Pertanika J. Soc. Sci. & Hum. 29 (S3): 401 - 419 (2021)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

# The Effects of Different Rater Training Procedures on ESL Essay Raters' Rating Accuracy

**Souba Rethinasamy**

*Faculty of Language and Communication, Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak, Malaysia*

## ABSTRACT

The study investigated the effects of three commonly employed rater training procedures on the rating accuracy of novice ESL essay raters. The first training procedure involved going through a set of benchmarked scripts with scores, the second involved assessing benchmarked scripts before viewing the scores. The third was a combination of the former and latter. A pre, post and delayed post-experimental research design was employed. Data were collected before, immediately after and one month after training. Actual IELTS scripts with benchmarked scores determined by subjecting expert IELTS raters' scores through Multi-Faceted Rasch (MFR) analysis were used for the training and data collection purposes. Sixty-three TESL trainees were selected based on their pre-training rating accuracy to form three equally matched experimental groups. The trainees' scores for the essays before, immediately after and one month after the assigned training procedure were compared with the official scores for the operational essays. Although the findings indicate that generally, rater training improves raters' rating accuracy by narrowing the gap between their scores and the official scores, different training procedures seem to have different effects. The first training procedure significantly improved raters' rating accuracy but showed a decreasing effect with time. The second training procedure showed immediate as well as delayed positive effects on raters' rating accuracy. The third training did not lead to significant immediate improvement, but rating accuracy improved significantly after some time. This paper discusses the implications of the findings in planning efficient and effective rater training programmes.

*Keywords:* Assessing writing, rater training, rating accuracy, standardisation, validity and reliability

ARTICLE INFO

*Article history:*
Received: 16 July 2021
Accepted: 04 October 2021
Published: 30 November 2021

DOI: https://doi.org/10.47836/pjssh.29.S3.21

*E-mail address*:
rsouba@unimas.my

ISSN: 0128-7702
e-ISSN: 2231-8534

## INTRODUCTION

Assessment is often seen as a crucial and integral part of teaching and learning. Assessment in education has been going through a major shift from traditional assessment of cognitive knowledge only to performance-based assessments. The scores derived from assessments conducted by educational institutions and testing bodies usually have critical implications on both the test takers and the stakeholders. However, subjectivity in assessing performance assessments, including written essays, poses a major threat to validity (Barkaoui, 2011; Lumley, 2002; Messick, 1994; Shabani & Panahi, 2020; Xie, 2015).

While a common yardstick referred to as rating scale or rubrics help reduce subjectivity in scoring even when multiple assessors are involved (Ragupathi & Lee, 2020), rubrics alone are insufficient to improve standardisation in scoring (Brown, 2009; Reddy & Andrade, 2010). Rater training has been an important component of assessment literary and is often recommended to increase the validity and reliability of scoring from a rubric. Rater training is also known as 'standardisation', 'moderation', 'calibration', 'parity' and 'norming' sessions (Hamilton et al., 2001; Hodges et al., 2019; Kondo, 2010; McIntyre, 1993; Schoepp et al., 2018). During rater training, raters are calibrated towards a common rubric with exemplar scripts to guide them to interpret the rubrics in a similar manner (Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010). Rater training is suggested to decrease subjectivity in rating, keep score variations within acceptable limits and assist raters to assess according to standards set by the testing organisation. According to Alderson et al. (1995), "the training of examiners is a crucial component in any testing programme, since if the marking of a test is not valid and reliable then all of the other work undertaken earlier to construct a 'quality' instrument will have been a waste of time" (p. 105).

In the last two to three decades, rater training has become widely accepted and implemented by many educational institutions and language testing organisations such as Cambridge ESOL which is responsible for the International English Language Testing System (IELTS), Educational Testing Services (ETS), which is responsible Test of English as a Foreign Language (TOEFL), and Malaysian Examination Council which is responsible for the Malaysian University English Test (MUET) (Brown, 2000; Chan & Wong, 2004; Furneaux & Rignall, 2002; Wei & Llosa, (2015).

A literature review shows that empirical studies on rater training only started to gain some attention in the 1990s. For example, Shohamy et al. (1992) and McIntyre (1993) reported that training improves raters' ratings, particularly inter-rater and intra-rater reliability. Weigle (1998) found the training to be more beneficial to improving intra-rater reliability than inter-rater reliability. On the other hand, Engelhard (1992, 1996) reported significant differences in rater severity and accuracy even among highly trained raters. Myford

and Wolfe (2009) found significant positive and negative drift in rater accuracy over time for a small proportion of the raters. Despite similarities, the studies have reported some rather contradictory findings.

While examining rater, essay and environment effects, Freedman (1981) unexpectedly found that subtle differences in approach and input during training could lead to significant differences in rating. For example, the training that raters in Weigle's (1994, 1998, 1999) study went through consisted of the following procedures:

- reading through exemplar essays with their official scores
- assessing a set of essays and compare own scores with the official scores
- explaining reasons for own scores that differ from the official scores and reaching an understanding of the reason for the official score

Weigle (1994) also added that a complete description of the training session was not possible and that a certain amount of 'informal training' also took place as the ratings were done in a group setting where raters could see the scores given by the previous rater and receive feedback on their ratings. Trainers also did speak to the raters whose ratings were aberrant in some ways.

On the other hand, in Lumley's (2000, 2002) study, the rater training involved the following two major procedures.

- practise assessing several sample essays using the rating scale
- discuss the scores and the reasons for and against different scores, by the trainers and/or the other members of the group

An extensive review of studies shows that rater training programmes seem to employ several procedures in various ways (Attali, 2015). The most common procedures utilised in rater training programmes are

- going through exemplar essays with their official scores (Furneaux & Rignall 2002; Knoch et al., 2007; McIntyre 1993; O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Raczynski et al., 2015; Wang et al., 2017; Weigle, 1998; Weigle, 1999)
- practise rating exemplar essays and comparing own scores with official scores (Erlam et al., 2013; Furneaux & Rignall 2002; Knoch et al., 2007; Lumley, 2000; Lumley, 2002; O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Weigle, 1998; Weigle, 1999; Wolfe & McVay, 2010)
- discuss reasons for scores (Kim et al., 2017; Knoch et al., 2007; Lumley, 2000; Lumley, 2002; O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Shaw, 2002; Weigle, 1994; Weigle, 1998; Weigle, 1999)

As cautioned by Freedman (1981), the differences in the training procedures employed during rater training sessions raise the question of whether they had contributed to the inconsistencies in the effects of rater training in language performance assessment.

Although many studies have compared the effects of different rating procedures, especially in the field of performance appraisal (Ellington & Wilson, 2017; Rosales-Sánchez et al., 2019; Tziner et al., 2000), studies comparing the effects of different rater training procedures in assessing language performances seem scarce. Despite several calls to investigate the effect of procedures used for training language performance raters so that these procedures could be put to best use (Freedman 1981; Furneaux & Rignall, 2002; Hamp-Lyons, 1990; McIntyre 1993; O'Sullivan and Rignall, 2001), only one type of training procedure, i.e. feedback, that too as a post-training procedure, has received some attention (O'Sullivan & Rignall, 2001; O'Sullivan & Rignall, 2002; Shaw 2002; Wigglesworth 1993). Although studies by Leckie and Baird (2011) and Gyagenda and Engelhard (2009) have focused on rater training in language performance assessment, they did not study the effects of the training.

Wigglesworth (1993) experimented with the potential effect of Multi-Faceted Rasch (MFR) based bias analysis feedback as a form of post-training procedure on a speaking test. The study found some evidence of improvement in rater consistency following the feedback and recommended the implementation of the bias analysis feedback into rater training. O'Sullivan and Rignall (2001) conducted an experimental study to explore Wigglesworth's (1993) suggested use of MFR based bias analysis feedback as a form of post-training procedure in the context of writing assessment. The MFR bias analysis feedback had an additional brief written description to make it self-explanatory. Twenty trained and experienced IELTS examiners with at least two years of rating experience were involved in this study. The study utilised 81 essays written by candidates who sat for the IELTS Writing Module in 2002. The findings showed that only written feedback had a limited effect on the Feedback Group's rating performance.

O'Sullivan and Rignall (2001) hypothesised that feedback delivered systematically over a period may result in more consistent and reliable examiner performance. Shaw (2002) investigated the effect of feedback delivered over a period. The feedback given to the participants in this study was based on the official scores for the essays, with notes explaining the reasons for the scores. The participants were the Certificate of Proficiency in English (CPE) examiners. Data were collected on five successive rating occasions. The results showed a small gain in the percentage of rating with 0 band difference and a small gain in the percentage of rating with 0 and 1 band difference. Shaw (2002) attributed the small improvement in accuracy of the experienced raters to the possible inherent standardisation quality of the revised scoring rubrics.

While the studies above investigated the effect of feedback as a post-training procedure, the effects of the different procedures employed 'during' training on raters' rating have not been addressed

sufficiently. In addition, researchers have highlighted that a great deal remains unknown about the effects of different rater training procedures on raters' rating accuracy (Azizah et al., 2020; Leckie & Baird, 2011; Raczynski et al., 2015; Wolfe & McVay, 2010).

The emphasis of research on rater training in the 21st century shifted to the emergence of web-based rater training programmes. An early study by Hamilton et al. (2001) described a pilot online rater training programme and investigated the raters' attitudes toward the programme. A similar study conducted by Elder et al. (2007) also canvassed raters' responses towards the effectiveness of an online rater training programme. Knoch et al. (2007) compared the effectiveness of an online rater training programme and face-to-face rater training in a large-scale writing assessment. On the other hand, Attali (2015) compared the effect of web-based rater training between inexperienced and experienced raters. While these studies indicate the practical alternative to face-to-face rater training, the effects of the different rater training procedures employed during training to train the raters remain unanswered. Thus, it creates a huge gap in designing effective rater training courses and calls for focused investigation in this area (Shabani & Panahi, 2020).

The present study attempts to address this gap and shed light on the effects of some of the commonly employed procedures during essay rater training on raters' rating accuracy. The general research question that the study aimed to address is;

"How do the different rater training procedures affect raters' rating accuracy?":

This study investigated the effects of different essay rater training procedures on the rating accuracy of novice ESL raters.

The study attempted to answer the following research questions

RQ1. To what extent do the different rater training procedures affect ESL raters' rating accuracy immediately after training?

$H_{01a}$: There will be no significant difference between the rating accuracy of the Training Procedure A group immediately after training compared to before training.

$H_{o1b}$: There will be no significant difference between the rating accuracy of the Training Procedure B group immediately after training compared to before training.

$H_{o1c}$: There will be no significant difference between the rating accuracy of the Training Procedure C group immediately after training compared to before training.

RQ2. To what extent do the different rater training procedures affect ESL raters' rating accuracy stability over time?

$H_{o2a}$: There will be no significant difference between the rating accuracy of the Training Procedure A group one month

after training compared to before training.

$H_{o2b}$: There will be no significant difference between the rating accuracy of the Training Procedure B group one month after training compared to before training.

$H_{o2c}$: There will be no significant difference between the rating accuracy of the Training Procedure C group one month after training compared to before training.

## Materials and Methods

The study employed a matched pairs quasi-experimental design with three rater training procedures, three rating occasions and three experimental groups. The first rating was done before training, the second was completed immediately after training, and the third rating was done one month after training.

## Participants

Shohamy et al. (1992) and Weigle (1998) highlighted that raters' background could influence their rating. Thus, a homogeneous group of raters with similar backgrounds were selected as participants for this study. The study involved all the penultimate and final year undergraduates taking a degree in Teaching English as a Second Language (TESL) at a public university in Malaysia. The requirements to be accepted into the TESL programme are that an applicant must have obtained good grades in the Malaysian equivalence of the GCSE and A-Level examinations. In addition, candidates also must have a distinction in the GCSE English language papers and at least a band 3 in the Malaysian University English Test (MUET). The demographic data obtained from the participants confirmed that the participants of the study fulfilled these language requirements. Also, they have not had any formal training in assessing written essays. Thus, the participants could be classified as novice essay raters.

## Instruments

The materials used in carrying out the planned study must go through several proper construction stages, vetting and testing. In turn, it would help ensure that the findings from the study are not affected by the problems related to the writing task, scoring rubrics or the scripts. The International English Language Testing System (IELTS) test was mainly chosen because it is a long-established high-stakes test used in assessing international students' English language proficiency. In addition, the tasks and rating scales of the test have gone through several years of rigorous experimentation and validation.

The IELTS is designed to assess "the language ability of candidates who intend to study or work where English is used as the language of communication" (IELTS, 2003, p. 3). The IELTS test's ability of test-takers all four language skills—Reading, Writing, Listening and Speaking. IELTS provides a nine defined band level that ranges from Non-User to Expert-User as a guide for

interpreting the band scores (Green, 2003; O'Sullivan & Rignall, 2001)

The IELTS test for Writing Task 2 Version 42 (a retired version) was utilised for this study. The topic for the writing Task 2 Version 42 reads as below:

> Only parents can offer the care and attention that is necessary to a child's development. It is, therefore, wrong for both parents in a family to expect to pursue a career: one of them, whether it is the father or mother, should stay at home and look after the children.
> Do you agree or disagree?
>
> Give reasons for your answer.
>
> You should write at least 250 words.

**Benchmarked Scripts**

The study utilised IELTS essays as benchmarked scripts for training. On request, 81 essays written by candidates on the topic and the rating scale were provided by Cambridge ESOL. The official scores of the essays were determined by subjecting the certified IELTS raters' scores to Multifaceted Rasch analyses. The essays were rated using the IELTS rating scales. The rating scales are not made public; however, a public version is available at https://www.ieltsanswers.com/wp-content/uploads/2016/09/Essay-writing-criteria-official.pdf .

From the 81 essays, a total of 36 essays were selected for the study. For training purposes, two parallel sets (matched in terms of their scores) consisting of nine essays each were selected. First, the two sets were labelled as Set A and Set B. Then, 18 essays were selected to form the operational set (Set C) used for actual rating purposes. Table 1 and Table 2 provide the list of essays and their band scores.

Table 1
*List of the parallel Set A and Set B and their official bands*

| SET A | | SET B | |
|---|---|---|---|
| Essay ID | Global band score | Essay ID | Global band score |
| 21 | 3 | 08 | 3 |
| 13 | 4 | 53 | 4 |
| 03 | 5 | 67 | 5 |
| 16 | 5 | 54 | 5 |
| 84 | 6 | 65 | 6 |
| 66 | 6 | 40 | 6 |
| 22 | 7 | 39 | 7 |
| 38 | 7 | 82 | 7 |
| 33 | 8 | 69 | 8 |

Table 2
*List of operational essays (Set C)*

| Essay ID | Global band score |
|---|---|
| 43 | 3 |
| 27 | 4 |
| 87 | 4 |
| 02 | 5 |
| 37 | 5 |
| 78 | 5 |
| 85 | 5 |
| 04 | 6 |
| 07 | 6 |
| 14 | 6 |
| 36 | 6 |
| 30 | 7 |
| 31 | 7 |
| 44 | 7 |
| 64 | 7 |
| 24 | 8 |
| 28 | 8 |
| 29 | 8 |

## Data Collection Procedure

This experimental study involved 3 rating occasions. Figure 1 illustrates the data collection procedure for the study.

Prior to the training rating occasion, a total of 103 TESL trainees assessed the operational essays (Set C). The scores given by the trainees were compared with the official scores for the essays to determine their rating accuracy. Based on the before training rating accuracy of the trainees, 63 of them were selected. The selected participants were randomly divided into three equally matched experimental groups consisting of 21 raters each. Thus, each group had the same rating accuracy before training. Then, each group was assigned to different training procedures. Finally, the training for each group was conducted two weeks after the first rating occasion.

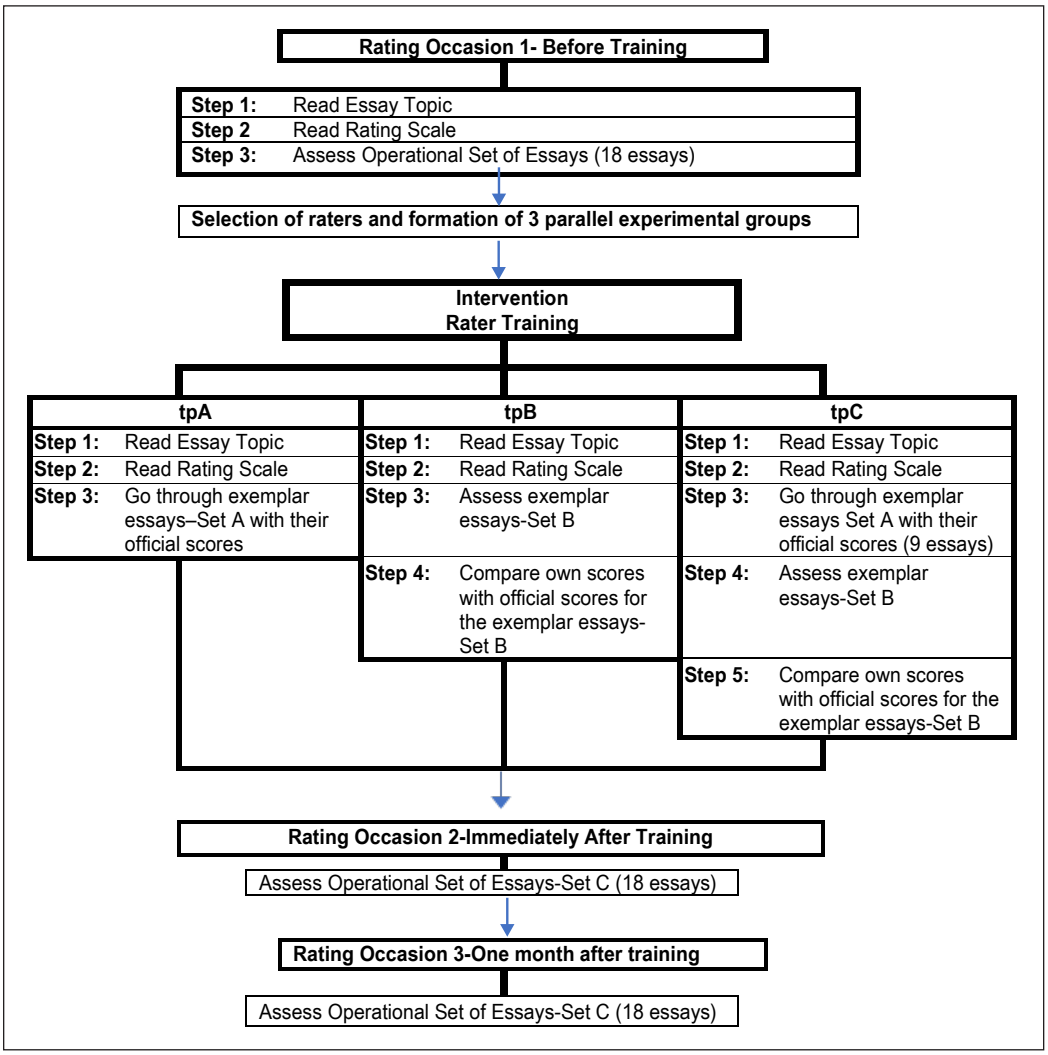


*Figure 1.* Data collection procedure

The first group, the Training Procedure A group (TpA), first read the topic for the essay and scoring rubrics. Then, they went through a set of exemplar essays (Set A) consisting of 9 benchmarked essays with their official scores. It took them approximately 45 minutes to complete TpA. The second group, the Training Procedure B (TpB) group, also read the essay topic and the scoring rubrics. Then, they assessed a set of benchmarked essays (Set B) and compared their scores with the official scores for the essays. TpB group took approximately 1.5 hours to complete the training. The third group, the Training Procedure C (TpC), went through a combination of TpA, followed by TpB. TpC took approximately 2.5 hours. Since the time taken to complete the training and assess the scripts is rather long, the raters were supplied drinks, snacks, and were allowed to have short breaks.

After the assigned training, each group assessed the same set of operational essays (Set C). Then, one month after training, each group assessed the operational essays again. The raters' rating accuracy before training, immediately after training and one month after training formed the dataset.

**Data Analysis**

For rating performance, two categories were initially formed using the data on the band difference distribution. The categories were Rating Accuracy and Rating Deviation.

Rating accuracy category refers to scores with no difference or one band difference with the Official Band. This category included all essays that differed by '0', '-1' and '+1 band from the Official Band. The reason for this category is to provide an alternative measure for accuracy by allowing a small variation from 'On-Standard' (0 band difference) as practised by most test organisations (Weigle, 2002). The number of essays scored 0–1 was calculated and converted to percentage [(number of essays within 0-1 difference/ total number of essays scored)*100]. The higher the percentage of essays in this category indicates that the rater's, or in this study, the group's rating accuracy is high as the differences in scores are within an acceptable range, which in turn suggests that the quality of the group's scoring is good.

Rating Deviation refers to scores with '≥2 band difference'. This category included all essays that differed by two and more bands, regardless of whether the band difference is 'minus (-)' as when assessed harshly or 'plus (+)' when assessed leniently. When the percentage of essays in this category is high, the group's rating is deviant from the acceptable range. Thus, the accuracy is low, suggesting that the quantitative quality of the group's scoring is poor.

The 'Within 0–1' and '≥2' categories comprise the number of essays assessed in each rating occasion. In other words, when the percentage of essays in these two categories are added, they make up 100%. Thus, the two categories dovetail with each other, and so an increase in one of these categories corresponds to a decrease in the other.

The percentage of essays scored for the 'within 0-1' category and '≥2' category was calculated to determine rating quality for descriptive statistical analysis. In addition, inferential statistics were performed to examine how each of the training procedures affects rating accuracy. For this purpose, the data were input into an SPSS file and subjected to One-Way Repeated Measures ANOVA analysis with three levels of rating occasions. They are Before Training (BT), Immediate After Training (IAT), and One Month After Training (OMAT) as the within subject-factor. In addition, a post-hoc test using the Bonferroni Adjusted Pairwise Comparison procedure was also performed to determine the extent of differences between the rating occasions. The threshold *p* value for this study was pre-determined at .05 ($p < .05$) because the commonly used *p* value is .05 for educational studies (Best, 1977; Lodico et al., 2010).

## RESULTS

The purpose of this experimental study was to investigate the effect of the different rater training procedures on raters' immediate and delayed rating performance compared to before training. Thus, the rating performance for each experimental group was calculated after every rating occasion, i.e., before training, immediately after training and one month after training. All the three experimental groups had baseline similarity, as indicated by the rating accuracy percentage before training (BT) in Table 3 and Table 4. The experimental groups had the same rating accuracy for the "within 0–1" category before training, i.e., 63% for within 0-1 band difference and 37% with '≥2 band difference. The three experimental groups were equally matched in terms of rating accuracy before training. Each experimental group's rating accuracy immediately after training and one month after training were compared to the rating accuracy before training to determine the effect of the different rater training procedures on raters' rating.

### Rating Accuracy

As shown in Table 3, immediately after training (IAT), the rating accuracy for 'within 0-1" for TpA increased to 81%, TpB to 83% and TpC to 74%. However, one month after training (OMAT), TpA's rating for within 0-1 accuracy dropped 4% to 77%, TpB's increased 1% to 84%, while TpC's increased to 78%.

Table 3
*Rating accuracy (%) 'Within 0-1 band difference*'

| Training Procedure | BT | IAT | OMAT |
|---|---|---|---|
| TpA | 63 | 81 | 77 |
| TpB | 63 | 83 | 84 |
| TpC | 63 | 74 | 78 |

### Rating Deviancy

As shown in Table 4, immediately after training (IAT), all three groups' rating deviancy for essays that were scored with two or more band differences with the official scores decreased to 19%, TpB to 17% and Tp C 26%. One month after training, TpA's rating deviancy was 23%, TpB's 16% and TpC 22%. As mentioned

earlier, the results for Rating Deviancy dovetails with the results for rating accuracy. Thus, the increase in rating accuracy within the 0–1 category corresponds with the decrease in rating deviancy.

Table 4
*Rating Deviancy (%) '≥2 band difference'*

| Training Procedure | BT | IAT | OMAT |
|---|---|---|---|
| TpA | 37 | 19 | 23 |
| TpB | 37 | 17 | 16 |
| TpC | 37 | 26 | 22 |

Figure 2 illustrates the results for rating accuracy and deviancy in graphical form.

Additionally, to examine how each training procedure affects Rating Accuracy, employing an alpha level of 0.05, the data for within 0-1 band difference were subjected to One-Way Repeated Measures ANOVA analysis, with three levels of rating occasions (BT, IAT and OMAT) as the within subject-factor. A post-hoc test using the Bonferroni Adjusted Pairwise Comparison procedure was also performed to determine the extent of differences between the rating occasions. Tables 5 shows the post hoc test results.

The results in Table 5 show that TpA's rating accuracy was at p = .005 ($p < .01$), indicating a highly significant difference immediately after training compared to before training. Furthermore, one month after training, the p value was at p =0.037 ($p < 0.05$), indicating a significant difference one month after training. Thus, $H_{01a}$ and $H_{02a}$ are rejected.

Although the rating accuracy on both occasions post-training was significantly different compared to before training, a clear inspection of Table 5 shows that the

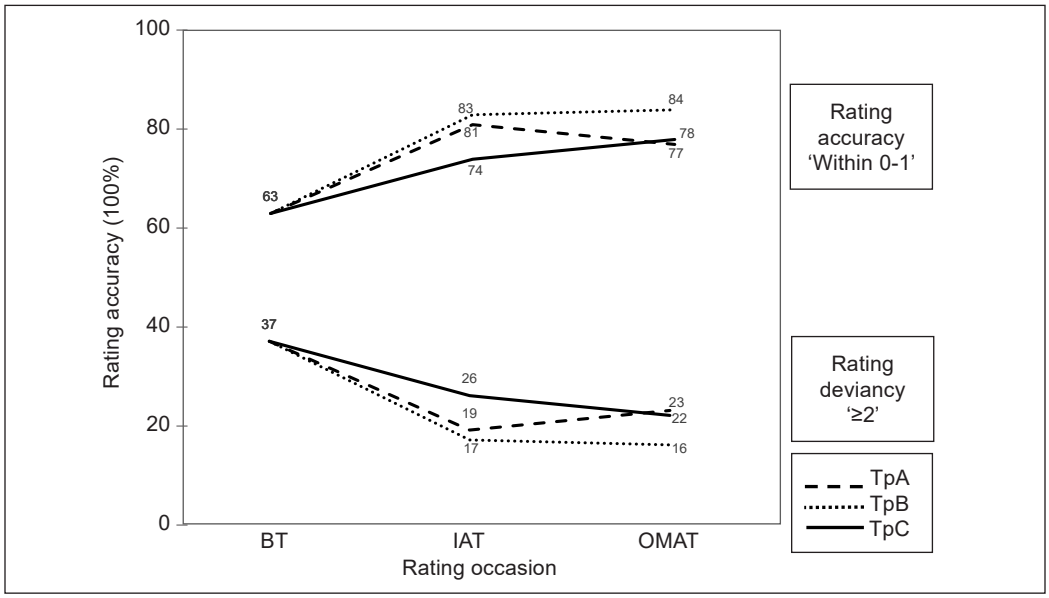


*Figure 2.* Graphical representation of TpA, TpB and TPC's rating accuracy before (BT), immediately after (IAT)and one-month after training (OMAT)

Table 5
*Post-hoc results for rating accuracy*

| Training Procedure | (I) OCCASION | (J) OCCASION | Mean Difference (I-J) | Std. Error | Sig.(a) | 99 Confidence Interval for Difference(a) | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| TpA | IAT | BT | 18.254(**) | 5.059 | .005 | 5.037 | 31.471 |
| | OMAT | BT | 13.492(*) | 4.903 | .037 | .682 | 26.302 |
| TpB | IAT | BT | 20.370(**) | 4.360 | .000 | 8.980 | 31.761 |
| | OMAT | BT | 21.164(**) | 5.081 | .001 | 7.890 | 34.437 |
| TpC | IAT | BT | 11.376 | 5.207 | .123 | 2.228 | 24.980 |
| | OMAT | BT | 12.170(**) | 3.378 | .005 | 3.346 | 20.994 |

significance level dropped from highly significant ($p < .01$) immediately after training to significant ($p < .05$) at one month after training. It suggests that the effect of TpA is showing signs of fading.

TpB results show that raters' rating accuracy at $p = .000$ ($p < .01$) was highly significant immediately after training and the $p$ value was at $p = .001$ (($p < .01$) indicating that the improvement remained highly significant one month after training. Thus, $H_{01b}$ and $H_{02b}$ are rejected.

On the other hand, TpC results show that the rating accuracy was at $p = .123$ ($p > .05$), indicating no statistical significance immediately after training compared to before training. However, the rating accuracy became highly significant one month after training with a $p$ value at .005 ($p < .01$). Therefore, based on the results for TpC, $H_{01c}$ is accepted, whereas $H_{02c}$ is rejected.

## DISCUSSION

The descriptive statistics results (frequency results) showed that all three rating procedures helped improve the raters' accuracy immediately after training and reduced the number of deviant scripts. Behind the improvement in rating accuracy lies a reduction in rating deviancy. The improvement in rating accuracy and decrease in rating deviancy indicate that raters can better understanding the scoring rubrics and standards for each band level after going through rater training. Thus, it enables them to score closer to the standard set by the testing body. These findings are consistent with previous findings that training improves raters' rating performance as reported by Attali (2015), Fahim and Bijani (2011), Furneaux and Rignall (2002), Knoch et al. (2007), Tajeddin and Alemi (2014), Wang et al. (2017) and Weigle (1994, 1998, 1999). Therefore, the results from the present study affirm that for writing performance assessment, rater training is a must to ensure better validity and reliability of the scores awarded.

The post hoc test results revealed the differences in the effect of the training procedures. The results from the scores given immediately post-training showed that TpA and TpB groups' rating accuracy

improved significantly immediately after training, surprisingly the improvement for TpC group's rating accuracy was not significant. However, one month after training, all three experimental groups' rating accuracy was significantly high, indicating a positive delayed post-training effect.

It is also interesting to note that the post host test results for TpA showed highly significant ($p$ <0 .01) improvement immediately post-training but dropped slightly and became significant ($p$ <0.05) during the delayed post-training. It suggests that TpA, which involved raters going through a set of exemplar scripts and official scores for each script, has an immediate positive and delayed effect, but it also revealed signs of fading. The results for TpA suggest that the positive effect of TpA may not be retained over a long time. Thus, retraining would be required to maintain stability in scoring. Lumley (2000, 2002), Shohamy et al. (1992) and Wang et al. (2017) have also highlighted that the effect of training may not last long and reinforced the need for retraining.

In contrast, TpB seemed to have a highly significant immediate and delayed effect on raters' rating accuracy. TpB required raters to assess the exemplar scripts and then compare their scores with the official scores. In TpA, raters were only asked to go through the exemplar essays with the official scores without rating the essays. In contrast, TpB is more hands-on because the TpB group had to score the exemplar scripts before the official scores were shown to compare scores. The results indicate that rating the scripts and comparing their scores with the official scores acts as a feedback to raters on their rating performance. Although the study's feedback is not in verbal form, Hoskens and Wilson (2001) and Leckie and Baird (2011) reported a similar effect of verbal feedback on raters' drift toward the mean leading to the homogeneity of raters' scoring.

The post hoc results for TpC were unexpected. TpC, which is a combination of TpA and TpB involved longer training and more exposure to the standards. However, TpC did not seem to improve raters' rating significantly immediately after training but showed a highly significant follow-up effect. This result is rather puzzling. The reason for this could be lethargy. Since TpC is a long training session because it is a combination of TpA and TpB, it is likely that the raters became tired and could not fully concentrate on their operational scoring immediately after training. However, the significant increase in rating accuracy during delayed post-training suggests that the effect of what the TpC group have learned during the training seem to surface sometime after training. Although the improvement in rater performance over months of rating was also reported by Lim (2011), the finding on the post-training effect of TpC in this study needs further investigation, perhaps with longer breaks during the training.

Among the three training procedures, TpB, which involved raters rating the exemplar scripts before comparing their scores with the official scores seems to have a greater effect on raters' immediate and follow-up rating accuracy. Thus, the effect of training that involves more hands on or active involvement of raters tend to have more immediate as well as longer effect on raters' rating accuracy. It is consistent with the recommendation made by Wang et al. (2017).

Overall, the results for rating accuracy seem to suggest that TpB, which is longer and more hands-on than TpA and but less time consuming than TpC, appears to be more effective for immediate and follow-up positive effects. Nevertheless, TpA would be sufficiently effective and adequate for immediate rating that does not prolong over a long period. It is also crucial to remember that long training may be exhausting to the raters and detrimental to their immediate operational rating performance. Thus, if training takes long, raters should not be asked to assess operational scripts immediately. These findings have significant implications for practical and effective rater training courses, as Shabani and Panahi (2020) emphasised.

## CONCLUSION

According to Reed and Cohen (2001), the rating is itself a performance, just as important as the test-takers performance and is thus worthy of investigation. This study adopted an experimental pre–post-follow-up approach to investigate the effects of different procedures employed during the training of writing raters on immediate and delayed rating accuracy. The findings from the study show that although different rater training procedures have a different effect on raters' rating accuracy, rater training does help raters to assess more accurately according to the standard set by the organisation, especially assessments that involve multiple raters. Considering the important decisions that educational institutions and organisations make using assessment scores, perhaps it is not an exaggeration to say that test organisations must train their raters not only to meet their professional obligations but more so for moral reasons.

In this study, the scoring rubrics, the exemplar scripts for training and operation scripts for scoring were chosen from a set of scripts from an established examination, i.e., IELTS. In addition, the official scores for the scripts were determined through MFR analysis of the scores given by expert IELTS raters. These could have contributed to the effectiveness of the training and consequently the raters' understanding of the standard required for scoring. It also highlights that for rater training to be effective, such careful and meticulous selection of benchmarked exemplar scripts are crucial. Nevertheless, the findings from this study offers crucial insights on the effects of different rater training procedures on ESL essay raters' ability to access according to the standard set by the testing organisation. Since rater training is not only time consuming but a costly process

(McIntyre 1993; Hamilton et al., 2001), it is hoped that the empirical evidence the study provides will help inform practitioners, test developers and test administration organisations in designing training for raters effectively and efficiently way. It is also hoped that the study will broaden the scope of research in the direct assessment of writing and other performance assessments such as speaking, in which similar rating procedures are typically used.

Previous research by Eckes (2008), Cumming (1990), and Wolfe et al. (1998) reported that more experienced raters considered factors that were not in the scoring rubrics. However, the raters in this study were novice ESL raters. Therefore, the exposure to the rubrics and the benchmarked scripts to these novice raters may have contributed to their adherence to the standards they were exposed to during the training. Consequently, this could have contributed to the positive effect of the rater training procedures. However, the effect may not be the same with expert raters. Thus, further research with raters of different rating backgrounds and test contexts are necessary.

Finally, the present study employed a quantitative approach to investigate the effects of rater training. However, it cannot be denied that quantitative similarities may camouflage differences in rating judgement, i.e., the reasons for awarding the scores. Thus, future studies could focus on investigating rater training effects on raters' qualitative judgement.

## ACKNOWLEDGEMENT

This study was funded by Universiti Malaysia Sarawak (UNIMAS). The author is grateful to Prof Dr Barry O'Sullivan for his guidance and feedback throughout the study.

## REFERENCES

Alderson, J. C., Clapman, C., & Wall, D. (1995). *Language test construction and evaluation*. Ernst Klett Sprachen.

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. https://doi.org/10.1177/0265532215582283

Azizah, N., Suseno, M., & Hayat, B. (2020). Severity-leniency in writing assessment and its causal factors. In *International Conference on Humanities, Education and Social Sciences (IC-HEDS) 2019* (pp. 173-185). Knowledge E. https://doi.org/10.18502/kss.v4i14.7870

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy and Practice, 18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585

Best, J. W. (1977). *Research in education* (3rd ed.). Prentice-Hall Inc.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In *International English Language Testing System (IELTS) Research Reports 20003* (Vol. 3, pp 49-84). IELTS Australia

Brown, G. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In S. D. L. H. Meyer, H. Anderson, R. Fletcher, P. M. Johnston & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40-48). Ako Aotearoa.

Chan, S. H., & Wong, B. E. (2004). Assessing oral skills of pre-tertiary students: The nature of the communicative act. In *Proceedings of the International Conference on English Instruction and Assessment* (pp. 33-48). National Chung Cheng University.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51. https://doi.org/10.1177/026553229000700104

Eckes, T. (2008). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64. https://doi.org/10.1177/0265532207071511

Ellington, J. K., & Wilson, M. A. (2017). The performance appraisal milieu: A multilevel analysis of context effects in performance ratings. *Journal of Business and Psychology*, *32*(1), 87-100. https://doi.org/10.1007/s10869-016-9437-x

Engelhard, J. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191. https://doi.org/10.1207/s15324818ame0503_1

Engelhard, J. (1996) Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

Erlam, R., Ellis, R., & Batstone, R. (2013). Oral corrective feedback on L2 writing: Two approaches compared. *System, 41*(2), 257-268. https://doi.org/10.1016/j.system.2013.03.004

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing*, *1*(1), 1-16.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-255.

Furneaux, C., & Rignall, M. (2002) *The effect of standardisation-training on rater judgments for the IELTS Writing Module*. Cambridge University Press.

Green, A. (2003). *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university professional courses* [Unpublished Doctoral dissertation]. University of Surrey.

Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement, 10*(3), 225-246.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of on-line rater training and monitoring. *System, 29*(2001), 505-520. https://doi.org/10.1016/S0346-251X(01)00036-7

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.009

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assessing Writing, 40*, 1-13. https://doi.org/10.1016/j.asw.2019.03.001

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden state examination. *Journal of Educational Measurement, 38*, 121-

145. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

IELTS. (2003). *IELTS annual review 2001/2002*. IELTS Australia.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(20), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kim, Y. S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and writing, 30*(6), 1287-1310. https://doi.org/10.1007/s11145-017-9724-6

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing*, *12*(1), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics, 14*(2), 1-23.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

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. https://doi.org/10.1177/0265532211406422

Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2010). *Methods in educational research: From theory to practice* (Vol. 28). John Wiley & Sons.

Lumley, T. (2000). *The process of assessment of writing performance: The rater's perspective.* [Unpublished Doctoral dissertation]. University of Melbourne.

Lumley, T. (2002). Assessment criteria in large scale writing test: What do they really mean to raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples* [Master's Research thesis, University of Melbourne]. Minerva Access. http://cat.lib.unimelb.edu.au/record=b1849170

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

O'Sullivan, B., & Rignall, M. (2002). *A longitudinal analysis of the effect of feedback on rater performance on the IELTS General Training writing module*. Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report.

O'Sullivan, B., & Rignall, M. (2001). *Assessing the value of bias analysis feedback to raters for the IELTS writing module*. Cambridge ESOL/The British Council/ IDA Australia: IELTS Research Report.

Raczynski, K. R., Cohen, A. S., Engelhard Jr, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301-318. https://doi.org/10.1111/jedm.12079

Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. In *Diversity and inclusion in global higher education* (pp. 73-95). Palgrave Macmillan. https://doi.org/10.1007/978-981-15-1628-3_3

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language testing essays in honour of Alan Davies* (pp. 82-96*)*. Cambridge University Press.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Rosales-Sánchez, C., Díaz-Cabrera, D., & Hernández-Fernaud, E. (2019). Does effectiveness in performance appraisal improve with rater training? *PloS one*, *14*(9), Article e0222694. https://doi.org/10.1371/journal.pone.0222694

Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process. *Practical Assessment, Research, and Evaluation, 23*(1), 1-12.

Shabani, E. A., & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing Asia, 10(1),* 1-25. https://doi.org/10.1186/s40468-020-00111-4

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes, 8*, 13-17.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Tajeddin, Z., & Alemi, M. (2014). Pragmatic rater training: Does it affect non-native L2 teachers' rating accuracy and bias. *Iranian Journal of Language Testing*, *4*(1), 66-83.

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group & Organization Management*, *25*(2), 175-190. https://doi.org/10.1177/1059601100252005

Wang, J., Engelhard Jr, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47. https://doi.org/10.1016/j.asw.2017.03.003

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283-304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating Rater/Prompt Interactions in Writing assessment: Quantitative and Qualitative Approaches. *Assessing Writing, 6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing writing*. Ernst Klett Sprachen. https://doi.org/10.1017/CBO9780511732997

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

*Testing, 10*(3), 305-335. https://doi.org/10.1177/026553229301000306

Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context* (White paper). Pearson Assessments.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*, 465-492. https://doi.org/10.1177/0741088398015004002

Xie, Q. (2015). "I must impress the raters!" An investigation of Chinese test-takers' strategies to manage rater impressions. *Assessing Writing*, *25*, 22-37. https://doi.org/10.1016/j.asw.2015.05.001