The impact of rater training on rater reliability in an English oral test

Shiknesvary Karuppaiah[1], Abdul Halim Abdul Raof[2]
[1]SMK Canossian Convent, Kluang, Johor
[2]Language Academy, Faculty of Social Sciences and Humanities,
Universiti Teknologi Malaysia
m-halim@utm.my[2]

Received: 27 November 2020; Accepted: 5 December 2020; Published: 16 December 2020

To cite this article (APA): Karuppaiah, S., & Abdul Raof, A. H. (2020). The impact of rater training on rater reliability in an English oral test. *Asian Journal of Assessment in Teaching and Learning*, *10*(2), 94-105. https://doi.org/10.37134/ajatel.vol10.2.10.2020

To link to this article: https://doi.org/10.37134/ajatel.vol10.2.10.2020

Abstract

Speaking skill assessment is gaining great interest in the field of assessment nowadays. Literature has highlighted reliability of raters in rating a speaking performance as one of the challenges due to human's subjective nature. This study has attempted to explore the influence of rater training on rater reliability in the assessment of a spoken task. A qualitative research design was used and, semi-structured interview was employed to obtain data for this study. A total of 21 secondary school teachers participated in the study. They were raters trained to assess an oral English interaction test. Data were analyzed using thematic content analysis which resulted in three main categories i.e. importance of rater training, effects of rater training on rater reliability, and improvement of rater training. The results show that rater training is essential before any rating is to be done, and its effects include, among others, maintaining rating consistency, exposure to test task, and criteria for grading. While suggestions to improve rater training sessions are related to the length, frequency, and quality of training.

Keywords: Rater Training, Rater Reliability, Oral Interaction Test, Speaking Skill, Rater

INTRODUCTION

It is indisputable that language assessment forms an integral part of any education system. In schools, decisions about a student's language ability often lies upon the results of language test(s) that students take, be it school-based or national level. This is especially true with the productive skills of writing and speaking. Of late, the assessment of the speaking skill has gained interest among researchers, teachers and those in the education system due to the applicability of its results as a threshold for pursuing studies or job applying purposes (Jenkins & Parra, 2003).

However, literature has identified several limitations in assessing students' speaking skill due to its subjective nature. One of them is reliability of raters. A rater is someone who uses a scoring rubric to measure a performance in assessment (Davies et al., 1999) while rater reliability is the extent of two or more raters agreeing on each others' scores of the same candidate (Fulcher (2003). The issue in question is what could be the cause of low rater reliability (or rating inconsistencies) among raters and how could this problem be overcome. This paper reports on a study which explored the influence of rater training on reliability of raters in an oral interaction test.

LITERATURE REVIEW

Achieving rater reliability is difficult as raters have their own rating principles and may be reluctant to deviate from those principles to adopt a new set of rating standard (Stahl and Lunz, 1991). Additionally, rater bias, which according to Hoyt (2000) is the disagreement due to raters' different interpretations of the rating scale or contradictory perceptions of candidates' performance, could result in inconsistency in evaluation.

One practical way to attain rater reliability is rater training (Elder et al., 2005). Rater training revolves around activities conducted by professionals or assessment experts to provide exposure, practice, and techniques of grading to raters. This would normally involve raters coming together to rate a performance, discuss their rating agreement, and modify their rating, if necessary, before reaching an agreement. This would contribute to a more reliable rating (Luoma, 2004). In addition, Cook (1989) views rater training as a process which helps raters to understand rating scales and test tasks better while Weigle (1994) regards it as a practice to minimise rating variations.

A study by Kang, Rubin & Kermad (2019) involving 82 inexperienced raters who assessed 112 speech samples shows that about 20% of the score variance was due to rater background and attitudinal factors. They also found out that having a user-friendly online rater training programme greatly lessened the impact rater background and attitudinal variables had on assessment. Thus implying that rater reliability could be improved with rater training.

One of the main enquiries of a study carried out by Bijani (2018) was investigating the validity of oral assessment rater training programme focusing on raters' perceptions and attitudes. On the whole, it was found that raters' perception of the feedback and training programme received were positive and encouraging. This was evident when a majority of the raters claimed that the training session had helped them to think more carefully and make changes to their rating behaviour accordingly. Findings of the study also revealed that raters with positive attitudes tended to achieve higher consistency in subsequent ratings. Similar findings were discovered by Davies (2016) with greater impact of rater training seen among raters who showed a more optimistic and positive view about it.

METHODOLOGY

This study employed a qualitative research design. A total of 21 English optionist teachers from secondary schools in Johor were selected using purposeful sampling method to be participants of the study. All participants had at least one year of teaching English and one year oral assessment rating experience. Details of the background of participants are shown in Table 1 below. The participants had undergone a rater training session prior to this study as they were among the teachers appointed as raters of an English oral test. The test that the teachers had to assess included a spoken interaction task between a candidate and an interlocutor.

Table 1. Background of participants

Participant/ Rater	Gender	Qualification	Teaching English experience (years)	Oral assessment rating experience (years)
R1	Female	Degree	3	3
R2	Female	Degree	30	4
R3	Female	Degree	3	1
R4	Female	Degree	27	4
R5	Female	Degree	23	1
R6	Female	Degree	3	2
R7	Male	Degree	9	4
R8	Female	Degree	3	3
R9	Female	Degree	1	1
R10	Female	Degree	6	4
R11	Female	Degree	1	1
R12	Female	Degree	25	3
R13	Female	Degree	3	3
R14	Female	Degree	3	2
R15	Female	Degree	19	4
R16	Female	Degree	3	3
R17	Female	Degree	18	3
R18	Female	Master	6	3
R19	Female	Master	7	3
R20	Male	Degree	10	3
R21	Female	Degree	3	3

As can be seen from Table 1, the majority of the participants were female degree holders and had between one and three years of teaching English experience. In addition, there was another group, slightly less in number, comprising senior teachers with ten or more years of teaching English experience. Many of the participants had at least three years of experience in oral assessment rating.

The participants were interviewed to gain insights on the influence of rater training on their practice in assessing an oral test. A face-to-face semi-structured interview was carried out to gather answers to the enquiry of the study. In instances where a participant was not available for a face-to-face meeting, the interview was done via a telephone call. Each interview was done individually and lasted for about 30 minutes. The responses given were recorded and later transcribed. Some of the questions used were as follows:

1. Do you think rater training is important? Why or why not?
2. How does rater training help you maintain your consistency in giving marks to students?
3. How does rater training help you generate agreement among raters?
4. How does rater training help you understand the spoken task(s)?
5. How does rater training help you understand the scoring rubric?
6. How does rater training help you with the understanding of allocation of marks?
7. How does group discussion during rater training help you?
8. Do you revise scores given to candidates? If so, why?
9. What criteria do you use to revise scores?
10. How can rater training be improved?
11. Who do you refer to to clarify doubts in assessment made? Why? How?

RESULTS AND DISCUSSION

Data gathered from the interview were analyzed and grouped under three categories. They are importance of rater training, effects of rater training on rater reliability, and improvement of rater training. The following sections present empirical data from the interview followed by a discussion related to each category.

Importance of Rater Training

Rater training is found to be essential. First is the provision of knowledge in such training. A total of 16 out of 21 participants agreed that rater training provided them with knowledge on rating criteria, and how to rate students' performance. As one participant (R1) put it, "In my opinion, I think rater training is important for teachers to know the standard way of rating.".

Rating consistency can also be improved by rater training as suggested by R12 who said "It is important because the training ensures that there is consistency in marks given to the candidates." **.** Besides, rater training acts as a platform for teachers to share ideas and build their rating confidence.

> "First of all, it is to encourage raters' confidence . . ." (R17)
> "Other than new knowledge, I think the teachers would probably get new insights or new experience, and sharing new ideas with each other." (R6)

Training also helps raters to rate without bias thus maintaining their grading reliability.

> "Rater training is important because to avoid bias and to understand what is required to us, which area for us to focus on evaluation." (R19)

The second reason why rater training is essential is the creation of a support group which could serve as a referral after training is concluded. In responding to the question on how raters would clarify their doubts, majority of the participants (17 out of 21) stated that they would discuss with their fellow raters or colleagues if they had any doubts about a candidate's performance so that they could provide a reliable score reflecting the candidate's actual speaking ability.

> "I would always discuss with my partners or other teachers." (R10)

On the other hand, four participants revealed that they preferred referring to the Area Appraiser (*PKW*) to clarify doubts as they are more knowledgeable on oral assessment.

> "I will probably ask someone who is in authority. Someone who would give me satisfactory answers, maybe the '*penaksir kawasan*'." (R12)

Although a participant agreed that doubts should be discussed during training as it would be helpful for other raters too, seven of them were uncertain about the efficiency of this idea but according to them guidance by mentors or trainers is helpful.

> "Yeah, by discussing and doing reflection." (R5)
> "I am not surely agree on that but it is just that up to themselves. But I think guidance by mentors will influence their judgment.".(R7)

The rater training seems to prepare raters to accommodate for any doubts raised by fellow raters pertaining to the assessment process. This could be done through discussions with other raters or colleagues or the trainers themselves. Additionally, the Area Appraiser (PKW) would be someone to consult with to resolve issues that may arise in line with Wang (2010) who stated that experienced raters are able to equip novice raters with adequate knowledge.

Based on the results, rater training is found to be essential for oral assessment due to several reasons. It enhances raters' knowledge and competency of rating besides gaining experience. Training also seems to help raters to rate without bias thus maintaining their grading reliability as well as paves the way to get support when in doubt.

Literature related to rater training highlights it as a fundamental factor in assisting raters to improve their rating skills. Kang (2012) and Kang, Rubin & Kermad (2019) asserted that rater training is vital to lessen rater variability while Xi & Mollaun (2009) found out that rater training functioned in identifying and reducing rater bias. Despite that, research in L2 performance suggests that leniency in rating can be reduced but may not be diminished completely (Weigle, 1998).

Thus, it is arguable whether rater training can eliminate bias in rating which has an effect on rater's reliability. Furthermore, training functions as a platform to share ideas on rating besides building up raters' confidence to evaluate candidates. This is true to a certain extent as novice raters need more guidance on the techniques of evaluation. Besides that, some novice raters may feel that they are not capable and inefficient in assessing the candidates as they are still new raters. Therefore, rater training is beneficial for novice raters specifically as the knowledge that raters gain from rater training guides them to rate without bias as Weigle (1994) attested that it is crucial that raters adhere to certain standards in subjective grading but this is unfeasible without training as rating will be unreliable.

Consistency is an important key point to help raters in arriving at a reliable scoring pattern as oral assessment is subjective and the evaluation relies wholly on raters' judgment. At the same time it is also a major threat to reliability (Brown, Bull & Pendlebury, 1997). According to KesharavMehr (2011), problems in rating happen due to inconsistent rating or raters' interaction with test items such as not adhering to rating criteria completely. Hence, rater training helps to overcome this problem by making raters to be more self consistent (Lumley & McNamara, 1995; Kondo-Brown, 2002). According to Wiseman (1949) as cited in Lumley and McNamara (1995), consistency is the main aspect used to judge effectiveness of rating done by raters however, it is important to train the raters to be self consistent and not forcing them to be in agreement with each other (Fahim & Bijani, 2011) because there is a tendency for raters to be more lenient than other raters when they are assessing candidates (McNamara, 1993).

Although studies show that rater training can help raters to be more self consistent, Wang (2010) claimed that training is not able to reduce rating bias wholly and this is in line with McNamara (1993). The proverb "Practice makes perfect" was also mentioned by raters as they believed that practice during training or hands-on experience is essential to promote consistency. Through practice, they are able to improve their rating techniques by clarifying any uncertainty during training. Practice helps raters to understand the rubric well (Davis, 2016). Besides practice, participants also stated that consistency can be achieved through various approaches used in training such as familiarizing to rubric pattern, discussion, group task, videos and explanation by trainers. Based on the responses gathered, it

is obvious that trainers employed a variety of techniques to enlighten raters' on rating process so rater training becomes a means of reducing rater error besides improving consistency (Farrokhi, Esfandiari & Schaefer, 2012).

Although many participants agreed that rater training improves rating consistency, there were also participants who disagreed and argued that only isolated practice enabled them to achieve high level of consistency. This may be due to their application of criteria during assessment. Joe (2008) found that both experienced and novice raters vary in terms of rating qualities and rating discrepancies can be minimized through practice. Weigle (1998) further attested that novice raters tend to apply scoring rubric more strictly than experienced raters which results in strict rating but studies show that training is effective to reduce rating strictness because novice raters rate just like other experienced raters after training.

However, this finding is not in line with Stahl and Lunz (1991). They argued that rater training is not successful in eliminating any discrepancies among raters who rate strictly. A possible reason on why participants believed that isolated practice helped them in maintaining consistency is because it enabled them to identify their weaknesses and they had ample time to improve on it before the actual rating of candidates. For instance, raters can rate the students during school-based assessment task (PBS). However, practice done in a one-day training session failed to identify raters' weaknesses as in isolated practice. Apart from that, these participants also need to adhere to the rubric as it is an important means to be consistent. Feedback also will be helpful especially for novice raters and they should get involved with retraining activities to learn how to use rubric in a consistent way (Myford & Wolfe, 2004). Elder et al (2005) further stated that if raters failed to be consistent with their ratings, they should be given follow up training after some time and peer feedback from group work.

Rater training also provides raters with the support they need when faced with difficulties with rating even after the session has ended. Fellow colleagues would normally be the first group that raters would consult. The area appraisers and the trainers too are resource persons whom a rater can refer to when having doubts about his or her assessment practices. Raters could learn from how they interpret and apply rating criteria to reduce scoring bias (Davies, 2016).

Effects of Rater Training

Rating consistency

Does rater training help raters to maintain their rating consistency? Out of 21 participants, 14 agreed. They thought that this was possible through the approaches used by trainers in the rater training such as familiarizing with the rubric patterns, having group discussion, doing group task, and watching video conversation on top of the explanation given by the trainers.

> "Each group will be given an assignment…So the rest of the group members can evaluate various answers and learn about consistency there" (R16)

Despite that, three participants, R3, R10, R17, disagreed. As for them, practice is the only key that enables raters to achieve their rating consistency and it cannot be achieved by just attending a one day training and one participant thought that rubric is important to ensure consistency. These raters have a range of teaching experience from one to 18 years but one to four years of rating experience. It can be inferred that their rating experience may have an influence on their perspective about rater training.

> "I don't think so because I have my own judgement for the criteria to evaluate oral assessment task." (R3)
> "I think consistency comes from practice itself. Practice makes perfect. " (R10)
> "That's why I said that the guideline is very important." (R17)

Unlike the raters who agreed on isolated practice, two participants also highlighted the vitality of practice during training as to promote their rating consistency.

> "In rater training teachers are given some practice on how to rate students. So with that they can learn about it." (R15)

> "Based on my experience, we have a mock thingy like we take one student as an example then we'll discuss what is the minimum requirement." (R19)

Generating Agreement

Almost all of the participants (20 of 21 participants) reported that rater training helped in generating agreement among raters and this is achieved through explanation, discussion, role play, referring to rubric and video samples. This supports Weigle (1999) and Bijani (2018) viewpoint however, Lumley (1998) cautioned that each rater has his or her own interpretation of what a good performance should be which is based on a candidate's abilities. The only one participant (R6) in this study who disagreed was a novice, someone with two years of rating experience and three years of teaching experience, and had this to say "I don't think so because each person has their own reasons on why they mark these students with these marks.".

It was reported that training helps to generate agreement through explanation, discussion, role play, referring to rubric and video samples. This is supported by Weigle (1999) that agreement can be achieved by training. However, a novice rater disagreed that rater training can help in generating agreement as training may not reduce rating differences since each rater has his or her own interpretation of how a good performance should based on candidates' abilities (Lumley, 1998). Brown (2003) also stated that novice raters have their own way of evaluation. Besides, there are also several factors such as different experiences or 'lack of agreement upon scoring routines' which influence raters' judgment that make raters to arrive at different scores when rating the same candidate (Davidson, Howell & Hoekema, 2000).

Task Exposure

Rater training seems to help expose raters to the spoken tasks through discussion, videos, print outs, Power Point Presentation, modules and sample questions during the training and this was agreed by 18 participants. The following sum up their point of view.

> "This is done in rater training by having suitable and various topics to be asked and to be dealt with different levels of proficiency." (R1)
> "They erm show them videos or of course they can have printouts or use power point to help them." (R12)

Conversely, three participants partially disagreed that rater training was able to enlighten them with the task format. According to them, experience, practical and in-house training could aid raters to comprehend the task better. A rater (R13) sums this up by saying "New teachers can be given in-house training to understand the task better such as in LET meting.".

It is obvious that rater training help in comprehension and exposure of task through discussion, videos, print outs, power point presentation, modules and sample questions. This is because; raters get the opportunity to experience how oral assessment is carried out in real examination setting through the use of these materials. Nonetheless, four participants were not in agreement that rater training enlightens them with task format. According to them, rating experience, practical and in house training work the best in accomplishing this purpose. Shahomy (1983) believed that extensive rating and experience are required to rate speaking proficiency reliably. Hence, rater training content should be modified to cater this purpose.

Marks Allocation

Marks allocation is a key element to ensure that a rater's scoring is reliable. Rater training seems to give exposure to raters in allocation of marks to candidates. The majority of the participants (19 out of 21) agreed that rater training helped in enhancing their knowledge of marks allocation through the use of examples, sharing of rating experience, explanation and by referring to the rubric.

"I would say yes and it is by having the guidelines and examples of mark allocation." (R1)
"We will discuss with other teachers and refer to rubric." (R8)

Marks allocation is a key element to ensure that raters' scoring is reliable and majority of the participants agreed that training aids in enhancing knowledge of marks allocation through examples, explanation, experience, group teaching and referring to rubric. The use of samples in rater training is emphasized by McCallean (2010) as it was suggested that samples used in training should be clear according to performance level of raters and different kinds of responses should be included in the samples. Moreover, studies suggested that effectiveness of rater training can be enhanced through ongoing discussions among raters. Retaining is recommended by Lunz, Wright & Linacre (1990) if raters could not comprehend the process of reliable marks allocation.

Changing Marks

It is normal for raters to revise the scores that they have given to candidates. Revision is necessary to ensure that raters do not leave out any evidence which could affect the marks of a candidate. From the interview, 17 participants confessed that they revised/changed the score after their first evaluation. The reasons given were; to ensure that reliable scores are given and to be fair to the students. The changes made were based on the knowledge that they had gained from the training they received. The following were some of the participants' responses including those citing the criteria they used to revise a candidate's marks.

"Yes, I revise to check their score." (R6)
"The criteria are proficiency, content, personal opinion so to be fair." (R2)

Conversely, there were others who stated that they did not revise scores or had a different reason for doing it.

"No, because the first thing is because we are lack of time and when we award the marks to evaluate and since we have discussed earlier with teachers, there is no need to revise the marks." (R4)
"I don't really revise because it is difficult to recall for me to revise." (R8)
"Sometimes I revise the marks because I pity the students." (R15)

There seems to be a possibility that raters revise the marks of candidates not based on criteria stipulated in the rubric. According to Calham and Spandel (1993), raters may rate students based on sympathy and not actual performance as they may have noticed the effort put in by students. Further analysis reveals that raters revised the marks because they had prior knowledge on the candidates' background as they had taught the candidates before.

Wang (2010) reported that discrepancies in rating happen because raters examined the same candidate whom they had examined in other settings, as in this case, candidates were evaluated by their own English teachers who know too well of the candidate's language ability. However, as Kang, Rubin & Kermad (2019) found out, the impact rater background and attitudinal variables could be reduced dramatically with rater training.

Criteria for Grading

There were four criteria for grading the oral test as stated in the rubric; 1. personal response and ideas, 2. fluency, 3. language accuracy, and 4. pronunciation. These were used in the rater training session attended by all raters. However, based on the interview conducted, it was found that participants had focused on two criteria for grading which were language proficiency and personal opinion (about a candidate). While language proficiency is related to criteria number 3 and to some extent criteria 2 and 4, personal opinion is not related to personal response and ideas which is basically the content of a talk. One possible reason for this could be because of the presence of a gap between the criteria in scoring rubric with raters' vagueness of the rating criteria (KesharavMehr, 2011). The following paragraphs

provide empirical evidence on this in the context of rater reliability.

According to a participant (R21), proficiency is an important criterion which influenced her grading. She further added that if students can speak well, they would be awarded with good marks. Another rater particularly considered students' language mastery, and understanding of the topic when grading a candidate's performance.

> "Actually I will look at proficiency level and tense. Sometimes, personal opinion about the students too." (R20)
> "In terms of proficiency, we have to look at the mastery of the language, must be how proficient they are to bring forth the points, the content point is understanding the subject matter" (R5)

Three participants (R3, R7 and R20) admitted that their personal opinion on candidates influenced their grading as to them by doing so it would be fair to the candidates. The following is what one of them (R3) said "Yes, but it is going to be fair for the students you see.". This practice could result in unreliable grading since apparently not all raters had used the (same) criteria as stipulated in the rubric. In addition, R7 also stated that by considering the personal opinion of raters in rating, the objective of the oral test might not be achieved.

> "Yes for assessment purposes I will hinder the achievement of the objective." (R7)

On the other hand, another three participants (R5, R8, R11) revealed that their personal opinion on the candidates did not influence their grading as they strictly adhered to the scoring rubric to help them provide the marks that students deserve.

> "No, it did not." (R5)
> "I don't think so because normally there are two raters, we can discuss." (R8)
> "I don't think the criterion (personal opinion) is reliable because students are taking exams on that day and I should assess the students based on the topic given on that day." (R11)

At the same time, eight participants touched on how to prevent bias due to raters' personal opinion of the candidates. A participant suggested inviting raters from other schools to evaluate students' performance as they would have no preconceived ideas about the candidates. Another participant put forward the idea that the oral test should be made as an ongoing assessment to ensure raters would not be directly influenced by their personal opinion about the candidates.

> "Maybe can get raters from other schools or having monitored by experienced teacher." (R3)
> "Perhaps the oral test should not be done on one exact date but ongoing assessment. . . let's say three months." (R7)

Some other suggestions on how to overcome the issue of bias among raters were also made. The following quotes highlight some of these:

> "OK maybe we can change the teachers instead of the same teacher rating students, maybe they can ask any other English teachers who are not teaching them to rate" (R16)
> "Whatever it is we have to look at the rubrics." (R17)
> "The area appraiser (the PKW) can help monitor raters to rate without bias." (R18)
> "This is where discussion comes in." (R19)
> "You cannot let the school teachers to evaluate their students. They need to ask other teachers or maybe form a committee." (R20)

Forming a committee is also a good idea as according to Bachman Lynch & Mason (1995), multiple raters can rate consistently, reduce variability (Lumley & McNamara, 1995) and increase grading reliability (Swartz et al,1999). In addition, Schaefer (2008) emphasized on the importance of having consistent and objective raters as they can reflect students' ability. Conversely, Schneider (2001) claimed that having two raters to score candidates increases reliability of grading but this decreases practicality of test in terms of assessment time as raters need to spend ample of time to rate the candidates when in reality they are only given limited time to assess the oral performance in a particular test.

Participants also suggested that marking with reference to rubric will reduce bias due to personal opinion. However, it is crucial to ensure that raters can comprehend the rubric and its expectations clearly before using it (Jonsson & Svingby, 2007). Besides, raters also need to be provided with detailed rubric as to improve the quality of task (Baldwin et al, 2009). Thus, clarity and appropriateness of a rubric is crucial for grading as it can avoid uncertainty (Reddy and Andrade, 2010). Furthermore, it should be a highly structured rubric to reduce raters' personal choice of criteria (Goulden ,1994).

At the same time, three participants stated that their ratings were not influenced by personal opinion. This is however not in line with Clark and Lett's (1988) findings that examiners may be influenced by functional and interpersonal elements as they have little time to give emphasize on candidates' language.

Although raters are provided with an appropriately designed rubric to evaluate candidates, the findings clearly prove that raters' choice of assessment criteria and score revision criteria varies depending on other factors. Therefore, rater training should be modified so that a uniform application of rating criteria can be employed in the oral assessment.

Improvement of Rater Training

Charney (1984) cited in Fahim and Bijani (2011) reported that the aim of rater training is mainly to prevent raters from using their own judgments as differences between raters may vary. A number of suggestions were made in response to the question of improving oral assessment rater training.

More video samples

The first suggestion revolves around having more video samples of candidates' oral performance from different proficiency levels. Ten participants recommended this idea.

> "Maybe can discuss more videos of candidates and try to rate them to see if raters rate the same way." (R1)

By watching videos of students' performance from different proficiency levels, raters can identify a number of problems that may arise. In relation to this two participants highlighted the importance of having clearer examples during the training which would further enhance raters' understanding of the assessment process.

> "I would prefer a clearer example be given during training." (R6)

Longer duration and more training sessions

The second suggestion as recommended by six participants was to conduct longer training sessions. Currently, the training session lasts for only one day which seems hardly enough for raters to enhance their knowledge on oral assessment rating.

> "Longer time for training the raters instead of a one-day course and should have an on and off discussion rather than having a one shot training." (R2)

A one-day training seems insufficient to train raters from the entire district as each rater may have different previous rating experiences that require more guidance. According to Wang (2010) training is an "ongoing business and not a once for all matter". A related suggestion was to have more training sessions per year.

> "Training should be done two to three times a year to fully prepare raters." (R11)
> "We need to have more training from time to time because we got training for few days before assessment." (R20)

This is in support of Wigglesworth (1993) who proposed that training should be conducted regularly and in detailed to help raters not to assess the candidates harshly. Besides, Wang (2010), and Lumley & McNamara (1995) found that training effect may not persist for a long time after each session thus constant training needs to be conducted to maintain rating quality. Thus, it is crucial to hold more training sessions before each test to allow raters to re-establish their rating criteria. While this could be true in most cases, findings of the study by Kang, Rubin & Kermad (2019) imply that short, user-friendly online rater training sessions would lessen rater bias, at least within a certain period.

Conduct mock sessions

A third suggestion given by three participants was to have mock training session or hands on experience as a practice. In addition, two participants suggested to have more group activities before they get on to the real assessment.

> "I think we can call the PKW to do mock training where there will also be students." (R14)

Mock training sessions and group activities are seen as initiatives to provide raters with close to authentic rating experience before the real assessment takes place. Mock training will be beneficial for raters especially if it is carried out by experts so that raters can comprehend how experts rate and justify their grading and compare with their rating style (Orr, 2002). Besides, group activity also could help raters to have the opportunity to assess and discuss simulated tasks using specific criteria (Wigglesworth, 1993).

Through hands on experience and group discussion, raters are able to exchange viewpoints and suggestions which will be helpful for them. Some of the raters are actually teachers from schools which are performing well. Hence, they will have different rating experience which can benefit all the raters when shared.

Engage experts as trainers

The final suggestion to improve rater training was to engage experts to conduct the sessions. The following are what the participants said,

> ". . . to have *guru bertauliah* who really knows what he or she is doing so it will be more reliable." (R11)
> "Yes I think an improvised training should be given to help teachers especially in oral assessment. Experts are the best people to conduct training." (R16)
> "More often and more qualified trainers as sometimes they don't have answers to our questions." (R20)

The suggestion to improvise the present rater training by engaging experts or more experienced raters to conduct the training is good. As McCallen (2010) found that it is crucial to have raters who have expertise in the content of assessment. Thus, with this knowledge, trainers can conduct the training successfully.

CONCLUSION

Rater reliability is a major concern in assessing the speaking skill as it is not easy to achieve reliability in rating. Rater training is seen as a way to attain high rater reliability however, rater training may not always produce good results. This study has explored the influence of rater training on rater reliability. Findings reveal that the majority of raters were in agreement that rater training is essential as it promotes the provision of knowledge to enhance reliability of raters. It is also functional in the sense that it gives rise to a support group which could serve as a referral even after training is concluded. In addition, rater training seems to have positive effects on raters including maintaining rating consistency, deeper understanding of test task and rating criteria, as well as better application of scoring rubric. While suggestions for improvement of rater training relate to length, frequency and quality of training.

This study also points to the need to further investigate other variables and perspectives in the relationship between rater training and rater reliability to shed more light on oral assessment. This is especially important now as the outcome of a rating of a spoken task is used more than before in determining the language ability and the readiness of candidates for further studies or for the workplace.

REFERENCES

Bachman, L.F., Lynch, B.K. and Mason,M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*. 12, 238-257.

Baldwin, S.G., Harik, P., Keller, L.A., Clauser, B.E., Baldwin, P., Rebbecchi, T.A.(2009). Assessing the impact of modifications to the documentation component's scoring rubric and rater training on USMLE integrated clinical encounter scores. *Acad Med*, 84, 97-100.

Bijani, H. (2018). Investigating the validity of oral assessment training program: A mixed-methods study of ratrer' perceptions and attitudes before and after training. *Cogent Education*, 5(1), 1-20.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20, 1- 25.

Brown, G., Bull, J., and Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

Clark, J.L.D. and Lett, J. (1988). A research agenda. In Pardee,L,J.,and Charles,W.S,(Ed) *Second Language Proficiency Assessment:Current Issues.*(pp. 54 -82). Englewood Cliffs,NJ:Prentice Hall.

Culham, R.and Spandel, V. (1993). *Problems and Pitfalls Encountered by Raters*. Developed at the Northwest Regional Educational Laboratory for the Oregan Department of Education.

Cook, S.S. (1989). Improving the quality of student ratings of instructions: A look at two strategies. *Research in Higher Education*, 30(1), 31-45.

Davidson, M., Howell, K. W. and Hoekema, P. (2000). Effects of ethnicity and violent content on rubric scores in writing samples. *Journal of Educational Research*. 93, 367-373.

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

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Elder, C, Knoch, U., Barhuizen, G. and Von Randow, J. (2005). Individual feedback To enhance rater training: Does it work? *Language Assessment Quarterly.* 2(3), 175-196.

Fahim, M. and Bijani, H. (2011). The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment. *Iranian Journal of Language Testing.* 1, 1 – 16.

Farrokhi, F., Esfandiari, R. and Schaefer, E. (2012). 'A Many-Facet Rasch Measurement ofDifferential Rater Severity / Leniency and Teacher Assessment', *Journal of Basic and Applied Scientific Research.* 2(9), 8786–8798.

Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Longman.

Goulden,N.R., (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Developmentin Education*, 27, 73 – 82.

Guest,G., Bunce, A. and Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*. 18, 59-82.

Hoyt, W. T. (2000). Rater bias in psychological research:When is it a problem and what can we do about it? *Psychological Methods*. 4, 64-86.

Jenkins, S. and Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: Thecomplementary roles of non-verbal, paralinguistic and verbal behaviors in assessment decisions. *The Modern Language Journal*. 67, 90-107.

Joe,J.N. (2008). *Using Verbal Reports to Explore Rater Perceptual Processes in Scoring: An Application to Oral Communication Assessment.* PhD Thesis, James Madison University,US.

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

Kang, O. (2012). 'Impact of Rater Characteristics and Prosodic Features of Speaker Accentedness on Ratings of International Teaching Assistants' Oral Performance'. *Language Assessment Quarterly*. 9(3), 249–269.

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481–504.

KeshavarzMehr, N. (2011). *The critical role of subjectivity at the item level in a test of spoken English: variability in rater estimations*. PhD Thesis, Melbourne Graduate School of Education, The University of Melbourne.

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

Lumley, T. (1998). Perceptions of language – trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purpose.* 17, 347 – 367.

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

Luoma, S. (2004) *Assessing Speaking*. Cambridge: Cambridge University Press.

Lunz, M. E., Wright, B. D. and Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied measurement in education.*3, 331-345.

McClellan, C.A. (2010). Constructed-response scoring: Doing it right. *R and D Connections,*13, 1-7.

McNamara, T.F. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessment of ESL writing samples.* Unpublished master's thesis, Faculty of Education, the University of Melbourne, Melbourne, Australia.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores.*System.* 3(1), 143-154.

Reddy,Y.M., and Andrade, H. (2010). A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, 35(4), 435-448.

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

Schneider, P. (2001). Microstructure analyses: Referential cohesion. Presented as part of seminar: Development and implementation of a narrative norms project", *American Speech-Language-Hearing Association Convention*, New Orleans, LA, November.

Shohamy, E. (1983). Rater reliability of the oral interview speaking test. *Foreign Language Annals*, 16(3), 219-222.

Stahl, J. A. and Lunz, M. E. (1991). *Judge performance reports: Media and message.*Paper presented at the annual meeting of the American Educational Research Asscoiation,SanFrancisco , CA.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E. L., Reed, M., et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, 59, 492–506.

Wang, H. (2010). *Investigating the justifiability of an additional test use: An application of assessment use argument to an English as a foreign language test.* Doctoral dissertation, University of California, Los Angeles.

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

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing,*15(2), 263-287.

Weigle, S. C. (1999).Investigating rater prompt interactions in writing assessment:Quantitative and qualitative approaches. *Assessing Writing,*6, 145 -178.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing,* 10, 305-335.

Xi,X. and Mollaun,P. (2009). *How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps.* Princeton,NJ: Educational Testing Service.