2011; Isaacs & Thomson, 2013; Lim, 2011; Lumley, 2002; Schaefer, 2008); task type (Ahmadi & Sadeghi, 2016; In'nami & Koizumi; 2016; Eckes, 2005); educational background (Kim, 2015); rating occasion (Lumley, & McNamara, 1995), and test takers (Eckes, 2005; Kondo-Brown, 2002; Yan, 2014). Such studies have highlighted rater subjectivity as an indispensable part of performance assessment. To reduce this subjectivity, assessment practitioners have employed a number of methods which include (a) implementing rater training, (b) developing rating criteria, (c) identifying anchor papers to function as benchmarks for each proficiency level, (d) identifying inconsistent raters, and (e) using the many-faceted Rasch program to study rater behavior (Johnson, Penny, Gordon, Shumate, & Fisher, 2005).

Despite all the efforts made to reduce rater subjectivity, raters still are found to be subjective in their ratings. Some studies have found variations in raters' subjectivity even after rater training (e.g., Bonk & Ockey, 2003; Eckes, 2005, 2011; Lumley, 2002, 2005; Papajohn, 2002; Yan, 2014). As such, resolution methods are often used to resolve discrepancies in rating and to increase inter-rater agreement. Resolution methods are of two types (Johnson, Penny, Fisher, & Kuhs, 2003): (a) resolution methods that involve a third rater as the adjudicator (tertium quid model, expert judgment model, and parity model), and (b) resolution methods that rely on group discussion (the discussion method). In the tertium quid model the expert rater's (adjudicator's) score is compared with the scores assigned by the two original raters. Then the score which is closer to the adjudicator's score is combined with the adjudicator's score and the average is reported as the final score. In the expert judgment model, the expert's score replaces the discrepant scores reported by the two raters. In other words, it is the expert's score which is reported to the public. In the parity model, all the raters' scores are of the same value; therefore, the expert's score is added to the other scores and then the average score is computed and reported. In the discussion method, raters resolve discrepancies in rating through discussion. In this method, at first, raters score a sample independently; then together they study the sample, express reasons for their scores, challenge other scores, exchange ideas, review different sources of information and finally reach a conclusion (Johnson et al., 2005). The discussion method was "originally adopted due to resource constraints-in particular the lack of trained raters" (Trace, Janssen & Meier, 2017, p. 2). In fact, in contexts where access to trained raters is limited, both rater training and the resolution methods which rely on a trained rater are impractical. The discussion method would be a promising alternative then.

An important point about discussion is that all the raters should make their own contributions; they need to be equally engaged in the process of exchanging and challenging ideas and provide reasons for their scoring

(Johnson et al., 2005; Moss, 1996). This requires raters to have "not only an understanding of the particular rubric” but also “an understanding of the role that discussion plays in the scoring process” (Johnson et al., 2005, p. 143). Raters can through discussion co-construct an understanding of the scoring criteria at each level of proficiency and match a candidate’s performance to the appropriate level. The problem arises when one rater tends to dominate others; that is, the view of one rater might become the dominant view of the group, and therefore the members may change their scores to agree with the original scores of this rater. Two possibilities may happen. One is that the dominant rater is more knowledgeable and experienced than the other raters. So through discussion other raters may most frequently change their original scores to agree with the scores of this rater. This may mean more accuracy for the final scores reported, on the assumption that the more experienced and knowledgeable rater is more precise in scoring. The second possibility is that the dominant rater is less experienced and knowledgeable, so the score reliability may drop. However, even the first possibility, which can lead to higher accuracy of the resultant scores, “runs counter to the intent of discussion” (Johnson et al., 2005, p. 127). The discussion method is meant to bring about additional information about a sample by including different voices from different raters. So it may lose its value when dominance occurs. This was the objective followed in the current study. It aimed at investigating rater dominance in discussions among EFL raters rating speaking.

LITERATURE REVIEW

The studies conducted on the efficiency of resolution methods (Johnson et al., 2003; Johnson, Penny, & Gordon, 2000, 2001) have come up with different and sometimes contradictory findings. The ones specifically focusing on the discussion method have evidenced the efficiency of this method in improving score accuracy (Clauser, Clyman, & Swanson, 1999; Johnson et al., 2005), contributing to the raters’ understanding of the scoring criteria, reducing rater bias, and increasing positive washback (Trace et al., 2017), though contradictory evidence has also been reported (e.g., Smolik, 2008). The success of the discussion method depends on all the raters getting engaged in the discussions and making their own contributions to the rating of spoken or written samples (Moss, 1996); otherwise, one rater may gain the dominant voice and affect other raters’ scoring.

The literature lacks research on rater dominance. The exceptions are two studies by Johnson et al. (2005) and Trace et al. (2017). In both of these studies rater dominance is investigated as a minor objective. This is because the studies have specifically focused on other issues as their major objectives, and rater dominance is only studied as a relevant issue, not as the main purpose of the study. In the first study, Johnson et al. (2005) focused on the efficacy of discussion as a resolution method in increasing the accuracy of scores obtained from writing samples. To this end, they compared scores coming from two resolution methods: averaging the two discrepant scores versus using discussion to reach common ground and report a consensus score. They further tried to explore rater dominance in the discussion group. The Chi-square test results indicated that dominance existed when the raters used a holistic rubric for rating. No dominance, however, was found when they used an analytic rubric. The writers argued that probably the higher cognitive demands of making a holistic decision provide the opportunity for some raters to dominate.

In a recent study, Trace et al. (2017) investigated the effect of negotiation (discussion) on scoring consistency in L2 writing assessment. They also explored how negotiation may improve raters' understanding of scoring rubrics. Finally, the study focused on rater dominance in negotiations. The results provided evidence for the positive effects of negotiation on improving raters' understanding of the rubrics, increasing rating consistency and reducing bias. No evidence of rater dominance was found "suggesting that in revising their judgments, raters are engaged in an equitable process" (p.16).

Obviously, further studies are still required to provide a clear picture of rater dominance in the discussion method. Although the significance of the discussion method in resolving rating discrepancies and the benefits attached to it are highlighted in studies such as Johnson et al. (2005), Smolik (2008), and Trace et al. (2017), the fact that rater dominance may threaten the benefits of this method requires further attention. The two studies explained above found contradictory results concerning rater dominance. Moreover, they failed to provide any insights into how dominance occurs. This is because rater dominance was studied as a minor objective in these studies. In other words, since their primary focus was on other issues such as comparing resolution methods (the first study) or exploring the impact of discussion method on rating consistency and bias (the second study), they were not specifically designed to study rater dominance and the topic was not, therefore, explored deeply. Furthermore, the study of dominance in these studies was limited to a quantitative analysis with score changing as the only index of dominance. Therefore, only the Chi-square test was used to study the frequency of times changes in scoring agreed with the original scores of a rater. A significant difference in the distribution of changes would indicate dominance, but this is a very limited perspective toward dominance. It can be argued that score changing is not the only index of dominance though it is practically the most convenient one needing just a simple frequency analysis. However, to study rater dominance one is required to focus on other indices of dominance such as the number of turns taken and the amount of speech produced by a rater as well. Furthermore, the study of rater dominance can more effectively be served by a qualitative study delving into interactions to see how dominance occurs when raters attempt to co-construct meaning for the rating criteria. Analysis of interactions can reveal invaluable data about dominance and how it influences and is influenced by discussions. Additionally, what is left unnoticed in the studies on rater dominance is how score changing is related to dominance and whether dominance is necessarily reflected in score changing. This was the reason the present study employed a mixed-methods research design to specifically focus on the analysis of negotiations and explore rater dominance with regard to issues neglected in previous studies, namely, turn taking and amount of talk. Finally, it should also be noted that both of the studies reviewed above were conducted on writing; however, the present study explores rater dominance in speaking assessment.

Research Questions

The following research questions were put forward in light of the above discussion:

1. Does rater dominance occur when discussion is used as the resolution method?
2. Can score changing provide evidence for rater dominance?
3. How is rater dominance related to the number of turns taken and number of words produced?

METHOD

Participants

The participants included five Iranian PhD students of Teaching English as a Foreign Language (TEFL). They were all female and ranged in age from 28 to 33. Besides a similar educational and cultural background, they had similar experiences concerning teaching English and rating speaking. Their teaching experience ranged from 5 to 9 years and was mostly limited to teaching English in language institutes. Their rating experience came from rating their students' speaking or writing in class. Concerning their familiarity with internationally-known rubrics, one of them stated that she had used the TOEFL Rubric twice in her classes, and two of them stated that they had the experience of rating candidates based on IELTS in a few preparatory IELTS classes. None of them, however, had ever received any training in rating. The reason for selecting such a homogeneous sample was that rater dominance seems to be unavoidable when raters are of different backgrounds and levels of expertise. So the impetus was to see whether raters with similar backgrounds would tend to dominate in discussions.

Materials

The materials used in this study were speech samples in the form of monologues of about one to two minutes produced by EFL learners of different proficiency levels at Shiraz University, Shiraz, Iran. The tasks used to elicit such speech samples were actually similar to TOEFL independent speaking tasks in which the test takers are given a topic to talk about in a limited time, and their performance is audio-recorded. The samples served as the materials for the discussion sessions as explained in the next section. Overall, 20 samples were rated and discussed; however, since one of the raters was absent for two of the samples, only 18 samples were analyzed in this study.

Rating Rubric

Since the test used in this study was similar to the independent speaking task of TOEFL, the TOEFL iBT's independent speaking rubric was employed to rate the speech samples. Like analytic rubrics, it contains specific rating criteria (delivery, language use and topic development);

however, raters are instructed to make holistic decisions following the general description provided for each level. The scoring scale ranges from 0 to 4.

Data collection and analysis

Before the study, the researcher met the raters to explain the purpose of the study. The details were not disclosed in order to avoid any potential effects on the results. The data for the study came from three discussion sessions in three weeks. At the beginning of the first session, the researcher explained the purpose of the study. Then the raters were briefly instructed about the rating scale they were expected to use. Finally, they were informed about the procedures to be followed in discussion sessions. It was explained to them that at first they need to rate each sample individually and independently. Then in a group discussion they should express their ideas about the performance sample, state reasons for their scoring, challenge others' scoring, review the scoring criteria together, and finally make their decision about the score. Based on the instructions given to them, achieving consensus on a score was not an aim. So, they could change their scores based on the feedback from other raters or alternatively could keep their original scores if they were not persuaded by the discussions to change their scores. Raters followed this procedure for all the samples. Six or seven samples were rated and discussed this way in each session. Each sample took between 10 to 20 minutes to discuss, and each session lasted for about one and a half hours. All the discussions were conducted in English as the participants were advanced users of English. The researcher was present in the first discussion session but did not intervene in the discussions. He was only an observer to make sure that everything went well. As there were no problems with the first session, he did not attend the second and third sessions so that the participants would feel more relaxed when discussing the samples.

All the sessions were audio-recorded and then transcribed. The transcripts were analyzed qualitatively to get insights into the rating process and rater dominance in discussions. The guidelines suggested by Strauss and Corbin (1998) were followed for qualitative analysis. The data were coded and labeled thorough careful reading and rereading of the interactions. Then the coded parts were reread and organized into meaningful categories. Finally, the relationships were checked and the final adjustments were made. Attention was also paid to each rater’s number of turns taken, number of words uttered and score changes in relation to rater dominance.

RESULTS

General Findings

Table 1 below presents the descriptive statistics about the raters' behavior in discussions. The bold numbers indicate dominance based on the relevant index. The final column indicates the direction of change in scoring after discussion. Those raters whose original scores are selected by other raters after discussion are considered dominant.

As the table indicates, rater 1 has had the highest number of turns in rating 11 samples, and overall, she has had the highest number of turns (n=280). She is obviously the dominant rater in this regard making the highest contribution to the discussions. Other raters have been dominant in a few cases, rater 4 in three cases and raters 2 and 3 in two cases. Rater 5 has not been dominant at all when the number of turns is the criterion. The mean values also indicate that rater 5 has a mean of 6.66 turns in discussing each sample. This is very low when the other raters have a mean of at least 11 turns, and the dominant rater has a mean of about 16 turns. So the difference which is depicted here between rater 1 and rater 5 is very noticeable with the other three raters being very similar to each other. The rank-ordering based on the overall number of turns is rater 1, 2, 3, 4, and 5.

When the criterion for dominance changes from the number of turns to the number of words (amount of speech production) a relatively different pattern is observed. Rater 3 is the most dominant rater generating about 30% of the total production and is followed by raters 1, 4, and 2, respectively. The least amount of production goes to rater 5 with about 10%. As there were five raters, a production of about 20% on the part of each rater could indicate that all the raters were equally engaged in discussions, and no dominance existed.

Table 1

Descriptive Statistics about Rater Dominance

Samples	Number of turns					Number of words					Percent of words					Changes in scores
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	
1	22	13	22	**23**	7	327	150	**716**	381	107	19.45	8.92	**42.59**	22.67	6.37	R2 (3 to 4)/**R3 & R4**
2	**13**	8	10	10	6	130	99	**207**	179	81	18.68	14.22	29.74	25.72	11.64	----------
3	9	**23**	21	13	5	214	273	**394**	187	81	18.62	23.76	34.29	16.28	7.05	R2(3 to 2)/R1, R2, R4, R5
4	8	17	15	**18**	7	157	149	**226**	219	56	19.45	18.46	28.00	27.14	6.94	R1(1 to 2/R2, R3, R4, R5
5	23	16	**28**	12	11	218	173	**478**	172	163	18.11	14.37	39.70	14.29	13.54	R1(4 to 3)/R2, R3, R4, R5
6	**16**	12	6	3	10	**165**	104	145	20	44	34.52	21.76	30.33	4.18	9.21	----------
7	**27**	15	10	23	8	**232**	116	96	268	82	29.22	14.61	12.09	33.75	10.33	R3(3to2) R4(3to2)/R1, R2, R5
8	**35**	28	19	25	10	229	263	226	**331**	83	20.23	23.23	19.96	29.24	7.33	----------
9	**12**	5	1	5	6	**120**	24	1	27	43	55.81	11.16	.47	12.56	20.00	----------
10	**10**	8	8	6	5	42	53	**58**	48	19	19.09	24.09	26.36	21.82	8.64	----------
11	**16**	13	9	8	8	146	87	**161**	148	68	23.93	14.26	26.39	24.26	11.15	----------
12	**32**	19	18	16	14	**314**	103	275	172	123	31.81	10.44	27.86	17.43	12.46	R1(3to4)/R5
13	15	9	8	**16**	4	**160**	85	89	145	69	29.20	15.51	16.24	26.46	12.59	R3(3to2) R4(3to2)/R1, R2, R5
14	8	**14**	10	10	4	79	101	**270**	201	48	11.30	14.45	38.63	28.76	6.87	R4(4to3)/**R3**
15	**7**	3	1	2	4	**80**	9	1	20	21	61.07	6.87	.76	15.27	16.03	----------
16	5	5	**7**	3	1	19	33	**46**	31	9	13.77	23.91	33.33	22.46	6.52	----------
17	**7**	5	6	5	3	62	32	101	**91**	44	18.79	9.70	30.61	27.58	13.33	----------
18	**15**	5	12	5	7	**200**	40	193	53	102	34.01	6.80	32.82	9.01	17.35	R1(3to4)/**R5**
Total	**280**	218	211	203	120	2894	1894	**3683**	26.93	1243						
Mean	**15.56**	12.11	11.72	11.28	6.67	160.78	105.22	**204.61**	149.61	69.06	23.33	15.27	**29.68**	21.71	10.02	

Table 1 also reveals that there is not necessarily a one to one correspondence between the number of turns and words; that is, the raters who had more turns did not necessarily produce more speech. More discrepancies are observed between the raters when the number of words rather than turns is considered.

As indicated in half of the cases (9 out of 18) no change was observed because either the scores assigned by all the raters were the same (this happened in eight cases), and after discussions they decided to keep the score, or each rater considered her score to be accurate and did not change it (this happened in one case). Not all the raters changed their scores concerning the remaining nine cases. Actually, the overall rate of change, as indicated in Table 2, was very low (12.22%); that is, the raters kept their original scores 87.77% of the time. The highest resistance to change was related to rater 5 who maintained her original score 100% of the time; in other words, she did not change her score at all. This is very interesting as in most cases she was the rater who was more of a listener type in discussions, and she also had the lowest contribution. More interestingly, the highest change is related to rater 1 who changed her score 22.22% of the time (n = 4). This rater was the most dominant based on the number of turns and the second most dominant based on the number of words.

Table 2

Rater Dominance Based on Score Changing

Raters	Frequency of changing one's original score	Percentage of changing one's original score	Frequency of dominance based on the direction of change
R1	4	22.22	3
R2	2	11.11	6
R3	2	11.11	5
R4	3	16.66	5
R5	0	0	6
Total	2.2	12.22	---

When the results of the number of turns and words are compared, rater 5 obviously indicates no sign of dominance in any of the sessions. Her production in both cases is lower than what was expected. However, quite surprisingly she indicates her dominance when the changes in scoring are considered. In fact, together with rater 2, she is the most dominant rater based on the direction of changes. Table 2 indicates that the discussion change scores most frequently agreed with the original scores of these two raters (54.55% of the time, n = 6 out of 11). Surprisingly, the table also indicates that the changed scores least frequently (27.27% of the time, n = 3) agreed with the original scores of rater 1 (the most dominant rater in terms of the number of turns and words together). Also in 45.46% of the time (n = 5), the changed scores agreed with the scores of raters 3 and 4. Although the differences depicted in Table 2 are overall low, the fact that the results depicted here are not in line with or are even contradictory to those of the number of turns and words should be noted.

An important point is the score level at which discussions occurred. As stated before, in eight cases the raters had 100% agreement on the scores assigned and when they discussed the reasons, they still decided to keep their original scores. Out of the remaining ten cases in which raters had discrepancies in scoring, in six cases the raters disagreed as to whether a score of 3 or 4 should be assigned; in three cases they were hesitant between a score of 2 and 3, and only in one case they were doubtful whether to assign a score of 1 or 2. This obviously indicates that more agreement was found on lower level scores and more disagreement at higher levels, especially the score of 3.

Analysis of Individual Samples

All the discussions were both quantitatively and qualitatively analyzed to see how dominance emerged through discussions. In what follows, three of the discussions are specifically analyzed to provide evidence for the findings of the study. The first sample in Table 1 is a good example of what is usually expected by dominance; that is, the one who keeps the floor is the dominant rater. As the table depicts, rater 4 has the highest number of turns (n=23), and raters 1 and 3 with only one less turn occupy the second position. The number of turns suddenly drops to almost half for rater 2 and to one third for rater 5. So it could be stated that raters 4, 1, and 3 are similarly dominant; however, when the number of words is considered, rater 3 is strikingly the most dominant rater producing 716 words (42% of the whole production). Rater 4 is the second dominant rater producing 381 words, about half of the production by rater 3. Two of the raters have very low production, rater 2 with less than 9% and rater 5 with about 6%. So based on the number of turns and words, rater 3 is the dominant rater and then with a noticeable difference rater 4 is the second dominant rater. The two dominant raters have also scored the sample differently from the other raters. Before discussions, they have assigned a score of 4, whereas all the other raters have assigned 3. So after this long discussion one expects to see a score change from 3 to 4; that is, a change toward the dominant raters' scores. Table 1 indicates that this is the case only for rater 2 with raters 1 and 5 maintaining their scores after discussion. So the huge production by raters 3 and 4 (about 65% of the talk is produced by them) could not convince raters 1 and 5 (the least productive rater) to change their scores. This indicates that dominance cannot be simply explained by the number of turns or words. Neither can it be explained by studying the direction of score changes alone. Actually, rater dominance is a multidimensional issue which cannot be simply defined by a linear relationship among the number of turns, number of words and score changes; that is, the dominant rater is not necessarily the one with more turns or words or the one whose scoring is accepted by the other raters. The following excerpts can shed more light on this point.

Sample 1

> 5. R5: I couldn't decide between 3 and 4 but I think 3 is better because he was not so much fluent.
>
> 6. R1: I had problem with topic development. I think development of idea was limited. He didn't discuss so many ideas. He also didn't elaborate on his ideas. He just mentioned one basic idea; and with delivery I think he had a little bit difficulty with pacing in the speech. That's why I scored him 3 not 4.
>
> 7. R2: me too. Because for the score of 4 you need all delivery, language use and topic development to be all fairly good. To me that was 3 because topic development was a bit problematic. I mean it wasn't enough…
>
> 14. R2: he just touched upon one point.
>
> 15. R1: yes, he could discuss two sides of the argument, but he only discussed only one part very basically…

22. R1: but I thought more elaboration was needed.

23. R5: but he talked just one minute. How could he elaborate?

25. R3: … I think this speaker would have done better if he was given, I mean, he was given more time. That was like introductory part for him. That's more or less the same for all of us, I think…

33. R4: you know… if you consider the words he uses, he is quite proficient. I found him proficient enough. And these pauses weren't for finding the words, he was just trying to express himself. If we consider the positive sides, he did acknowledge the positive side as well that technology's supposed to make our life and the world a better place to live and then he went to one negative side as well. I find him proficient enough.

38. R1: yeah, I changed my mind about pacing the second time I heard it. But topic development …

58. R4: you know if we had 3.5 I would give him 3.5 or 3.75 but I don't think he should be given 3.

59. R1: so you do agree that he is not 4.

60. R2: a little bit less than 4.

61. R4: he is not 3.

62. R2: he is not 3 either, he is not 4. It's close to 4.

All the raters have stated their scores. Rater 5 is hesitant between 3 and 4. She reports 3 as she thinks the test taker is not very fluent. However, in turn 6, rater 1 states that the problem is with topic development and delivery as well. So these two raters have assigned the same score for different reasons. In turn 7, rater 2 agrees that topic development was problematic so a score of 3 is good for this sample. In explaining topic development, raters 1 and 2 believe that the test taker has focused only on a single point, whereas more elaboration is needed (turns 14, 15 and 22). However, this idea is rejected by rater 5 (turn 23) arguing that more elaboration is not possible in the short time of one minute. The same idea is expressed in turn 25 by rater 3. After that, rater 4 (turn 33) argues for the score of 4 pointing to fluency, using non-basic words, and natural pauses. Other raters express similar ideas until in turn 38 rater 1 states that she has changed her mind. The discussion continues and disagreement is still there. In turn 58, rater 4 states that 3 is not fair and the test taker should be given a score between 3 and 4, an idea which is welcomed by rater 2 (turns 60 & 62). This discussion continues for 16 more turns. However, except for rater 2, all the raters keep their original scores.

Sample 12

Sample 12 could be an example of how you say something is more important than what you say. As the table indicates, rater 1 has kept the floor in this discussion by producing the highest number of turns (32) and the highest amount of speech (more than 30% of the whole production). Other raters are quite markedly different from this rater in terms of both the number of turns and words. The difference becomes very large when this rater is compared with rater 5. However, it is very interesting that finally rater 1 is convinced to change her score to agree with rater 5. So based on the direction of change, the least productive rater is dominant.

But why does rater 1 change her score when she is the dominant rater in the discussion? Actually, what happens in this lengthy discussion, including 99 turns and 987 words, is that the meaning of the scoring criteria is co-constructed through discourse. It is through this joint discourse construction that rater 1 notices her wrong perception of language use as a scoring criterion. Of course, rater 5 plays her important role in this regard by creating doubts in other raters about their scoring, but finally it is the co-construction of discourse which creates the final effect. This can be illuminated by analyzing the discussion on this sample.

All the raters have scored this sample 3 except for rater 5 who has scored it 4. In turn 27, rater 3 expresses doubts concerning her score of 3 and thinks 4 could also be a good score for such a performance. But in turn 29 she repeats her score of 3 looking confident. Rater 1 also repeats her score of 3 in turn 28. The discussion continues for 22 more turns during which rater 5 is mostly quiet and just listening to her partners. In turn 52, rater 1 explains why the sample cannot receive a score of 4 for all the criteria, so the overall score cannot be 4 either, until in turn 53, rater 5 triggers the idea of giving 4 to language use which makes rater 1 think about it (turn 54). Here the doubt is created by rater 5, the least productive rater. Then rater 1 begins to ponder the speech sample and tries to match the performance to the rubric description for score 4. Then rater 3 who was doubtful about her score from the very beginning expresses her doubt again (turn 55) and states that 4 can be a good score too. The discussion continues until in turn 60, rater 1 explains that the performance deserves a score higher than 3. Again it is rater 5 who argues (turn 63) why a score of 4 should be assigned to this performance. Her explanation is completed by rater 3 in turn 64, and this is the turning point for rater 1who finally changes her score (turn 65). Following this, the raters discuss whether the overall score could be 4 or not. They exchange ideas for 34 more turns, but finally all of them decide to maintain their scores. Although they are in doubt about their scores, they are not convinced to change their scores either. So this discussion indicates that the rater with the least production through a simple question (turn 53) makes the dominant rater hesitate about her scoring, and then other raters lose confidence in their scoring too. They review the rubric description for each level and try to rematch the performance to the descriptions. Further elaboration by rater 5 and subsequently by rater 3 is enough to make rater 1, who has obviously kept the floor, change her score.

Sample 12

27. R3: I really like to say 4 but emm

28. R1: it's 3.

29. R3: 3.

52. R1: 3 for language use, and for delivery it could be 4. Could it be?

53. R5: I think even for language use it was 4.

54. R1: why not? Why not 4? Because it didn’t have main mistakes and it was fluent, you could easily understand him. Wasn’t it fluent enough?

55. R3: I think 4.

56. R2: his score is 4?

57. R3: yeah, I really cannot accept that…

58. R1: because I cannot…

59. R3:… 3 is not fair…

60. R1: because I cannot score all the three [criteria] 4, I scored him 3. But to me, 3 closer to 4. Three point something.

61. R5: can I explain?

62. R1: yes.

63. R5: I think topic development was very good. Although he compared two situations, he was quite rele, he was quite relevant. And with regard to language use he had a few mistakes, but it was quite intelligible, and with regard to delivery, he was quite fluent.

64. R3: let's not forget, in 4, minor errors are always there. But they do not obscure…

65. R1: sorry, I changed my mind, I first scored him 3, but I changed my mind to 4. …

Sample 8

This is an example of a sample which indicates no score changes after discussion. Rater 3 has assigned a score of 3, whereas other raters have unanimously, though unconfidently, reported a score of 4. So the discussion here is a dialogue between rater 3 and the other raters. Eventually in spite of all the discussions made, nobody is persuaded to change her score.

Sample 8

40. R3: I didn't say 4 because 4, it needs to, I mean, requires all three parts to be rewarded 4 separately. Right? So to me language use, I mean it's not to say that he had fairly high degree of automaticity. I mean

41. R1: at the same time, but it says at the end emmm

42. R3: he was automatic user but emmm

43. R1: errors are noticeable but do not obscure meaning.

44. R2: did it obscure meaning?

45. R5: it was intelligible.

58. R3: that was to attract your attention and he could win it. I mean, if he had another pronunciation he couldn’t win a topic as such and the examiner, I mean like it or not, would be affected the way he had, he was trying to put the words

59. R5: I don’t think so. Because he was quite fluent and…

60. R1: but I do agree with you. I doubt it for delivery.

64. R1: the way he talked was too artificial.

65. R3: yeah.

66. R1: it wasn’t natural.

67. R2: ok but still he had control over what he said.

68. R3: … 4 at least is to demonstrate a person who is near native or at least has got ability of natural speaking but this person was that odd in his performance that I couldn’t believe that if he was talked to another time, another topic, or something when he was, let's say, off this exam session he could have speak, I mean, spoken this topic the same.

69. R4: that's the reason I said. You know if… as if he had practiced this talk…

70. R1: like he had a preplanned lecture.

Raters 1, 2, and 4 have hesitantly reported a score of 4 for this sample because they think the right score is between 3 and 4, but the rubric does not let them report half scores. Only raters 3 and 5 look confident in their scoring. In turn 40, rater 3 explains that the sample cannot receive 4 because it lacks sufficient automaticity. Rater 1 interrupts her to emphasize that although she had errors they were not problematic as they did not obscure meaning, an idea which is verified by rater 2 (turn 44) and also rater 5 (turn 45) emphasizing intelligibility of the speech. In turn 58, rater 3 believes that the test taker has meant to impress the raters through an exaggerated accent. This idea is welcomed by rater 1 (turns 60, 64 and 66) but is rejected by rater 5 (turn 59) and rater 2 (turn 67) believing that the testee still has control over his speech. In turn 68 again rater 3 repeats the idea that the performance was not natural, that she doubts whether the candidate can repeat this performance if he is given another topic in a different situation. During the discussion, it is repeated several times that the performance seems rather artificial as if the person has memorized a text and is now reproducing it from memory (see Weigle, 2002 for the problem of rating a sample produced from memory). This idea is partially accepted by raters 1, 2 and 4 (e.g., in turns 69 and 70). The discussion continues for 93 more turns. At the end, although the ideas presented by rater 3 are mostly accepted by raters 1, 2, and 4, they are not convinced to change their original scores; neither are raters 3 and 5.

DISCUSSION

This study was conducted in response to the paucity of research on rater dominance. Unlike the few studies conducted on rater dominance which have only considered score changing as the sign of dominance, the present study tried to bridge the gap by focusing on three indices of dominance, namely turn taking, speech production and score changing. Lack of dominance was assumed to indicate itself in a similar contribution of raters both in terms of the number of turns and the amount of speech, and this in turn would mean that discussion change scores would equally agree with the original scores from different raters (Johnson et al., 2005; Moss, 1996). The findings, however, indicated that in most of the cases raters were *unequally* engaged in discussions, so rater dominance clearly existed. However, dominance did not necessarily display itself in more turn taking, a higher amount of speech production or changes in scoring. Analysis of discussions revealed that through interactions raters jointly construct a discourse which affects their understanding of the rating criteria and may accordingly impact upon the way they score a sample. It was found that the relationship among the three indices of dominance is not a simple linear relationship, rather it is a complex function of interactions in the discussion method. It was found that the dominance displayed in the number of turns or words may fail to lead to a score change. Or the change may not necessarily align with the dominant rater's score. In a few cases in the present study even the dominant rater changed her score to agree with the original score from the rater who had the least engagement (this was observed in samples 12 and 18). The three samples discussed above shed light on this relationship.

In the first sample the two raters who had more turns and produced more speech, therefore being dominant, could only convince one of the raters to change her score. However, the other two raters were not convinced by their discussion and maintained their original scores. So, in this case, rater dominance displayed in the number of turns and words appeared in score changing in one case and failed to appear in two cases. In the second sample, the most dominant rater with the highest number of turns and words changed her score to agree with the original score from the least productive rater. The discussion was hot and everybody made her contribution, though the amount of contribution was highly different. Finally, the dominant rater was convinced that the score assigned by the least productive rater was more logical. So considering score changing, rater 5 was clearly the game winner; however, she was not the only player in this game. The final result was a function of all the players playing their role in a complex interaction. So a mere focus on score changes and making judgment on that basis makes one fail to see the complexity of interactions that exist among the players in a certain context. The third case depicted a situation where only one rater was different from the others in her score. After a lengthy discussion which included rater dominance, nobody changed her score. So, dominance failed to indicate itself in score changes. The findings of the study indicated that dominance is a multidimensional issue which is realized differently in different contexts. Thus, considering each of the indices separately will produce an inaccurate picture of how rater dominance is formed and developed.

Several factors may account for the findings of this study. First is the rating scale. Previous studies have indicated that the type of scale affects raters' decision-making behaviors (e.g., Barkaoui, 2007, 2008, 2010; Wiseman, 2008). Unlike analytic rubrics, holistic rubrics require "a rater to consider many criteria when arriving at a score" that represents the construct displayed in the test (Johnson et al., 2005, p. 142). "One might ask whether the cognitive demands in holistic scoring provide more opportunity for one rater to build an argument that overwhelms another rater" (Johnson et al., 2005, p. 142). Li and He (2015) found that raters used different rating strategies depending on whether they used analytic or holistic scales. Furthermore, raters paid attention to different aspects of essays. The effect of rating scale could even be larger than raters' experience on their decision making (Barkaoui, 2010). The raters in the current study pinpointed the issue of scale in several cases. They specifically referred to the fact that the scale did not have enough score categories to enable them to make adequate distinctions among the testees. They noted the need for half scores or more levels to be added to the scale. This problem was more explicit at higher levels, especially the score of 3 in this scale. So the number of score levels is a factor that can affect raters' discussion and may lead to dominance.

The second factor that may have an influential role in rater dominance is the rater's personality; that is, various cognitive and affective factors can create a unique personality type which may make an individual show higher tendencies in surrendering to or rejecting opposing ideas. The literature on rating highlights the significance of personality issues in rating (e.g., Messick, 1984; Thunholm, 2004). Baker (2012) also argues that some rater variability can be explained by individual sociocognitive differences. For example, rater 5, who never gave up in the current study, had in most cases assigned a different score from the others. Although she was often the one who talked the least and was the most passive rater, as indicated in Table 1, she did not change her score even in a single case. Surprisingly, even more agreement was found between the changed scores and her original scores than other raters' scores; that is, although in none of the discussions she was the one to hold the floor, she displayed her dominance in score changes.

In line with personality, raters' "similar expertise" (Johnson et al., 2005, p.141) or perception about "both the expertise and the status" of other raters (p.142) could help explain the findings of this study. The reason why in most cases the raters in this study maintained their scores could be that they had similar expertise and did not believe in the higher expertise of the other raters. This seems logical as they had very similar backgrounds: the same L1 and cultural background, the same educational background (all of them were PhD students of TEFL and had passed the same courses on language testing), similar teaching experiences and low variations in rating experiences. This may have given them the impression that they are not much different and their level of expertise is similar. So they preferred to maintain their scores in most cases, and in a few cases in which they had a change, this change was not necessarily directed toward the dominant rater as the dominant rater was not considered to be of higher expertise. The literature has also referred to raters' background such as relevant education and teaching experience as a source of beneficial effect in rating (e.g., Attali, 2016; Davis, 2016).

CONCLUSION AND IMPLICATIONS

This study could provide insights into the complexity of rater dominance as a highly context-dependent issue. This complexity may arise from the fact that dominance is a social event that is interactively linked to many factors including the type of scale used, the raters' background and personality issues. Therefore to come up with a clear picture of how dominance functions, these factors should become the focus of attention. Furthermore, qualitative analysis of the interactions in the discussion method could help explain how the joint discourse constructed affects raters’ understanding of the scoring criteria and their rating behavior. As such, mere investigation of dominance in terms of quantitative features such as the number and direction of changes or even the number of turns and words cannot provide a clear picture of what is really happening in discussions when dominance occurs. Neither can it explain why dominance occurs the way it does in a certain context. Rater dominance is a multidimensional issue functioning differently and of course *unpredictably* in different contexts. Further qualitative research should investigate how the issues discussed above can specifically play their role in rater dominance and affect the results of discussion as a resolution method.

LIMITATIONS OF THE STUDY

Like any other study, the current study may suffer from a number of limitations. First of all, a small number of raters took part in the study. This may limit the generalizability of the findings and needs to be taken into account. Second, all the raters taking part in this study were female. It would be interesting to explore whether gender can play a role in how and why rater dominance occurs. Finally, for practicality reasons, only three discussion sessions were designed in the current study. Needless to say, increasing the number of sessions and samples to be discussed can provide a more comprehensive picture of how rater dominance occurs.

REFERENCES

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.

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

Baker, B. A. (2012).Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, *9*(3), 225–248.

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

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished doctoral dissertation). University of Toronto, Toronto, Canada.

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.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110.

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

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

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

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

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. (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., 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*, *2*(2), 117-146.

Johnson, R., 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.

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly*, *12*(2), 178–212.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560.

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

Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist*, *19*(2), 59–74.

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

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

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

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.

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

Thunholm, P. (2004). Decision-making style: Habit, style, or both? *Personality and Individual Differences*, *36*(4), 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.

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

Wiseman, C. (2008). Investigating selected facets in measuring second language writing ability using holistic and analytic scoring methods (Unpublished doctoral dissertation). Columbia University, New York, New York.

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.

ACKNOWLEDGMENT

The author would like to thank the editor and anonymous reviewers of the Taiwan Journal of TESOL for their constructive comments to help improve earlier versions of this paper.

CORRESPONDENCE

Alireza Ahmadi, Department of Foreign Language and Linguistics, Shiraz University, Shiraz, Iran
Email address: arahmadi@shirazu.ac.ir

PUBLISHING RECORD

Manuscript received: August 20, 2019; Revision received: December 16, 2019; Manuscript accepted: December 22, 2019.