Issues in Language Teaching (ILT), Vol. 8, No. 1, 195-224, June 2019

A Study of Raters' Behavior in Scoring L2 Speaking Performance: Using Rater Discussion as a Training Tool

Alireza Ahmadi
Associate Professor, Applied Linguistics, Shiraz University, Shiraz, Iran

Abstract
The studies conducted so far on the effectiveness of resolution methods including the discussion method in resolving discrepancies in rating have yielded mixed results. What is left unnoticed in the literature is the potential of discussion to be used as a training tool rather than a resolution method. The present study addresses this research gap by exploring the data coming from rating behaviors of 5 Iranian raters of English. Qualitative analysis of the data indicated that the discussion method can serve the function of training raters. It helped raters rate more easily, quickly and confidently. Furthermore, it helped them improve their understanding and application of the rating criteria and enabled them to justify their scoring decisions. Many-faceted Rasch analysis also supported the beneficial effects of discussion in terms of improvement in raters' severity, consistency in scoring, and the use of scale categories. The findings provide insight into the potential of discussion to be used as a training tool especially in EFL contexts in which lack of access to expert raters can be an obstacle to holding training programs. The author argues for future studies to focus on how discussion may function depending on the rating scale used.

***Keywords**:* Discussion, rater training, L2 speaking, many-faceted Rasch analysis, resolution method

Corresponding author's email: arahmadi@shirazu.ac.ir

INTRODUCTION

Rater subjectivity in performance assessment has always been a source of concern. Two important procedures employed to deal with this issue include rater training before the exam and resolution methods after the exam. In rater training, raters are introduced to the test and scoring criteria. Then individual raters are given sample test responses to rate independently. The ratings are then compared with reference scores (usually the already established scores by the experienced trained raters) and discrepancies are discussed. The purpose is to reduce rater subjectivity and increase rater agreement when it comes to real rating. Unlike training, resolution methods are employed after the exam when the rating is over. The purpose here is to increase agreement by resolving discrepancies in rating.

The significance of rater training as a preventive measure in minimizing raters' bias and inconsistency of scoring is almost well-established in the literature, and the studies conducted so far in this regard have provided evidence for the overall efficiency of training though the evidence has not been conclusive (see e.g., Bonk & Ockey, 2003; Elder, Barkhuizen, Knoch, & von Randow, 2007; Tajeddin, Alemi, & Pashmforoosh, 2011). For instance, in several recent studies, training was found to increase inter-rater reliability (Davis, 2016), help novice, developing and experienced raters develop their rating (Kim, 2015), improve raters' severity and consistency (Lim, 2011), and encourage non-native teachers to reconsider the criteria they attend to in rating speaking (Tajeddin, Alemi, & Pashmforoosh, 2011). Previously, Elder, Knoch, Barkhuizen, and von Randow (2005) had found that providing feedback on ratings can help increase raters' awareness and scoring consistency and reduce bias though individual variations may exist in raters' receptivity to feedback. Earlier studies had also supported the beneficial effects of training on increasing inter- and intra-rater reliability (Shohamy, Gordon, & Kraemer, 1992; Weigle, 1998), and improving raters' understanding and application of the scoring criteria (Weigle, 1994).

In spite of the evidence provided in the literature for the effectiveness of rater training, variation exists among raters even after training (e.g., Bonk & Ockey, 2003; Eckes, 2005, 2011; Lumley, 2002, 2005; Lumley & McNamara, 1995; McNamara, 1996; Papajohn, 2002; Weigle, 1998; Yan, 2014). Furthermore, the mixed findings of studies focusing on different approaches to training have complicated the issue (see e.g., Elder et al., 2007; Erlam, Randow, & Read, 2013; Knoch, Fairbairn, & Huisman, 2016; Knoch, Read, & Randow, 2007) and have necessitated the use of resolution methods to resolve the score discrepancies that exist among the raters after training.

LITERATURE REVIEW

Resolution methods are of two types: those involving a third rater as the adjudicator (tertium quid model, expert judgment model, and parity model) and the discussion method. In the tertium quid model, the score assigned by the adjudicator is combined with the closer score of the two original scores and the average score is then reported as the final score. In the expert judgment model, the expert's (adjudicator's) score replaces the discrepant scores and it is the score reported to the public. In the parity model, all the scores are treated equally as the adjudicator's score is combined with all the other scores and the average is reported.

A few studies have been conducted on the effectiveness of resolution methods leading to mixed results. For instance, Johnson, Penny, and Gordon (2000, 2001) found that holistic scores varied depending on the resolution methods used. The highest reliability was also found with the parity model. Johnson, Penny, Fisher, and Kuhs (2003) came up with similar reliability and validity estimates for parity and tertium quid models, though slightly higher inter-rater reliability was found for the parity model and lower validity estimates for tertium quid model.

As another resolution method, rater discussion (or rater negotiation, as referred to by Trace, Janssen & Meier, 2017), is aimed at bringing

consensus among raters through discussion. In this method, raters exchange ideas on the discrepant scores to find the reasons behind and resolve the discrepancy. It was "originally adopted due to resource constraints-in particular the lack of trained raters" (Trace et al., 2017, p. 2). In fact, in contexts where trained raters are not adequately and easily available, both the training program and resolution methods relying on a third trained rater would prove impractical. The discussion method as a post-exam procedure would be a helpful alternative, then.

Although there is not adequate empirical evidence on how and to what extent rater discussion can be fruitful in minimizing rater subjectivity, the few studies conducted in this regard have provided partial evidence for the positive effects of this method. For example, Clauser, Clyman, and Swanson (1999) found that discussion can only be minimally effective in increasing the precision of the resulting scores. Also, Johnson, Penny, Gordon, Shumate, and Fisher (2005) found that the scores obtained through discussion in comparison to resolution methods (averaging the scores) correlated more with experts' scores. However, Smolik (2008, p. 19) concluded that discussion "does not function satisfactorily as a score resolution method" though it can serve the function as a training tool. Finally, in a recent study, Trace et al. (2017) found that rater discussion could reduce rater bias and contribute to the raters' understanding of the scale. It could furthermore improve positive washback on teaching. However, it could not change rater severity.

PURPOSE OF THE STUDY

The studies reviewed above are far from being conclusive as to the effect of resolution methods. The paucity of research on resolution methods in general and the discussion method in particular calls for further studies in this regard. However, what is completely left understudied in the literature is the potential that rater discussion may hold as a pre-exam preventive measure rather than a post-exam resolution method. In all the variations of

training programs and resolution methods (except the discussion method), the presence of experienced trained raters is indispensable. In some contexts, especially EFL, this poses a big limitation as there could be a lack of access to such trained raters. Following Smolik (2008), the idea of using discussion as a training tool (and not a score resolution method appearing as a postmortem strategy) in which untrained raters come together to reflect and exchange ideas on their rating can be a promising alternative. The negotiations held among raters can serve a training function and help them develop their understanding of the scoring criteria and empower them in applying those criteria. As such, in this study, the use of rater discussion as a training tool was investigated to see how it may affect the rating behavior of raters when assessing speech samples from different proficiency levels. The following research question was put forward in this regard:

Can rater discussion function as a training tool? How does it affect raters' rating behavior while assessing L2 speaking performance?

METHOD

Participants

Five Iranian learners of English volunteered to take part in this study. They were all female and ranged in age from 28 to 33. They were Ph.D. students of Teaching English as a Foreign Language and had taken courses on language assessment both in MA and Ph.D. programs. Based on their experience in studying and teaching English especially at high levels, they could be roughly considered advanced in their English proficiency. All had similar experiences in teaching English at different levels and had the experience of rating speaking though mostly impressionistically without using any specific rating scale. That is, during their teaching career, they were at many times expected to rate their students' speaking performance. This was mostly done impressionistically. As for rating speaking using standardized rubrics such as TOEFL, IELTS, etc., Melika had the

experience of using TOEFL's speaking rubric though her experience was limited to only two times using this rubric to rate some Iranian learners' speaking performance. Also, Sarah and Ziba claimed some basic familiarity with the IELTS rating rubric and experience of using it in a few preparatory classes for IELTS. However, none of the raters had ever received formal training on rating. Table 1 summarizes demographic information about the participants.

Table 1: Demographic information about the participants

Raters	Age	Education	Teaching	General Rating XP	Rating with TOEFL/IELTS/etc.	Training
Melika	29	Ph.D. in TEFL	7 yrs.	Yes	TOEFL	No
Atefeh	32	PhD in TEFL	5 yrs.	Yes	_____	No
Nasim	28	PhD in TEFL	7 yrs.	Yes	_____	No
Sarah	33	Ph.D. in TEFL	5 yrs.	Yes	IELTS	No
Ziba	28	Ph.D. in TEFL	9 yrs.	Yes	IELTS	No

Materials and Instruments

Speech Samples

Fifteen speech samples were purposefully selected from a database of about 50 speech samples previously collected by the researcher and a colleague from different proficiency levels. The speech samples were in a monologue format as the volunteer test takers were given a topic and asked to talk about it in about 2 minutes. Care was taken to select a topic (the effect of technology on our lives) which was easy to discuss and of general interest to test takers. The samples used in this study were selected in a way to make sure different levels of oral proficiency were included. No exact rating was done at this stage. Rather the researcher and another experienced rater

listened to the samples and selected examples from different levels. The purpose here was to make sure that the final samples were not limited to a certain proficiency level. This would provide a better picture of the raters' behavior in rating speaking performance.

Semi-structured Interviews

Semi-structured interviews were employed to explore how raters rated the speech samples. For each rater, two interviews were carried out, one before using rater discussions and the other after discussions. About ten questions were carefully thought out by the researcher to focus on issues such as the criteria the raters considered in scoring, their understanding of the rating criteria, the band scores and descriptors, and the ease and speed of rating. The interviews that were used after the discussion sessions further included questions about the raters' experience and perceptions of the discussion sessions as well. Examples of the interview questions appear below:

- What is your score for this sample?
- Why did you score the sample …?
- Do you think your score accurately represents the sample of performance? Why?
- Are you sure about your score? Did you have any hesitations in scoring?
- Could you easily match the performance with a scale level?
- What is your idea about the rubric?
- Did you find the scoring criteria easy to understand and apply? Were they logical?

There was no time limitation for the interviews so that the researcher could delve into the issues deeply. Overall, they lasted between 30 to 60 minutes for each rater. All the interviews were conducted in English as the participants were advanced users of English.

TOEFL iBT Independent Speaking Rubric

TOEFL iBT independent speaking rubric is a holistic rubric that is used to assess speaking on an independent speaking task. Although like analytic rubrics, it contains some specific rating criteria for scoring, namely, delivery (flow of speech and clarity of production), language use (effective use of grammar and vocabulary) and topic development (development and progression of ideas), at the end raters are expected to make a holistic decision based on the general description provided for each score. The rating scale runs from 0 to 4.

Data Collection Procedure

Before running the study, the researcher met each of the raters individually and talked about the overall purpose of the study; that is, studying raters' behavior in rating speaking. However, details were not conveyed to them to avoid any potential effect on the final results. Furthermore, adequate information was provided about how long the study may continue and how many sessions the raters are expected to attend. Although no written consent was obtained from the raters, they were all volunteers who personally selected to participate in the study after receiving adequate information about it. Arrangements were also made with the raters beforehand to rate the samples at the presence of the researcher. The data were collected in five stages as depicted in Table 2. In stage 1, the raters rated 15 speech samples individually; that is, each rater came to the researcher's office and rated the samples using TOEFL iBT's independent speaking rubric. Immediately after rating each sample, an interview was conducted with the rater. The rater was asked to explain her reasons for assigning a certain score. She was supposed to comment on the test taker's performance and justify her score. Hesitations in and speed of rating were also checked. Finally, the rater's attitude toward the rubric was investigated. This procedure was followed for all the speech samples. Each rater rated the samples in one session. The whole session was audio-recorded for further analysis.

Table 2: Schematic representation of the study

Stage 1	Stage 2	Stage 3
Raters rating 15 speech samples; Semi-structured interviews immediately after rating each sample	Raters attending three discussion sessions	Raters rating the 15 samples again; Semi-structured interviews immediately after rating each sample

In stage 2, the raters were asked to attend three discussion sessions. At the beginning of the first session, the researcher explained the purpose of discussions and procedures to be followed. During each session, they listened to 6 or 7 samples and assigned scores independently. After rating each sample of performance, they were supposed to review the scores together and provide reasons for their scores, discuss their understanding of the rating criteria, challenge each other and finally decide on a score. In addition to the fact that discussion in this study was used as a training tool rather than a resolution method, what made the study different from other studies on discussion was that the raters exchanged ideas on all the scores and not just the discrepant scores; that is, even if for a certain performance all the raters had assigned the same score, still they were supposed to express their reasons for that score and reflect on others'. This created interesting discussions when they could see that sometimes they had assigned similar scores for different reasons. Furthermore, achieving consensus on a score was not an aim. At the end of discussions, raters could change their scores based on the feedback they had received or they could stick to their original score if they still thought the score was the best estimate of the performance. The same procedure was followed in the second and third discussion sessions. Overall, 20 samples from different proficiency levels were rated in the three sessions. The samples used in these sessions were different from those used in stages 1 and 2, though like them they were selected from different levels of proficiency.

After the discussion sessions which took about three weeks, the raters took part in the third stage of the study during which they rated the same 15

samples they had rated in stage 1. The same procedure was also followed here. At the end of stage 3, the raters were also asked about their experience of the ratings and discussion sessions, their perception of how these sessions affected their rating, and the positive and weak points of the sessions.

Data Analysis

The study mainly benefited from qualitative analysis through which semi-structured interviews before and after the discussions were carefully transcribed, coded, and analyzed. Furthermore, the transcripts of discussion sessions were explored in detail. In addition, many-faceted Rasch measurement (FACETS 3.71.4) was employed to analyze the data quantitatively. Rasch model is a unique psychometric model which is widely used in studies on raters' behavior (see e.g., Green, 2013). Following Knoch et al. (2007), Rasch analysis was conducted separately for pre-discussion and post-discussion ratings. Raters and test-takers were considered as facets. An expert rater was also asked to rate the 15 samples used in pre- and post-discussion sessions, so that his scoring could be used as a benchmark for comparison.

RESULTS

Analysis of the Interview and Group Discussion Data

Analysis of the data collected from rater interviews and group discussions provided insights into the effect of the discussion method on raters' behavior. In what follows, the themes emerging from the qualitative analysis of the data are discussed. Overall, six different themes emerged from the qualitative analysis:

- *Easier, quicker and more confident scoring,*
- *Co-construction of the meaning of the scoring criteria and their application*
- *Leniency in scoring*

- *Making norm-referenced comparisons*
- *Inadequacy of the scale levels*
- *Limitations of discussion as a training tool*

Easier, Quicker and More Confident Scoring,

The first theme emerging from rater interviews was the effect of discussion on raters' scoring method. All perceived discussion as being helpful and affecting their scoring positively. After discussions, they thought they scored the samples more easily, quickly and confidently. This can be seen in the following statements:

- *As we moved on, it became easier for me to decide. From the previous session [third discussion session] it became easy for me to decide*. (Melika)
- *The first time* [before discussions] *I was hesitant all the time, I had to listen to the tracks even more than 2 times*. (Nasim)
- *We were, I mean, affected. Something like training happened to us. These sessions we had together with my friends, that was a chance of learning for all of us.* (Sarah)

These statements indicate how the raters perceived discussion could be effective in improving their scoring confidence and speed. Analysis of their behavior before and after discussions provided evidence for this perspective. For instance, before discussions Atefeh was doubtful as to whether assign a score of 3 or 4 to a candidate (candidate 7) because she stated, the structures were not complex enough or the development of ideas was not well-sophisticated in her production. So, she decided to score her 3, but then she thought perhaps 3 could not represent her performance so she finally changed the score to 4. However, after discussions, the same candidate was easily assigned a score of 4 by her and she expressed her confidence in this score.

Co-construction of the Meaning of the Scoring Criteria and Their Application

Analysis of the discussions and interview data indicated that discussions helped raters establish their understanding of the criteria, in applying those criteria to scoring and in justifying the scoring decisions. This can be considered the most important theme emerging from the data because whatever raters do is one way or another influenced by their understanding of the rubric and scoring criteria. As such, more space is dedicated to explaining different aspects of this theme. The following excerpt taken from the first discussion session indicates how different features of the same speech sample were salient to the raters and how they considered the same criteria differently.

- Atefeh: *I couldn't decide between 3 and 4 but I think 3 is better because he was not so much fluent.*
- Ziba: *I had a problem with topic development. I think the development of the idea was limited. He didn't discuss so many ideas. He also didn't elaborate on his ideas. He just mentioned one basic idea. … That's why I scored him 3, not 4.*
- Nasim: *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.*
- Sarah: *But I don't think so … I mean the pauses included weren't for, I mean, going after language. They were to make fluent speech non-monotonous speech. Pauses were about finding meaning; they are interpreted even positive. Because they are excluding, let's say, monotonous, let's say, way of presenting something.*
- Ziba: *And what about topic development?*
- *Sarah: And about topic development, let's not forget about timed speech. When it's like two minutes it is not possible to go for more*

than this. Let's imagine ourselves in their shoes. What would we say?

- Nasim: *He just touched upon one point.*
- Ziba: *Yes, he could discuss two sides of the argument but he only discussed one part very basically.*
- Sarah: *People put their argument in a framework they like it. Like they may go for two merits and one demerit. So, it is not necessary to go two by two let's say examples to compare…*
- Melika: *Generally, the questions [in standardized tests] are in the form of arguments. They ask you for example which side you go to. You have to select one and develop that particular side.*
- Atefeh: *But he talked just one minute. How could he elaborate?*
- Melika: *He could have provided more examples of the negative sides of technology but this particular idea was quite complete. And if his speech was timed, I think that was enough.*

This excerpt indicates that Atefeh is hesitant between assigning a score of 3 or 4 and finally because of fluency assigns 3. However, later on in the discussions, it became clear that to her, pauses were the main index of fluency. So, the candidate was awarded 3 because of pauses. However, Sarah differentiates between two types of pauses: language pauses and meaning pauses. While the former indicates that the candidate pauses to think about and retrieve appropriate vocabulary or to formulate structures and hence can be considered a lack of proficiency, the latter is positive as it is used to avoid monotonous language and to present the meaning in a better way. Thus, fluency is operationally defined by them differently.

Similarly, concerning topic development, although the raters seem to have a similar understanding of the concept, they differ as to the extent to which it is represented in the sample. While Ziba and Nasim think that the sample produced is problematic because the candidate has only discussed one side of the argument without sufficient elaboration, Sarah states that the test taker cannot be expected to do more than this in a timed test.

Furthermore, in any argument people may stick to one side of the argument and frame their speech based on that excluding the other side altogether, an idea which is approved by Atefeh and Melika as well.

Analysis of the interview data also indicated that raters believed discussions had a great impact on their conceptualization and understanding of the scoring criteria. For example, Nasim stated that:

> *The first session, I couldn't decide; I didn't use the rubrics in a good way; I couldn't separate delivery and language use and topic development easily. After discussions, I can easily separate these and score them one by one and then give the overall score*. (Nasim)

Ziba believed that discussions helped modify her wrong perception of delivery as a scoring criterion. She had conceptualized delivery as pronunciation, intonation, rate of speech and more specifically and importantly accent. Thus, she had scored most of the candidates very low on this criterion because of not having a strong American or British accent or because of having no clear accent at all. Then she explains that in discussions she learned from Sarah and Melika that intelligibility is an important criterion not having a strong accent.

Atefeh thought that topic development is not as important as the other criteria because she thought it is mostly related to the raters' personal taste. That is, it depends on how the ideas developed by a testee are subjectively welcomed by a rater. As such, in making her overall judgment she paid less attention to this criterion. However, after discussions, she stated that she got a clearer picture of topic development and tended to consider it more carefully in her scoring.

Also, Melika states that discussions helped them think about their scoring and understand that the scale is holistic, not analytic.

> *Maybe the justifications, the reasons we had to provide, and we had to convince each other, maybe this justification made me think more about the use of the scale. I think at the beginning, for example, we were, ummm, our approach was more analytic and we were analytic as if we*

were analytically scoring, but this scale is quite holistic. This was once, this was the point that I told my friends. (Melika)

Leniency in Scoring

Another point the raters referred to as the effect of discussion was a change in their leniency in scoring. They believed that extreme cases of leniency or strictness were mitigated. In the following excerpt Ziba clarifies how discussions changed her attitude toward the native speaker and native-like performance as a criterion:

The thing I know that happened to me is that I really became less strict and less perfectionist. I had the idea in my mind that a person who is a native speaker can be scored 4 because of this scale. But now I see that being native-like is not just a criterion here.

Making norm-referenced comparisons

An interesting finding about how raters used the criteria was about the high tendency they had to compare the individuals with each other. That is, while rating rubrics are employed in performance assessments so that raters can match a test taker's performance with a score which best estimates that performance (making a criterion-referenced decision), raters in this study tended to assign scores by comparing individuals with each other (making norm-referenced decisions). Discussions helped raise their awareness about this issue, and, as they moved on, they employed less and less comparison, though even after discussions still, comparison was a scoring strategy for them. The following comments exemplify how the raters became aware of this issue and reminded each other to avoid it.

It [comparison] happened a lot in previous sessions when we had the discussion. ...Sometimes we were to stop each other: do not focus on the previous one; that is not norm-referenced; it's not to get back to others. But this [comparison] happens. But today I could get more

focused on the speaker himself or herself not the others. And that was because … we were always advised not to do that. (Sarah)

Analysis of discussions and interviews indicated that the raters made norm-referenced decisions particularly when two test-takers whose performances were very similar appeared in sequential order. This point is also stressed in the following statement:

When two interviewees are adjacent to each other and they are very different [in terms of proficiency] it doesn't happen to me to compare them. But when they are close to each other … I compare with the previous one. (Ziba)

Inadequacy of the Scale Levels

Another point about how the raters were affected by the discussions is related to the levels of scoring of the rubric. All the raters were of the idea that TOEFL iBT's rubric with a limited number of scores fails to adequately discriminate among different levels of performance. They strongly believed that more levels (or half sores) are needed. Although discussions facilitated the use of the rubric; that is, after discussions, the raters applied the rubric more easily, the problem was only partially solved and they referred to this as a shortcoming of the rubric. The following excerpt from discussion session 2 can elucidate the point:

Ziba: *It is not 4, it is less than 4, and he is not 3. You know, the problem we have, you remember the last session, the problem is with the score 3. You cannot make any decision near 3*.

Sarah: *let's say, at least 3 should be divided into two parts. 3 and 3 plus*.

Melika: *maybe, the points, we are not trained raters of TOEFL, maybe they are trained to work on specific aspects which we are not aware of. I mean it is difficult for us to do that.*

Sarah: … *let's be a bit fair, let's say that's why I think IELTS has surpassed TOEFL on this because people could have room for*

themselves [using IELTS scale] to explicitly say who has a better performance.

This excerpt is about a participant whose language performance was quite good but the raters felt they could not assign 4 to it. Interestingly, all believed his performance was somewhere between 3 and 4, and, since there are no half scores in the rubric, they could not assign an accurate score. Ziba clearly states that the problem goes to score level 3; that is, any performance near 3 is problematic and cannot be accurately scored. As such, the scale requires more levels around 3 so that raters can accurately differentiate among levels of performance. Melika speculates that the problem could stem from the fact that they are not trained raters and therefore are not skilled enough to cope with such problems. However, Sarah believes that the problem goes to the scale itself as other scales such as IELTS provide sufficient score levels to differentiate among individuals' performances. This idea is also reiterated by Ziba:

I'd rather use the IELTS scale because we have 0 to 9 and we have half scores. Half scores can help especially for the score 3 [on TOEFL scale]. (Ziba)

Limitations of Discussion as a Training Tool

Raters were also aware of the limitations of discussion as a training tool. As the discussions went on, more agreement was found among the raters. While this could logically be expected to stem from the training function of discussions, Ziba believed that sometimes the raters tended to agree just because they did not want to appear different from others.

I don't know maybe because we were afraid of losing our face or such stuff… Yeah, we didn't want to have a very far score from them, very different score from them. (Ziba)

She commented that the raters sometimes preferred not to recite their scores if they were different from other raters'. Alternatively, they dishonestly reported a score that was the same or very close to those of

other raters. This was because they thought to be different meant being wrong and being criticized. She further added that sometimes the raters easily changed their scores without being convinced that others are right or the scores reported by them are more accurate. She also argued that discussions are good providing that you know you are moving on the right track. What if the raters' agreement is because of the wrong reasons?

Rasch Results

Rater Severity

Table 3 depicts the multifaceted Rasch results for raters' severity before and after discussions. The Model Fixed Chi-Square statistic indicates that raters rated differently in terms of severity. "This again means that the grade the test taker receives may vary under the rater who grades his/her performance" (Green, 2013, p. 303).

Table 3: Severity measures for raters before and after discussions

	Pre-discussion	Post-discussion
Chi-square	32.5, d.f.=5, *p*=.00	27.9, d.f.=5,*p*=.00
Separation	2.50	2.04
Reliability	0.86	0.81

The separation value indicates that the raters are different in their rating (there exist at least two levels of severity) and the reliability value confirms that this difference is significant and real (Green, 2013). According to these values, the raters rated significantly differently in terms of severity both before and after discussions; however, this difference diminished after discussions. Overall, there is a greater similarity in leniency/severity with the expert after discussions.

Analysis of severity measures for individual raters also provided support for the effect of discussions. Table 4 indicates how individual raters changed through discussions. The closer a severity measure is to zero, the closer the rater is to the average of the group (Knoch et al., 2007). For this

study, higher scores indicate more leniency. It is indicated that Ziba who was the strictest rater before the discussions turned out to be the second lenient rater after. As commented by this rater, before discussions she avoided scoring individuals 4 unless they were native-like. At the same time, she easily tended to assign the score of 1. Through discussions, she came to know that she was too strict and perfectionist and therefore decided to be more lenient. This made her appear as the second most lenient rater after discussions. An interesting finding is about Nasim who changed a lot through discussions and became the most lenient rater after noticing that before discussions she was a little strict toward the test takers. Melika became stricter after discussions. For Sarah and Atefeh, the changes are less noticeable. The overall mean value indicates about 40% of reduction which means an improvement in rater severity.

Table 4: Severity measures of individual raters

	Pre-discussion	Post-discussion
Expert	0.61	-1.16
Melika	0.07	-1.16
Sarah	-0.45	0.15
Atefeh	-0.96	-1.16
Nasim	-0.96	5.17
Ziba	-4.97	2.14
M	-1.11	0.66

Raters' Consistency

Several issues were checked concerning raters' consistency in rating. First of all, there was a longer range of logits at post-discussion than at pre-discussion. This suggests there is less noise in the post-data. The raters are rating more consistently (Linacre, pers. comm., July 14, 2017). Second, the distribution of the test-takers was lumpier at post-discussion than at pre-discussion. This suggests that there is more agreement among the raters about the levels of performance (John Linacre, pers. comm., July 14, 2017).

Mean-square (MnSq) and inter-rater reliability values were also investigated in studying raters' consistency in scoring. The infit and outfit MnSq indices between 0.5 to 1.5 are considered acceptable (Green, 2013; Linacre, 2014). Smaller ranges are considered by other researchers (e.g., 0.75 to 1.3 by McNamara, 1996). Values above or below this range indicate that the raters are too unpredictable (significant underfit) or too predictable (significant overfit) in their rating respectively. As for ZStd, values above +2 indicate greater variance in rating than expected (being too unpredictable) and values below -2 indicate less variance than expected (being too predictable). Values between -2 to +2 indicate acceptable variability. Table 5 indicates that both before and after discussions all the raters rated consistently. The only exception is related to Nasim who before discussions showed a slight inconsistency in rating (infit MnSq = 1.75).

Table 5: Consistency measures for raters before and after discussions

	Preintervention						Postintervention					
	Infit		Outfit		Exact agreement		Infit		Outfit		Exact agreement	
	MnSq	ZStd	MnSq	ZStd	Obs.%	Exp.%	MnSq	ZStd	MnSq	ZStd	Obs.%	Exp.%
Expert	.35	-1.8	.24	-.8	68.0	66.0	.77	-.2	.22	.4	80.0	80.5
Melika	.89	-.1	.64	-.2	69.3	68.0	.78	-.1	.23	.4	80.0	80.5
Sarah	.86	-.2	.68	-.2	68.0	69.1	1.56	.9	.76	.5	77.3.	81.5
Atefeh	1.23	.6	.87	.0	62.7	69.3	.78	-.1	.23	.4	80.0	80.5
Nasim	1.75	1.7	1.62	1.0	62.7	69.3	.85	.0	.50	.5	56.0	58.1
Ziba	.79	-.3	.52	.3	45.3	48.9	1.11	.3	.85	.5	74.7	77.8
Total agreement					62.7	65.1					74.7	76.5

Raters' consistency was also checked through the inter-rater agreement index which is determined by comparing the percentage of exact agreements observed in the data with the percentage of exact agreements expected by the model. According to the Rasch Model, the two values are expected to be very similar or the observed value is slightly bigger; that is, the difference should not be more than 0.5. Large differences are depicted in the infit MNSQ value. When the two values are equal, the raters are behaving like independent experts; where the observed value is bigger than the expected value, "the raters may be considered as being too predictable (they are rating

in a clone-like fashion rather than as independent experts)" (Green, 2013, p. 224). When the observed value is smaller than the expected value, the raters are behaving unpredictably.

Table 5 indicates that before discussions, Atefeh, Nasim, and Ziba were rating unpredictably. Sarah and Melika were better raters with Sarah being slightly unpredictable and Melika being slightly predictable in rating. But after discussions, Melika and Atefeh were like independent raters; Ziba and Nasim were still unpredictable in rating but showed rating improvement in rating. Nasim showed more improvement in this regard. However, Sarah was the only rater whose rating deteriorated after discussions as she tended to rate with higher unpredictability than before. Overall, the table indicates a better rating and higher similarity with the expert after discussions. The total exact and expected values in the last row indicate that overall both before and after discussions the raters were behaving unpredictably. However, the Table also indicates that the difference between these values has decreased after discussions which means an improvement in rating.

Using Scale Categories

The results of Rasch analysis (Table 6) indicated that all the categories were used by the raters both in pre- and post-discussion sessions. However, the Table indicates a noticeable reduction in the use of category 1 from the pre- to the post-discussion session and an increase in the use of higher scores (category 3). This is in line with the raters' tendency toward becoming less strict.

Table 6: Use of rating scale categories by raters before and after discussions

Category score	pre-discussion		Post-discussion	
	N	%	N	
	%			
1	13	18%	4	7%
2	33	46%	29	48%
3	16	22%	21	35%
4	10	14%	6	10%

DISCUSSION

This study aimed at investigating the potential of discussion as a training tool. Qualitative analysis of the data indicated that the discussion method can serve the function of training raters. Besides the fact that all the raters were satisfied with their experience of discussions, a study of their behavior before and after discussions and their comments provided evidence for the positive effects of discussion on rating. Moreover, quantitative analysis of the data through multi-faceted Rasch supported the positive effects of discussion. This is very interesting. Like other resolution methods, the discussion method has been used after the exam to resolve discrepancies in rating. But the results of resolution methods have been indicated to differ considerably (Johnson et al., 2000, 2001, 2003; Smolik, 2008; Trace et al., 2017) and are subject to criticism. However, this study came up with the finding that the discussion method has the potential to remove rating problems at an earlier level; that is, before the exam as a training tool and this can happen in the absence of an expert rater. The beneficial effects of discussion were found in several areas in this study:

First, qualitative analysis of the data revealed the effect of discussion on the scoring method; that is, the raters rated more easily, quickly and of course less hesitantly after discussions. Furthermore, Rasch analysis indicated partial improvement in inter-rater reliability after discussions. Relevant literature has reported higher inter-rater reliability after rater training (e.g., Davis, 2016; Shohamy et al., 1992; Weigle, 1994). Thus, discussion can be claimed to function as a training tool in improving raters' ability in rating.

Second, discussions helped raters co-construct meaning for the scoring criteria at different levels which in turn enabled them to justify their scoring decisions by referring to such criteria. From the very first session of discussions and of course analysis of pre-discussion data indicated that they had different perceptions of the criteria included in the scoring rubric.

Various studies have indicated that raters may considerably differ in their interpretation of the rating criteria (e.g., Eckes, 2008; Engelhard & Myford, 2003; Lumley & McNamara, 1995; Weigle, 1998) or may, in general, have similar understanding of the rating criteria but stress different aspects of these criteria or apply them differently (Lumley, 2005). Therefore, training is required in performance assessment to help raters with their perception of the criteria. This study indicated that discussion can serve this function well. Trace et al. (2017) also found evidence for the positive effects of negotiation on raters' understanding of the criteria.

Discussions also helped raters rate more reflectively as they came to know that the scale is a holistic one, and they need to consider the criteria together. Furthermore, they became aware of their leniency or strictness in rating and tended to modify their rating. Understandably, more changes were observed with raters who were more severe/lenient than their peers before discussions. This finding is in line with the literature (Lim, 2011). Rasch analysis of the group severity measures also showed some improvement in rating. However, severity measures at individual levels indicated variations in benefiting from discussions. This variability can be explained by differences in rater background and abilities (Davis, 2016) or by individual differences in socio-cognitive factors (Baker, 2012). Some studies have indicated that variability may continue to exist even after training and the use of inappropriate criteria may still be observed in raters' behavior (Elder et al., 2005; Kim, 2011; Papajohn, 2002).

Raters' understanding was also improved concerning how they used the scale. It was found that raters tended to score a sample by comparing it with the previous samples rated rather than by considering the scoring criteria stated in the rubric. Comparison is a strategy used by all human beings while making decisions. In all theories of decision making discussed in cognitive psychology "judgments and decisions result from the comparison of an attribute's value to a sample of other values, either from the decision context or from memory" (Stewart, Chater, & Brown, 2006, p. 2). In Davis's (2016) study this comparison was made by referring to the exemplar

responses that helped the raters conceptualize the rating criteria more effectively and map the scores on the scale to the examinees' performance more easily. However, the kind of comparison which is used in the current study is different as it is not made with exemplar responses that could be considered as benchmarks. Instead, the raters tended to compare each individual's performance with similar performances to make a decision which would be fair. Although such comparison may help better rank order individuals (norm-referenced decisions) especially when the sample of test-takers is small, it cannot be expected to help raters accurately map the sample performances with the score levels (criterion-referenced decisions). After discussions, this tendency to rate individuals by comparison dropped, and raters relied more on the rating criteria instead.

Rating scales play a key role in rating. The effect of rating scale could be larger than raters' experience on their decision making (Barkaoui, 2010). The raters in this study experienced problems using TOEFL iBT's rubrics, especially when making decisions at higher levels. Although discussions helped the raters use the scale more effectively, the problem was not alleviated altogether. A close analysis of how raters should assign scores using this scale can pinpoint the root of the problem. The general descriptions provided in this rubric for scoring different levels can be translated into the scoring requirements depicted in table 7. As indicated, there is only one possibility for assigning score 4 but four possibilities for score 3. This makes 4 the easiest and least-problematic score and 3 the most controversial score to be decided by raters. This is also why the raters believed that they had to unfairly assign score 3 to test takers of different proficiency levels.

Table 7: TOEFL iBT considerations for scoring

Scores	Requirement
4	4 on all the three criteria
3	4 on two criteria and 3 on one criterion, 4 on one criterion and 3 on two criteria, 3 on all the three criteria, 2 on one criterion and 3 on two criteria
2	3 on one criterion and 2 on two criteria 2 on all the three criteria 1 on one criterion and 2 on two criteria
1	2 on one criterion and 1 on two criteria, 1 on all the three criteria
0	No or unrelated response.

Finally, although the findings of this study provided evidence for the beneficial effects of discussion on raters' rating behavior, two problems related to discussion were also stated by one of the raters. She commented that sometimes the agreement among raters was due to unacceptable reasons, such as the desire not to appear too different from others and therefore avoiding criticism. It was stated that at times raters changed their scores without even being convinced that others were right or more accurate in scoring. The literature has indicated that when raters are under the pressure of agreement, they may simply tend to agree at a superficial level (Barrett, 2001; Charney, 1984). Although the raters in this study were not under pressure as they were not expected to necessarily agree at the end of the discussions, still they did so in some cases. Green (2013) states that "raters, in general, try to agree with each other" (p. 224). Another problem stated is that discussions are good providing that you know you are moving on the right track. What if the raters are getting more and closer in rating but for wrong reasons? This point is also highlighted in the literature (Shohamy et al., 1992). Therefore, while discussion could be argued to improve the validity of the inferences and decisions made based on scores by improving

raters' understanding of the criteria and application of those criteria, it may overestimate reliability defined in terms of the agreement among the raters.

CONCLUSION AND IMPLICATIONS

The raters in this study participated only in three discussion sessions. Although more sessions might have brought higher levels of improvement in rating, the fact that discussion improved their rating in such a short time is insightful. Unlike the previous studies that used discussion after the exam as a resolution method, this study delved into the training potential of discussion used by a group of untrained raters in the absence of an expert rater. The result was quite promising. This could be of paramount importance in contexts in which access to expert raters is not a utility. It could also be beneficial to teachers all around the world as such discussion may be a regular behavior for teachers to sit at a table discussing their rating without an expert rater being present. Raters' educational background and teaching experiences could justify this finding. Although the raters in this study had received no formal training and were unfamiliar with the rating system, all had solid theoretical knowledge of language testing and all were experienced teachers. This background can logically be effective in helping them control and regulate their learning through training. Previous studies have refereed to raters' background such as relevant education and teaching experience as a source of beneficial effect in rating (e.g., Attali, 2016; Davis, 2016).

Higher levels of the TOEFL iBT's speaking scale were found more difficult for raters to distinguish. Such categories can indicate constructs that are "difficult to conceptualize and are open to multiple interpretations and likewise could inform performance descriptors in future rubric design studies" (Trace et al., 2017, p. 17). Furthermore, the literature has indicated that different scales can activate the use of different strategies and consequently can lead to different decisions on the part of raters (Barkaoui, 2010). Future studies can focus on how discussions may vary depending on

the type of scale used. However, in spite of all the beneficial effects of discussion found in this study, the possibility that sometimes raters agree superficially or for wrong reasons requires special attention. Further studies can explore this issue particularly as it relates to socio-cognitive factors. Finally, the fact that a small sample of raters took part in the study may limit the generalizability of the findings and needs to be taken into account.

REFERENCES

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

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49-58.

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.

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

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

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

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online rater training program. *Language Testing*, *24*(1), 37-64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*(3), 175-196.

Erlam, R., Von Randow, J., & Read, J. (2013). Investigating an online rater training program: Product and process. *Papers in Language Testing and Assessment, 2*(1), 1-29.

Engelhard Jr., G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model (College Board Research Report No. 2003-1)*. New York, NY: College Entrance Examination Board.

Green, R. (2013). *Statistical analyses for language testers*. New York, NY: Palgrave Macmillan.

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.

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.

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

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *Papers in Language Testing and Assessment, 5*(1), 90-106.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

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: Peter Lang.

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

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

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

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*(1), 27-33.

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

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1-26.

Tajeddin, Z., Alemi, M., & Pashmforoosh, R. (2011). Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference? *TELL, 5*(1), 125-153.

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

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-387.

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.