Received: July 26, 2011 *Accepted: September 15, 2011*

The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment

Mansoor Fahim[1], Houman Bijani[2]

Abstract

The assessment of writing has always been threatened due to raters' biasedness. There is evidence that rater training can be effective in eliminating extreme differences in raters' severity, and increasing the self consistency of raters by reducing individual biases of raters (Weigle, 1994a). However, there is little research documenting the exact as well as the amount of the effectiveness of training in reducing this biasedness. The purpose of this study is to investigate how judgments of raters are biased towards certain criteria before and after the training program in assessing second language essay compositions. 12 EFL raters scored 40 pre-rated benchmark essay compositions rated by an authorized IELTS trainer. These essay compositions were scored before and after the training program. The results show that most raters were able to modify their scoring, resulting in greater intergroup consistency and reduced biasedness and severity / leniency. Raters who were identified as being highly severe / lenient and biased in particular categories of the rating scale were no longer biased after training.

Keywords: *Rater training; Writing assessment; Raters' Biasedness; Raters' severity; Raters' leniency*

1. Introduction

In the past 30 years, holistic writing assessment has become the norm in evaluating writing skills in both first and second language. In holistic assessment, examinees are asked to write compositions on one or more topics, and then they are scored by raters. Because such tests are scored subjectivity, it is essential that raters be carefully trained to stick to some standards (Weigle, 1994b). These standards are given to raters through scoring rubrics that describe the characteristics of writing samples at different levels. However, it is vivid that without training, ratings tend to be highly unreliable. A large body of literature beginning with the work of

1 English Department, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: dr.mfahim@yahoo.com

2 English Department, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: houman.bijani@gmail.com

Diederich, French, and Carlton (1998) and continuing through the present day with the work of Elder, Barkhuzien, Knoch and Randow (2007), states that training is one of the most important factors in reliability of composition ratings in both first and second language context.

Linacre (1989) believes that the phenomenon of rater variation is an inevitable part of rating process of essay compositions. He claims that raters cannot be trained to achieve similar levels of severity. Therefore, the function of rater training shouldn't necessarily be to force raters into agreement with each other (interrater reliability), but rather to train raters to be self-consistent (intrarater reliability). This view of the function of rater training allows for some variability in raters variation in scoring a text which is a natural part of rating process (Stock & Robinson, 1987). Multifaceted Rasch model (Linacre, 1989) can mathematically model raters' leniency and harshness and adjust scores. As long as the raters are self-consistent, variations in rater severity do not affect scores when multifaceted Rasch model is used for scoring.

It is clear that some raters are harsher in their assessment of candidates' ability than others. Therefore, it becomes a matter of luck for examinees whether they are assessed by a particular rater. Traditional theory regarded rater characteristics in terms of the difference between an idealized rater (a perfect rater) and an actual rater (an ordinary rater). These differences between raters could be understood in terms of overall severity or leniency. Linacre (1989, as cited in Lumly and McNamara, 1995) uses the term severity both to the overall severity of the rater and to differences between raters in the way they interpret rating scales.

Typically, rater training aims to reduce variability and randomness of overall severity or leniency. The most common way of fulfilling this goal is rater training sessions, where raters are introduced with a set of criteria and then they are asked to rate based on those criteria. The results show whether and to what extent they are in line with other raters and therefore getting a common interpretation of the rating criteria. In terms of overall severity, rater training can reduce but not eradicate raters' variability. Rater training reduces extreme scores in terms of harshness and leniency and brings them in line (McIntyre, 1993).

Despite discussions about the function of rater training, little is known about what actually happens during rater training and how it affects individual raters. Charney (1984) suggests that rater training functions primarily to prevent raters from applying their own judgments. However, as Freedman (1981) suggests, rater severity is a stable characteristics which differs from rater to rater and it is not clear to what extent rater training functions to bring raters into agreement in terms of severity and leniency in rating essay compositions. Rater training can increase overall consistency through increasing intrarater consistency (McNamara, 1993). McNamara (1993) suggests that differences between raters are various. One rater may be more lenient than another or a rater may be more lenient or harsher to a particular candidate or group. Although variability cannot be entirely eliminated, rater training can also have the effect of making raters more self consistent. This is a new technique called bias analysis which is a tool for providing feedback to raters in rating writing compositions (Stahl & Lunaz, 1992). This study also focuses on bias analysis for observing the feedback and the effectiveness of a rater training program in rating writing compositions.

There is evidence that rater training can be effective in that it eliminates extreme differences in rate severity, and increases the self consistency of raters by reducing individual biases of raters (Weigle, 1994a). However, another research has shown that the effects of this training may last only for a limited time. For example, Lumly and McNamara (1995) found that some raters have large differences in rating in terms of biasedness from the rating session to the operational rating session one month later.

2. Research questions

1. Are raters severer or more lenient following the face-to-face training program?
2. Is there a reduction of individual biases in relation to scoring of particular categories of the rating scale following the face-to-face training?

3. Research hypotheses

1. Raters are not severer or more lenient following the face-to-face training program.
2. There is no reduction of individual biases in relation to the scoring a particular categories of the rating scale following the face-to-face training.

4. Methodology

In order to investigate the research questions, a quantitative quasi-experimental research design was employed in this study. The quantitative part of this study explored the differences among raters before and after training.

4.1 Participants

60 adult Iranian advanced learners of English as a Foreign Language (EFL) studying at the Advanced level of the ILI (Iran Language Institute) voluntarily participated in this study. The reason for employing advanced learners of English was due to the fact that they have already studied some courses on paragraph and essay writing. The participants included 30 males and 30 females with an age range of 18 to 42.

12 Iranian EFL teachers voluntarily participated in this study as raters. They were undergraduates and graduates in English literature, translation, linguistics, and Teaching English as a Foreign Language (TEFL). The reason for employing volunteer raters was to ensure that they would participate eagerly in all three phases of the study. These raters were different in terms of level of teaching, ranging from basic to advanced with their age ranging from 24 to 48. It should also be stated that all the raters had high level of English language proficiency although none was a native speaker of English language.

A university professor holding a Ph.D. in TEFL participated in this study as a trainer. The trainer trained raters in two training sessions and also rated all students' writing papers in the two phases of the study to serve as benchmarks for further data analysis. It should be remarked that the trainer was authorized by the IELTS as a composition rater.

4.2 Instruments

A random sample of 45 compositions from all the 60 compositions was used in this study. The compositions were selected with the help of the trainer to represent different levels of writing proficiency based on the scores given by the trainer. The rating scale used in this study is the one used by International English Language Testing System (IELTS). In the IELTS scale, scripts are rated on four aspects of writing: organization, structure, vocabulary, and punctuation. Also each student was given some instructions which clarified what they were supposed to do in the exam session. Moreover, each rater was given some instructions which clarified what they were supposed to do in rating students' essays.

4.3 Procedures

Phase 1: Pre-training data collection

In the first step, writing data were collected from 60 advanced EFL learners. Having collected the data from the students, we typed the papers exactly like what the students had written and then they were given to the trainer to rate. The purpose of giving the composition papers to the trainer was to have them served as benchmarks for data analysis. The reason for typing essay composition papers was due to the fact that raters might be influenced by students' handwritings and this would influence the true effectiveness of the training program. Finally, 15 essay compositions were given to the raters to score students on each category of the IELTS rating scale.

Phase 2: Data collection and training (norming session)

In summer 2008, when the raters finished rating the papers, the training program started. The trainer taught the way and rules to rate essay compositions based on the IELTS rating scale. Moreover, the raters were also given five additional new writing papers during the norming session to rate in pairs or groups to increase the effectiveness of the training program through giving them appropriate hints when raters gave really different scores to an essay. In this step of the study, the videotaped recordings of the norming session were given to the raters on CDs so that they could watch the CDs at home and review the necessary points.

Phase 3: Post-training data collection

When the training program was over, the researcher immediately gave the raters another 15 essay compositions to rate based on what they had acquired during the norming session. The expectation was that the raters got the desired consistency following the training program. The data analysis and results of this phase will show raters' degree of severity/leniency and thereby their biasedness in the categories of the IELTS rating scale after training. The summary of data collection and the research procedures appear in Table 1.

Table 1. Summary of data collection and research procedures

Phase	Step	Date	Procedure
Phase 1	Step 1	May 24, 2008	Data collection from students
		May 25, 2008	Data collection from students
		May 26, 2008	Data collection from students
	Step 2	May 30, 2008	Composition papers were typed
		June 5, 2008	Compositions were rated by the trainer to serve as benchmarks
	Step 3	June 19,2008	15 papers were given to the raters to rate for pre-training data collection
Phase 2		July 24, 2008	The first norming session was held
		August 7, 2008	The second norming session was held
Phase 3		August 11, 2008	15 papers were given to the raters to rate for immediate post-training data collection

5. Data analysis

In order to answer both research questions of this study, the pre-and post-training data were analyzed to get raters' degree of severity/leniency and their biasedness in the particular categories of the IELTS rating scale. In this regard, the total discrepancy score for each rater as well as each category of the rating scale was measured and compared to that of the trainer, called the *benchmark*. Discrepancy score, according to Elder et al. (2007), is the score given by the raters minus the scores given by the trainer to that particular essay composition. However, discrepancy scores are raw scores and not analyzable. Therefore, in order to be able to analyze them, they were converted into z-scores and based on z-values the analyses were done (Stahl & Lunaz, 1992). Then through measuring the average z-scores of each rater holistically and that of the benchmark as well as each category of the rating scale for each rater and that of the benchmark, it was possible to get the final z-value.

5.1 Results

RQ1: Are raters severer or more lenient following the face-to-face training program?

In order to understand whether the raters became severer or more lenient after the training program, raters severity and leniency were measured for the two phases of the study i.e. before and after training.

Raters' severity/leniency prior to the training program

The pre-training data were collected through the rating of 15 essay papers by the raters. By comparing the obtained data to the benchmark the discrepancy between benchmark scores and raters' scores was calculated. Since the discrepancy scores were raw scores they were converted into z-scores and then the final discrepancies in z-values were obtained (Stahl & Lunaz, 1992). The results show whether and to what extent raters are severer or more lenient than the trainer. Positive values show that raters were more lenient than the benchmark and negative values show that they were severer than the benchmark. Table 2 shows the degree of raters' severity or leniency prior to the training program. It is vivid that all raters had some degrees of severity or leniency compared to the benchmark. The z-value indicates any departure from what is expected of that rater from normal variation.

Raters 2 and 12 were the most severe raters with the severity value of -3.1 and -3.8, respectively. Raters 5, 7 and 9 were the most lenient raters with the leniency value of 2.4, 3.1 and 2.9, respectively. Other raters have some degrees of severity or leniency too and deviate from the benchmark scores.

Table 2. Raters' degree of severity or leniency prior to the training program

Raters	Degree of severity / leniency
Rater 1	-0.4
Rater 2	-3.1
Rater 3	-0.8

Rater 4	-0.9
Rater 5	+2.4
Rater 6	-1.3
Rater 7	+3.1
Rater 8	+0.8
Rater 9	+2.9
Rater 10	-0.4
Rater 11	+1.1
Rater 12	-3.8

Raters' severity/leniency following the training program

Like the pre-training phase, the post-training data were analyzed to get the outcome of the training program. Another 15 essay compositions were rated at the post-training phase of the study and thereby the discrepancy scores were calculated. Table 3 shows the degree of raters' severity or leniency following the training program.

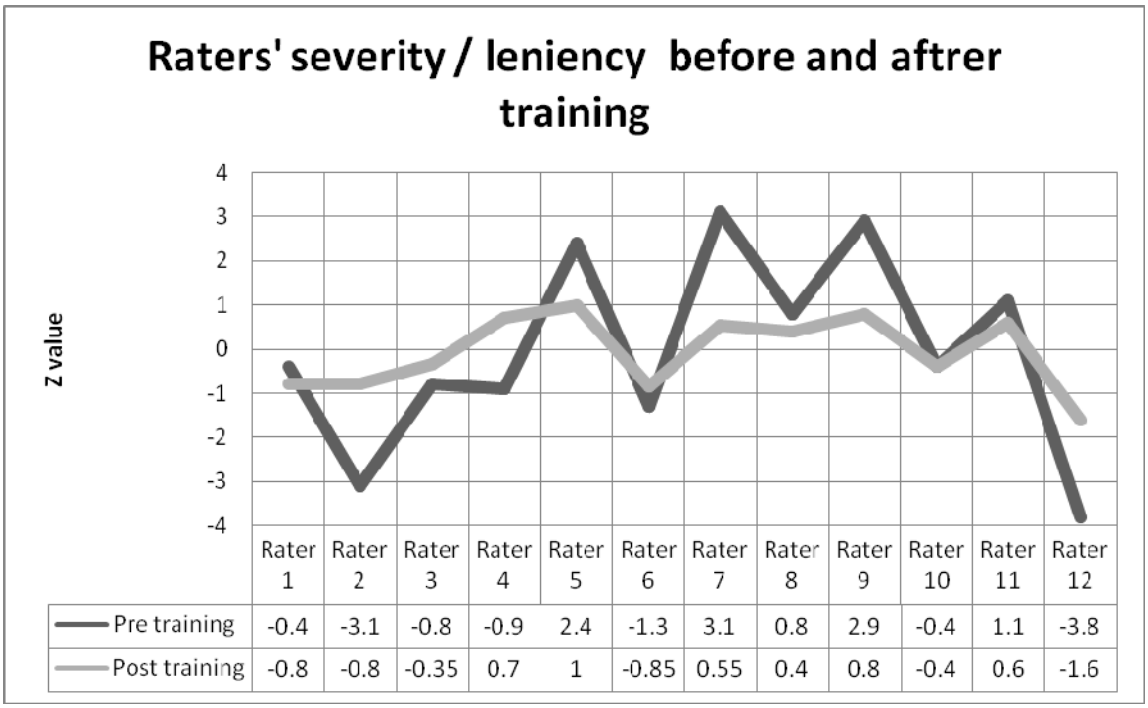
Table 3. Raters' degree of severity or leniency following the training program

Raters	Degree of severity / leniency
Rater 1	-0.8
Rater 2	-0.8
Rater 3	-0.35
Rater 4	+0.7
Rater 5	+1
Rater 6	-0.85
Rater 7	+0.55
Rater 8	+0.4
Rater 9	+0.8
Rater 10	-0.4
Rater 11	+0.6
Rater 12	-1.6

Through comparing the outcomes of Table 2 to those of Table 3, we find out about the effectiveness of the training program. The very first insightful thing is that the discrepancy scores for all raters are smaller and closer to the benchmark. Raters 2 and 12 were too severe before the training program with the severity of -3.1 and -3.8, respectively, however, after training their level of severity was lowered to -0.8 and -1.6. This is a great change and shows that the training program was very effective for them. Raters 5, 7, and 9 were too lenient with leniency of 2.4, 3.1 and 2.9. After training their level of leniency was lowered to 1, 0.55 and 0.8, respectively. The degree of severity or leniency for other raters including Rater 3, 6, 8 and 11 was also moderated and lowered to a great extent. Rater 4 was severe before training but moved

to be lenient after training; her ratings became closer to the benchmark though. Raters 1 and 10 showed a strange behavior compared to others. Rater 10 did not show any change from pre-training to post training in terms of severity. To our great surprise, Rater 1 became harsher after. She showed a moderate degree of severity of -0.4 before training but after training her level of harshness got severer to -0.8. Figure 1 shows raters' change of behavior in terms of severity or leniency in pre-training compared to that of post-training phase.

Figure 1. Raters' change of behavior in term of severity and leniency before and after the training program



	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10	Rater 11	Rater 12
Pre training	-0.4	-3.1	-0.8	-0.9	2.4	-1.3	3.1	0.8	2.9	-0.4	1.1	-3.8
Post training	-0.8	-0.8	-0.35	0.7	1	-0.85	0.55	0.4	0.8	-0.4	0.6	-1.6

Furthermore, the dispersion of rater severities decreased to a considerable extent. Through calculating the standard deviation of discrepancy scores in the pre-training stage, the dispersion index was lowered from 1.33 in pre-training ratings to 0.69 in post-training ratings. This shows the effectiveness of the training program in making the raters consistent with each other. The rater dispersion index shows that raters scored more in line with each other after training. According to Knoch, Read, and Randow (2007), the closer the dispersion index is to zero, the closer the raters are in terms of severity.

RQ2: Is there a reduction of individual biases in relation to scoring of particular categories of the rating scale following the face-to-face training?

Bias analysis shows there is a consistent interaction of a rater with a certain aspect of the rating scale. In this case, we tried to identify any significant biasedness with respect to a certain category of the rating scale (i.e. organization, structure, vocabulary, and punctuation). In order to understand whether and to what extent the raters were biased in the two phases of the study and

also whether their behavior in terms of biasedness changed or not, it was decided to study their biasedness in each particular category in pre-training and post-training.

Raters' biasedness prior to the training program

Having collected the pre-training data, the discrepancy scores for each category of the rating scale was calculated. To this point, each rater's score given to each category (organization, structure, vocabulary and punctuation) of the rating scale was compared to that of the benchmark. However, as mentioned in research question one, these discrepancy scores were converted into z-scores so that analyses could be applied to them. Hereby, by computing the average z-scores in each of the categories of the rating scale for each rater and the benchmark, the final z-value for each category was obtained. Table 4 shows the degree of raters' biasedness on each rating category prior to the training program.

Table 4. Raters' biasedness in each trait prior to the training program

Raters	Organization	Structure	Vocabulary	Punctuation
Rater 1	-0.3	-0.3	-0.4	-0.6
Rater 2	-3.6	-2.4	-2	-0.4
Rater 3	-0.6	-1.2	-0.8	+0.3
Rater 4	+0.6	+0.8	+0.8	+2.5
Rater 5	+2.4	+2	+2	+3.3
Rater 6	-1.5	-1.2	-1.6	0
Rater 7	+2.7	+2.4	+2.4	+3.6
Rater 8	+0.6	+0.2	+0.8	+1.5
Rater 9	+3	+2.4	+2	+3.2
Rater 10	-0.3	-1.2	-0.8	+2.7
Rater 11	+1.2	+0.4	+0.8	+1.8
Rater 12	-3.8	-2.4	-2.8	-2.4

The z-values show if individual raters rated a certain trait harshly or leniently. It is clear from the table that all the raters showed some biasedness before the training program. In terms of "organization", Raters 2 and 12 were too negatively biased and they were too severe before training with a degree of harshness of -3.6 and -3.8, respectively. Raters 5, 7 and 9, however, were too lenient with the degree of leniency of +2.4, +2.7 and +3, respectively. In terms of the "structure", Raters 2 and 12 were also too severe, with the degree of severity of -2.4 for both, and Raters 7 and 9 were too lenient with the degree of leniency of +2.4 for both. Rater 5 was a lenient rater with a discrepancy score of 2 before the training. Regarding "vocabulary", Raters 12 and 2 were severe with the degree of severity of -2.8 and -2, respectively. Rater 7 was too lenient in this regard with the degree of leniency of +2.4. Considering "punctuation", just Rater 12 was too severe with the degree of severity of -2.4 and Raters 4, 5, 7, 9, and 10 were too lenient with the degree of leniency of +2.5, +3.3, +3.6, +3.2, and +2.7, respectively. Surprisingly, Rater 6 was not biased at all on punctuation category before training.

Raters' biasedness following the training program

Like the pre-training phase, the post-training data were analyzed to get the outcome of the training program. To this point, another 15 essay compositions were rated at the post-training phase of the study and thereby the discrepancy scores were calculated. These discrepancy scores were converted into z-scores and then through measuring the average z-value of each rater in each particular category of the rating scale and that of the benchmark, the final z-value were obtained. Like before, these z- values indicate whether and to what extent raters were still biased in rating essay compositions after the training program. Table 5 shows the degree of raters' biasedness in each trait following the training program.

Table 5. Raters' biasedness in each trait following the training program

Raters	Organization	Structure	Vocabulary	Punctuation
Rater 1	-1.5	-0.4	-0.4	-1
Rater 2	-1.5	-0.4	-0.8	-0.15
Rater 3	-0.9	-0.3	-0.7	0
Rater 4	+0.9	+0.6	+0.3	+0.8
Rater 5	+0.9	+0.15	+0.6	+1.7
Rater 6	-1.2	-0.7	-0.7	-0.9
Rater 7	+0.3	+0.85	+0.2	+1.2
Rater 8	+0.3	+0.7	+0.4	+0.3
Rater 9	-0.45	+0.7	+0.3	+1.3
Rater 10	-0.3	+0.2	-0.4	0
Rater 11	+0.3	+0.3	+0.7	+1.2
Rater 12	-1.4	-1.3	-1.1	-1.5

Although still raters showed some biasedness after training, their level of biasedness reduced to a great extent. After training no rater showed any degree of biasedness more or less than ±2 z-score. In terms of "organization", Raters 2 and 12 were too harsh before training with the degree of harshness of -3.6 and -3.8. But after training their harshness was reduced to -1.5 and -1.4, respectively. Raters 5, 7 and 9 were too lenient with the degrees of leniency of +2.4, +2.7 and +3, but after training their leniency was reduced to +0.9, +0.3 and +0.45, respectively. In terms of "structure", Raters 2 and 12 were too harsh with a harshness of -2.4, but after training their degree of harshness was reduced to -0.4 and -1.3, respectively. Raters 7 and 9 ware also too lenient with the degree of leniency of +2.4 for both but after training their degree of leniency was reduced to +0.85 and +0.7, respectively. Rater 5 who was lenient turned to have a discrepancy score of +0.15 after training. Regarding, "vocabulary", Raters 12 and 2 were too severe but after training they changed their biasedness to -1.1 and -0.8, respectively. Rater 7 was also too lenient with the degree of leniency of +2.4 but after training it was moderated to +0.2. Considering "punctuation", Rater 12 was too severe with the degree of severity of -2.4 but she changed it to -1.5 after training. Raters 4, 5, 7, 9 and 10 were too lenient before training with the degree of leniency of +2.5, +3.3, +3.6, +3.2 and +2.7. However, they were moderated after training with the degree of leniency of +0.8, +1.7, +1.2, +1.3 and 0. Raters 3 and 10 did not show any biasedness on this category after training. To our great surprise, Rater 1 moved further away from before the training program and she became severer in organization, structure and punctuation after training. Other raters who moved away compared to pre-training stage are

Rater 3 in organization, who moved from -0.6 to -0.9, rater 6 who moved from 0 to -0.9 in punctuation and Rater 8 who moved from +0.2 to +0.7 in structure after training. All in all, except Rater 1, who did not reduce her biasedness and did not match to other raters' behavior after training, all the other raters turned to be less biased after training and it shows that the training was quite effective.

Furthermore, the dispersion of raters' biases decreased to a great extent. Through comparing the mean and standard deviation of discrepancy scores of each trait in the pre-training and post-training phases, we found out that the raters became less biased and more consistent in all the traits. Table 6 lists raters' dispersion index in all traits before and after training.

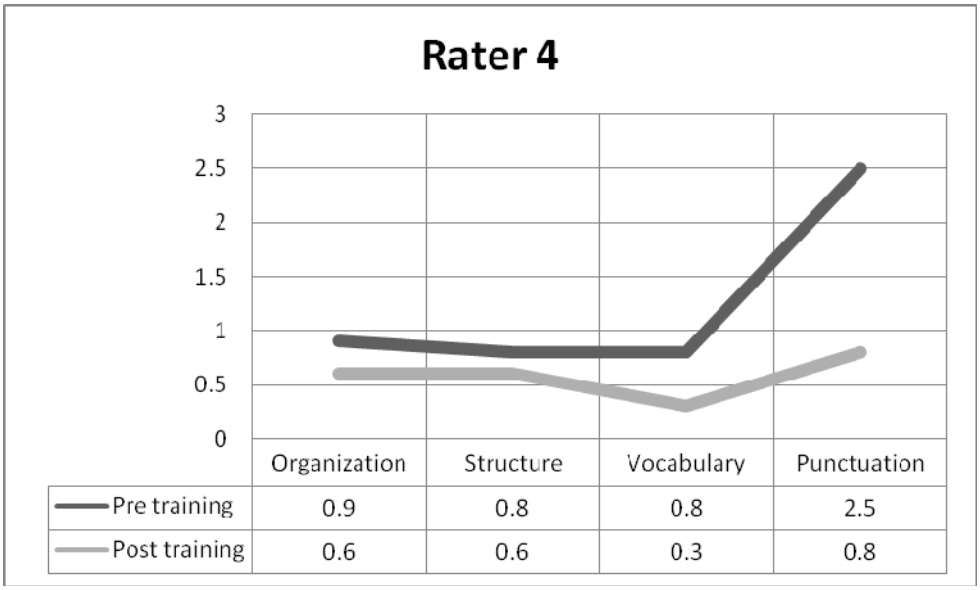
Table 6. Raters' dispersion index in rating scale traits before and after training

Study Phase	Rating trait	Dispersion index
Pre-training	Organization	0.78
	Structure	0.39
	Vocabulary	0.42
	Punctuation	0.27
Post-training	Organization	0.42
	Structure	0.29
	Vocabulary	0.26
	Punctuation	0.14

Studying Table 6 shows the effectiveness of the training program in aligning raters' ratings and therefore, increasing inter-rater reliability. Wigglesworth (1993) states that the closer the dispersion index is to zero, the less the raters are biased. Therefore, after training, raters' amount of biasedness was decreased greatly.

Figure 2 shows Rater 4's change of behavior in the four rating scale categories in terms of biasedness in pre and post-training stages.

Figure 2. Rating behavior for Rater 4 in pre and post training phases



Figures 3 to 13, provided in the Appendix, represent changes in the ratings of other raters in pre and post training phases. Training was all effective for Rater 4. She lost biasedness greatly in all the traits. Her degree of biasedness reduced to a considerable extent on all the categories of the rating scale. This rater was too lenient in punctuation at the pre-training phase; however, after training she was much closer to the benchmark. To see the effectiveness of training on other raters please refer to the Appendix.

6. Discussion and Conclusion

The major findings of this study include the followings. All raters became highly consistent after training. Training reduced raters' severity and harshness to a great extent but did not eliminate it. Training also reduced raters' biasedness but did not eliminate it altogether, that is, training seems to have brought the extreme scores within a moderate range of biasedness. This is in line with the findings of Stahl and Lunaz (1992) that rater training cannot eliminate differences among raters in terms of severity and biasedness. In terms of biasedness, although Rater 1's overall consistency improved after training, she ended up showing some more bias after training in organization, structure, and punctuation. Two other raters also showed more bias in a category of training, but they improved in reducing bias. These findings are consistent with the findings of Hamilton, Reddel, Spratt (2001). According to Shohamy, Gordon, and Kramer (1992), the raters who do not reduce their bias after training should be discarded because they deviate from the norm. In terms of raters' attitude to the training program, those raters whose rating behavior improved little after training tended to be somewhat less positive in their attitude to the training program compared to those whose rating was greatly developed. For example, Rater 11 did not have a positive view about training and its effectiveness and thus his improvement after training was little. Although causal connections between attitudes and outcomes cannot be assumed, it is said that if any training is done in a friendly atmosphere, it would be more effective (Hamilton et al., 2001). On the other hand, those raters who accepted authorities' comments tended to move more closely to the benchmark (as suggested by Reed & Cohen, 2001). Most raters had very

positive attitudes to the feedback received and considered it as a useful component of the training program. Most found improvements in their ratings as a result of face-to-face training.

Acknowledgements

We are indebted to Dr. Cushing Weigle and Dr. Elder for their helpful comments on this project. We would also like to thank Dr. Asadi for training the raters and all the raters who graciously gave their time and efforts for this research.

References

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

Diederich, P. B., French, J. E., & Carlton, S. T. (1998). *Factors in judgments of writing ability*. Educational Testing Service. Priceton, Nj.

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

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English*, 15 (3), 245-55.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of online rater training and monitoring . *System*, 29, 505-20.

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

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

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

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teachers' assessment of ESL writing samples*. Unpublished MA thesis. University of Melbourne.

McNamara, T. F. (1993). *Second language performance assessment*. Unpublished manuscript.

Reed, D. J. & Cohen, A. D. (2001). Revising 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: Essays in honor of Allan Davies*. Cambridge: Cambridge University Press.

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

Stahl, J. A. & Lunaz, M. E. (1992). *A comparison of generalizability theory and multi-faceted Rasch measurement*. Paper presented at the Midwest Objective Measurement Seminar, Chicago, IL.

Stock, P.L. & Robinson, J.L. (1987). Talking on Testing. *English Education*, 19, 93-121.

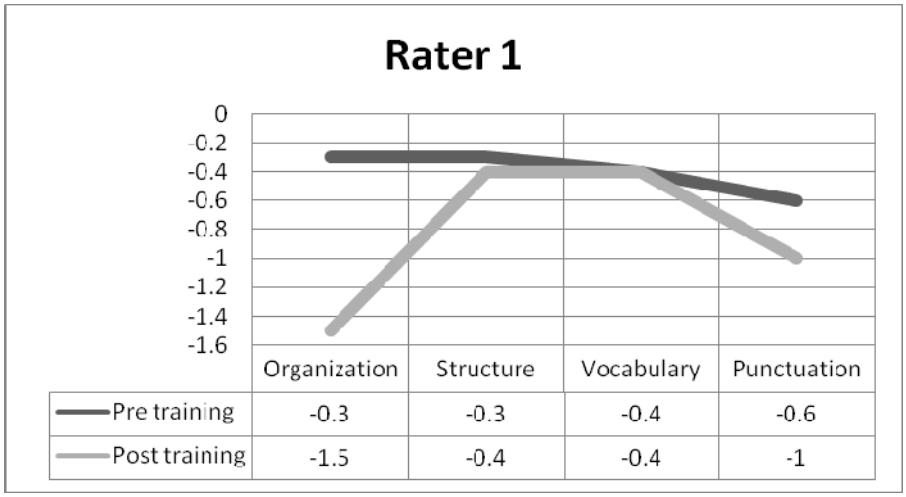
Weigle, S. C. (1994a). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223.

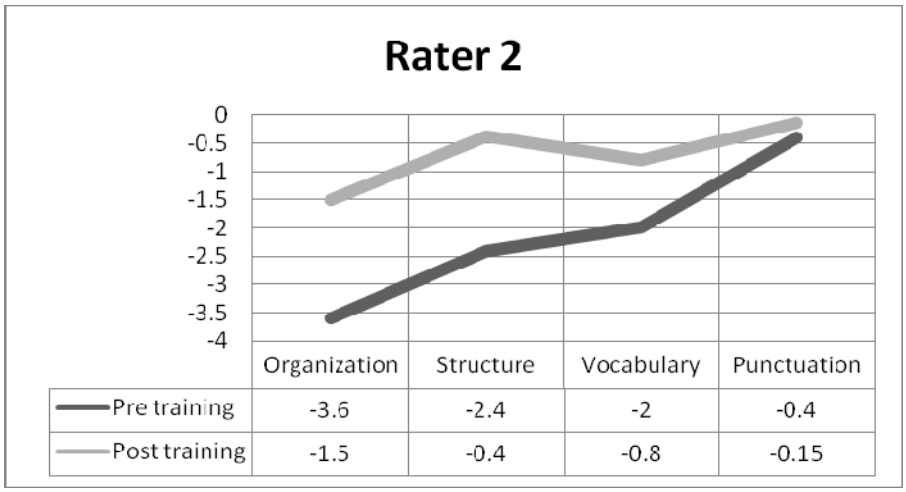
Weigle, S. C. (1994b). *Effects of training on raters of English as a second language compositions: Qualitative and quantitative approaches*. Unpublished Ph.D dissertation, University of California, Los Angeles.

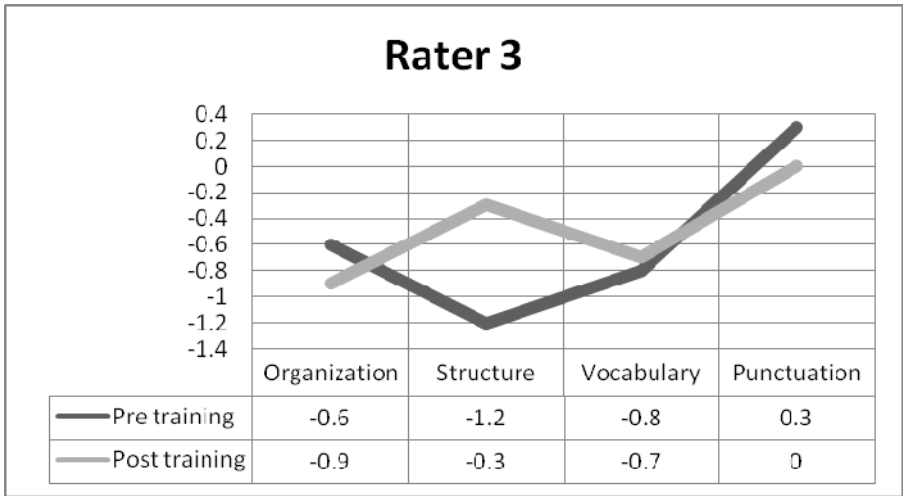
Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305-23.

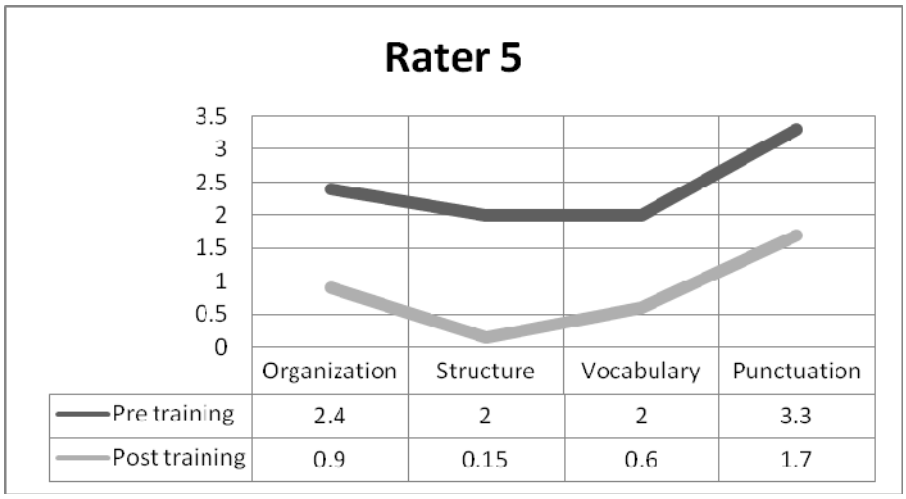
Appendix

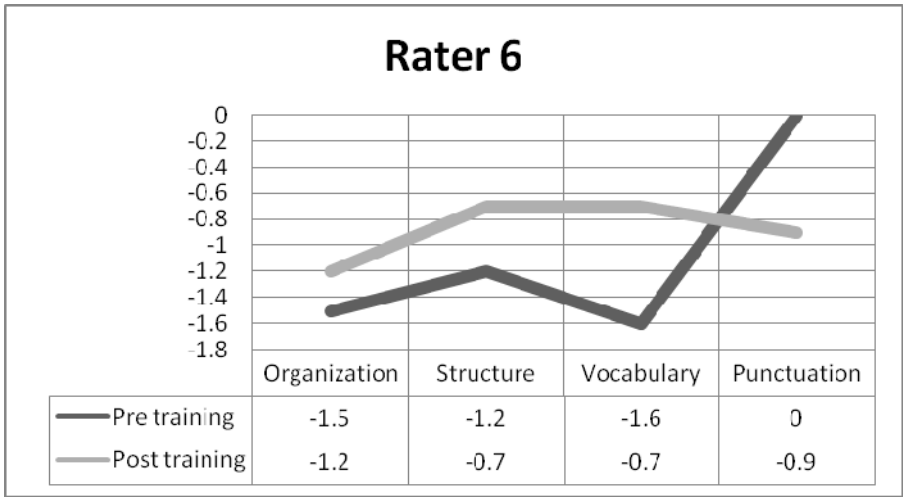
Rating behavior graphs for pre and post training phases

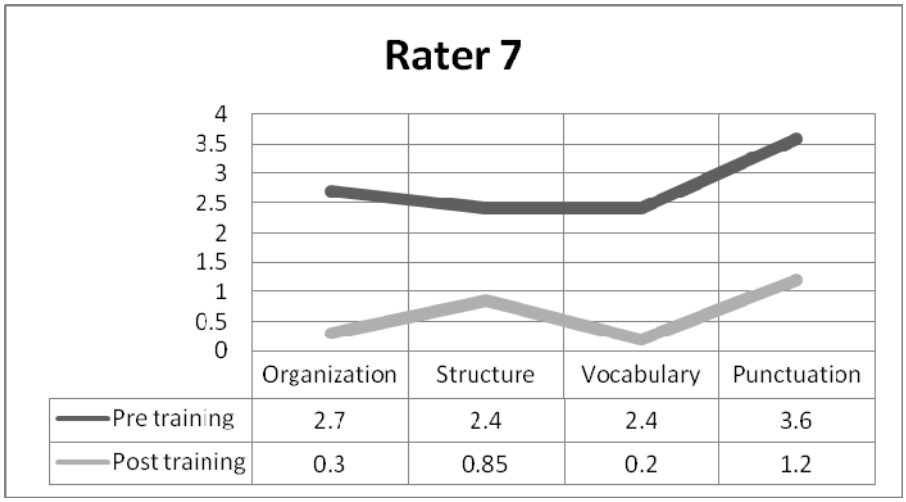


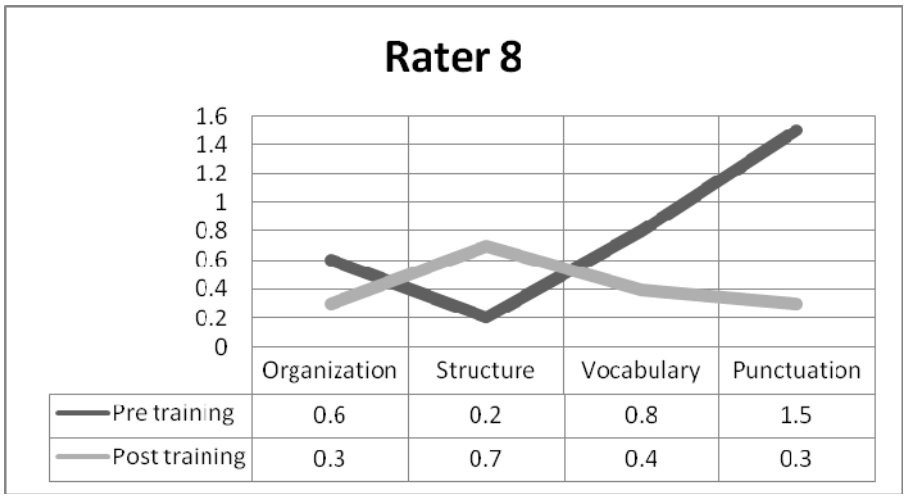


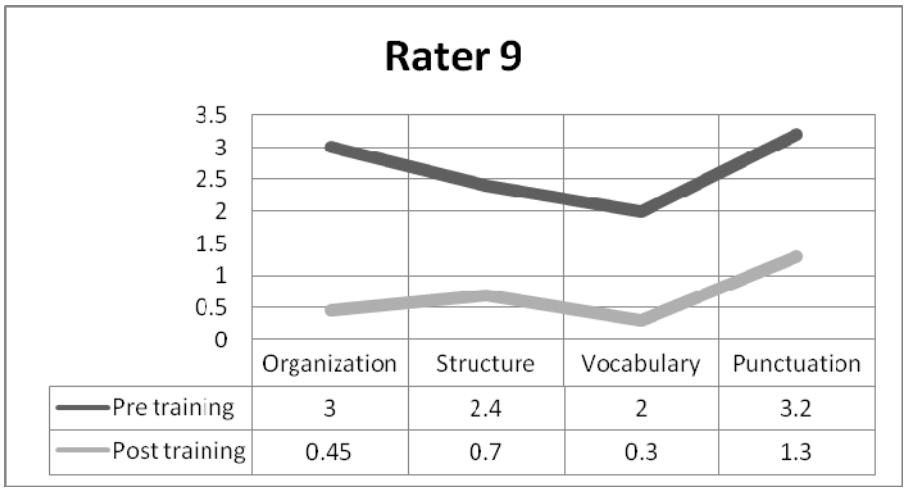

Rater 3
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
Organization
Structure
Vocabulary
Punctuation
Pre training
-0.6
-1.2
-0.8
0.3
Post training
-0.9
-0.3
-0.7
0

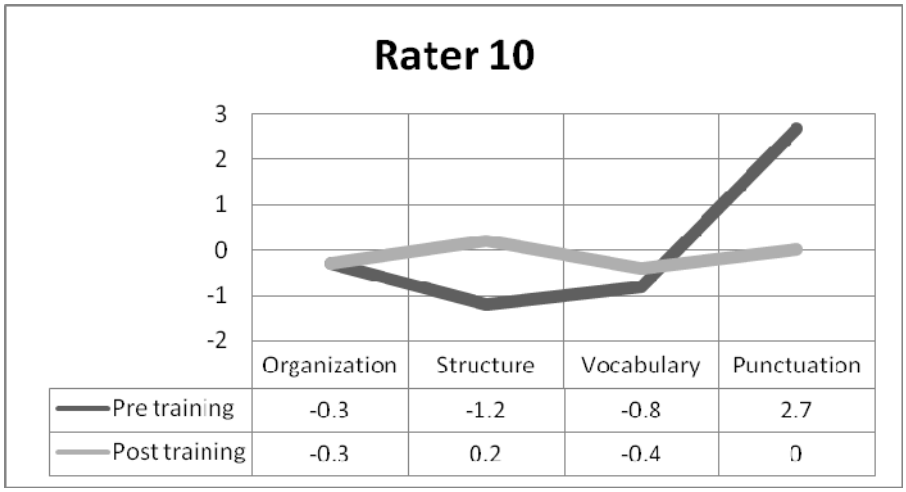

Rater 5
3.5
3
2.5
2
1.5
1
0.5
0
Organization
Structure
Vocabulary
Punctuation
Pre training
2.4
2
2
3.3
Post training
0.9
0.15
0.6
1.7

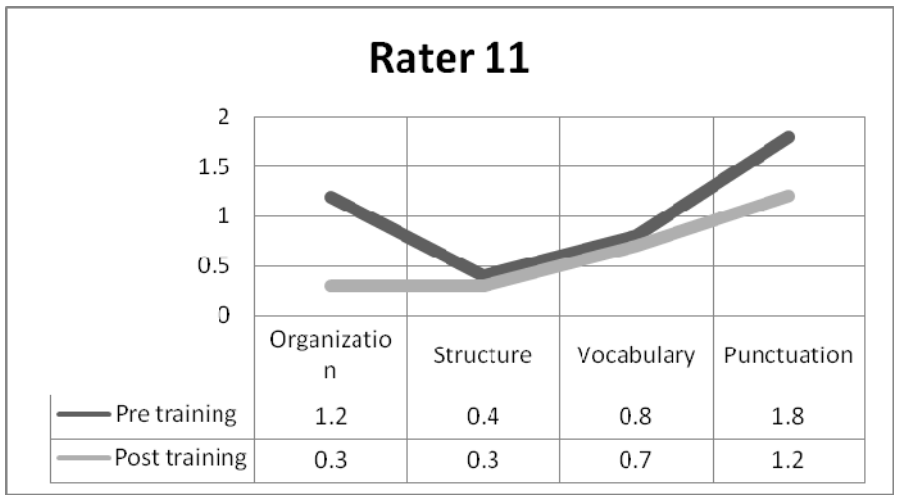

Rater 6
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
Organization
Structure
Vocabulary
Punctuation
Pre training
-1.5
-1.2
-1.6
0
Post training
-1.2
-0.7
-0.7
-0.9

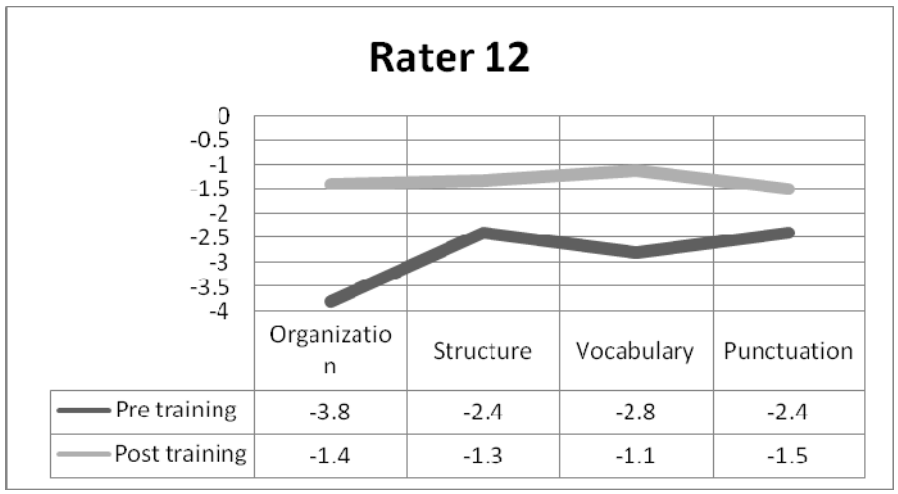

Rater 7
4
3.5
3
2.5
2
1.5
1
0.5
0
Organization
Structure
Vocabulary
Punctuation
Pre training
2.7
2.4
2.4
3.6
Post training
0.3
0.85
0.2
1.2


Rater 8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Organization
Structure
Vocabulary
Punctuation
Pre training
0.6
0.2
0.8
1.5
Post training
0.3
0.7
0.4
0.3


Rater 9
3.5
3
2.5
2
1.5
1
0.5
0
Organization
Structure
Vocabulary
Punctuation
Pre training
3
2.4
2
3.2
Post training
0.45
0.7
0.3
1.3


Rater 10
3
2
1
0
-1
-2
Organization
Structure
Vocabulary
Punctuation
Pre training
-0.3
-1.2
-0.8
2.7
Post training
-0.3
0.2
-0.4
0


Rater 11
2
1.5
1
0.5
0
Organizatio n
Structure
Vocabulary
Punctuation
Pre training
1.2
0.4
0.8
1.8
Post training
0.3
0.3
0.7
1.2


Rater 12
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Organizatio n
Structure
Vocabulary
Punctuation
Pre training
-3.8
-2.4
-2.8
-2.4
Post training
-1.4
-1.3
-1.1
-1.5