The Journal of Applied Linguistics Vol. 3, Issue 2
Fall 2010

Raters' Perception and Expertise in Evaluating Second Language Compositions

Houman Bijani[1]
Zanjan Branch, Islamic Azad University, Zanjan, Iran

The consideration of rater training is very important in construct validation of a writing test because it is through training that raters are adapted to the use of students' writing ability instead of their own criteria for assessing compositions (Charney, 1984). However, although training has been discussed in the literature of writing assessment, there is little research regarding raters' perceptions and understandings of the training program. Although a few studies have looked at the differences between trained and untrained raters in writing assessment (Cumming, 1990; Huot, 1990), few studies have used a pre-and post-training design. The purpose of this study is to investigate the effectiveness of the training program on experienced and inexperienced raters with regard to a pre-and post- training design. Twelve EFL raters scored 45 pre-rated benchmark essay compositions by an authorized IELTS trainer. These essay compositions were scored before, during and after the training program. The results regarding the comparison across raters showed that inexperienced raters had wider range of inconsistency before training but they became more consistent than experienced raters after training. Keywords: Experienced Raters, Inexperienced Raters; Interrater Reliability; Rater Training; Writing Assessment

In recent years, essay examination has become a standard way in assessing the writing skills of both first and second language speakers of English. Writing assessment needs subjective evaluation of writing skills by raters. This subjectivity is a

[1] Corresponding author. E-mail:

potential threat to test validity in that the score a candidate gets may have little to do with his/her real writing ability.

It is well-proved that raters do not always agree on composition scores. One such example is found in a study by Diederich, French, Carlton (1998), who had the 300 essays read by 53 untrained raters on a nine point scale. Of these essays, 94% received at least seven different scores. A great deal of research has been done on the raters' agreement regarding the scores they give to students' writings. Linacre (1989) introduces the "true score" approach to the problem of rater variation. True score approach is that a candidate's test score is composed of a true score, which is related to the candidates ability, and an error score, which is due to other factors (Bachman, 1990). The consideration of rater training is very important in construct validity of a writing test because it is through training that raters are conformed to the use of students' writing ability instead of their own criteria for assessing compositions (Charney, 1984).

The study of rater variables in writing assessment has got two important aspects: the attributes of writing on which raters focus and the effect of raters' background on the process of reading essays and the scores they give. During recent years, researchers have focused their attention on raters' characteristics which may affect their ratings. Ruth and Murphy (1988) compared the holistic ratings of high school students, novice teachers, and expert teachers on 114 student essays. They found that student ratings were significantly lower than expert teacher ratings with novice teacher ratings in between. Similarly, Cumming (1990), found that novice raters were significantly more lenient in their rating of rhetorical organization and content than expert raters. Moreover, in their study, Breland and Jones (1984) found that experienced raters were significantly severer in their ratings than inexperienced raters in essay scoring, i.e., experienced raters tended to be more strict and gave lower scores in rating students' compositions compared to inexperienced ones.

Shohamy, Gordon, and Kraemer (1992) compared the scoring of trained and untrained raters from two different backgrounds: experienced English teachers and nonteachers. They

found that the variable of training was more influential than the variable of background in terms of reliability. Stalnaker (1934, cited in Weigle, 1994b) stated that the rater who underwent a strict training program had reliabilities ranging from 0.73 to 0.98 whereas the reliability before training was as low as 0.30.

One important rater variable is their expectations about students writing. For example, Diederich et al. (1998) found that when raters were told that an essay was written by an honor student, they would give higher scores. Huot (1990) stated the same issue and further remarked that raters' expectations and experience are important parts of reading and rating process.

However, although training has been discussed in the literature of writing assessment, there is little research regarding of raters' perceptions of the training program. Only a few studies have considered the differences between trained and untrained raters in writing assessment (Cumming, 1990; Weigle, 1994a), and few ones have used pre-and post-training design (Elder, Barkhuizen, Knoch&Randow, 2007; Hamilton, Reddel& Spratt, 2001) . This study aims to explore the effects of training on raters' perceptions of the training program along with the effectiveness of the training program on experienced and inexperienced raters.

Research Questions

1. What are the raters' perceptions regarding the rater training program before and after training?
2. Do experienced and inexperienced raters differ in their scoring of compositions before and after training?

Method

In order to investigate the research questions, a quasi-experimental research design was employed in this study to compare the raters' perceptions, behaviors, and agreement before and after the training program. The quantitative part of this study explored the differences among the raters before and after rater training and the qualitative part explored the responses of the raters to the training questioners and interviews.

Participants

60 adult Iranian advanced learners of English as a Foreign Language (EFL) voluntarily participated in this study. These participants included 30 males and 30 females with an age range from 18 to 42. Also, 12 Iranian EFL teachers voluntarily participated in this study as raters. They were undergraduate and graduate in English literature, translation, linguistics, and Teaching English as a Foreign Language (TEFL). The reason for using 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. The raters were assigned to two groups based on their experiences in teaching and rating compositions.

A. The raters who had never rated compositions and had never been exposed to composition prompts or scoring guides or procedures. Hereinafter, we call these raters as **NEW**. This group included 6 raters with five to eight years of teaching experience.

B. The experienced raters were EFL teachers who had already taught and rated compositions using different scales. They were quite familiar with the rating scales and composition prompts. Hereinafter, we call these raters as **OLD**. These raters included 6 raters with at least two years of experience in teaching and rating compositions.

A university professor holding Ph.D. in TEFL, with about 16 years of experience in teaching and rating essay compositions professionally participated in this study as a trainer. The trainer trained raters in two training sessions and also rated all students' writing papers in two phases of the study, i.e., before and after the rater training program, to serve all as benchmarks for further data analysis. It should be remarked that the trainer was authorized by the IELTS as a composition rater. Table 1 summarizes the characteristics of all 12 raters.

Table 1.
Raters' Characteristics

Rater	Gender	Degree	Years of experience in teaching English	Years of experience in rating composition
NEW1	F	B.A. in translation	5	0
NEW2	F	B.A. in literature	5	0
NEW3	F	M.A. in linguistics	5	0
NEW4	F	B.A. in literature	6	0
NEW5	F	B.A. in literature	6	0
NEW6	F	B.A. in translation	8	0
OLD1	M	M.A. in literature	9	4
OLD2	F	M.A. in TEFL	8	4
OLD3	M	M.A. in TEFL	8	5
OLD4	M	B.A. in literature	13	6
OLD5	M	B.A. in translation	15	3
OLD6	F	M.A. in literature	19	10

Instruments

A stratified random sample of 45 compositions from all the 60 compositions was used in this study. The reason for omitting the remaining 15 compositions was that they were not in accordance with the instructions the students were supposed to write their compositions. 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. These 60 compositions were written by 30 male and 30 female students.

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. The four aspects are weighted on a 9 point Likert scale ranging from 1 to 9, scoring 1 to 4 as "seldom accurate", 5 as "occasionally accurate", 6 as "usually accurate", 7 as “often accurate”, 8 as “mostly accurate”, and 9 as “completely accurate”. The final score in each phase of the study was obtained via getting the average of the scored gained by the students in each category of the IELTS rating scale exactly as done by the ETS. The pre-and post training questionnaires used in this study were aimed to focus on individual rater’s degree of experience, attitude, expectations, effectiveness and evaluation of the training program. The pre-and post training questionnaires had originally been developed by Elder et al. (2007); however, in order to make them suitable for this study, they were modified. Also each student was given an instruction which clarified what they were supposed to do in the exam session. Moreover, each rater was given a written instruction which clarified what they were supposed to do in rating the students’ essays. The norming session was also videotaped and the videotaped recordings of the norming session were given to the raters in CDs so that they could watch the CDs at home and fully get the necessary hints they might have lost in the norming session.

Procedures

Phase 1:Pre-training data collection

Step 1: In the first step, data collection from students was done. This step took three days for the data collection job from the students was a big deal. In this regard, 60 advanced EFL learners participated. The reason for choosing advanced learners of English was that these students had already acquired the adequate knowledge regarding essay writing and paragraph development.

Step 2: Having collected the data from the students, the researchers had the papers typed exactly like what the students had written, and then, gave them to the trainer to rate. The purpose of giving the composition papers to the trainer was to have them served as benchmarks for the data analysis. Note that the reason for typing essay composition papers and then giving them to the raters to score was that the raters might be influenced by the students' handwritings, which would wrongly influence the effectiveness of the training program.

Step 3: Fifteen essay compositions were given to the raters to score each student's paper for each category of the IELTS rating scale. The scale was attached to each essay composition paper. The categorization of the papers was in a way that the 45 essay compositions were divided by three to be used at each phase of the study.

Phase 2: Data collection during training (norming session)

Step 1: The pre-training questionnaire was given to the raters to get their attitudes, feelings, and expectations of the training program. Meanwhile, as it was thought that some raters might not have answered all the items of the pre-training questionnaire in details, they were also interviewed. All the interviews were audio-taped and qualitatively analyzed.

Step 2: In summer 2008, the training program started. The trainer taught the rules for scoring compositions based on the IELTS scoring scale. Moreover, the raters were given five additional new writing papers during the norming session to rate in pairs or groups to increase the effectiveness of the training program. Appropriate hints were provided when the raters gave different scores to an essay. These papers were selected from among the 60 compositions already mentioned in Phase 1. In this phase of the study, the videotaped recordings of the norming session were given to the raters in CDs so that they could watch the CDs at home and fully get the necessary hints they might have lost in the norming session.

Table 2.
Summary of Data Collection and Research Procedures

Phase	Step	Date	Procedure
Phase 1	Step 1	May 24, 2008	Data were collected from students
		May 25, 2008	Data were collected from students
		May 26, 2008	Data were collected from students
	Step 2	May 30, 2008	Composition papers were typed
		June 5, 2008	Composition papers were given to the trainer to rate 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	Step 1	July 10, 2008	The 15 papers were collected and the pre-training questionnaires were given to the raters
		July 12, 2008	The raters were pre-interviewed
		July 13, 2008	The raters were pre-interviewed
	Step 2	July 24, 2008	The first norming session was administered
	Step 3	August 7, 2008	The second norming session was administered
Phase 3	Step 1	August 7, 2008	The post-training questionnaires were given to the raters
		August 9, 2008	The raters were post-interviewed
		August 10, 2008	The raters were post-interviewed
	Step 2	August 11, 2008	15 papers were given to the raters to rate for immediate post-training data collection

Phase 3: Post-training data collection

Step 1: After the training program finished, the post-training questionnaires were immediately given to the raters to get their attitudes, feelings, achievements, and evaluations of the training program. They were also interviewed and the interviews were audio-taped.

Step 2: The researcher gave another 15 essay compositions to the raters to score based on what they have 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 would show the degree of consistency among raters before and after training for both groups. The summary of the data collection and the research procedures appear in Table 2.

Data Analysis

In order to answer both research questions of this study, all the materials including the pre-and post training questionnaires, audio interview tapes before and after training and the video recordings of the norming session were carefully analyzed. Moreover, for the quantitative part of the study, Pearson's product moment correlation coefficient formula for calculating correlation among pairs of raters for both NEW and OLD groups was used. Then, to make the final interrater reliability estimate for both groups of the raters, the average correlation coefficient was adjusted using Spearman-Brown prophecy formula. All the quantitative data analyses in this study were done using SPSS ver. 15.

Results

5.1. The Analysis of the First Question

RQ1: What are the raters' perceptions regarding the rater training program before and after training?

The pre-training questionnaire consisted of 10 questions. The first nine questions required a fixed-choice response. Question 10 asked the raters about their understanding of the purpose of the

Table 3.
Fixed-Choice Responses to Questions on Pre-Training Questionnaire

Questions	Responses	Number
1. How much experience do you have in rating students' compositions?	Very	**4**
	To some extent	**2**
	A little	**4**
	Not at all	**2**
2. I feel comfortable with trying a face-to-face rater training program.	Strongly agree	**3**
	Agree	**9**
	Disagree	**0**
	Strongly disagree	**0**
3. I think I am going to enjoy the rater training experience.	Strongly agree	**7**
	Agree	**5**
	Disagree	**0**
	Strongly disagree	**0**
4. Generally, I support the notion that we need to assess the language proficiency of students' writings.	Strongly agree	**6**
	Agree	**6**
	Disagree	**0**
	Strongly disagree	**0**
5. It's <u>not</u> necessary to have some sort of formal assessment and training process to ensure comparability of standards.	Strongly agree	**0**
	Agree	**1**
	Disagree	**5**
	Strongly disagree	**5**
6. Do you anticipate any problems with the training program?	Yes	**0**
	Maybe	**5**
	No	**7**
7. How effective do you think the rater training program will be?	Very effective	**7**
	Quite effective	**3**
	A little effective	**1**
	Not at all effective	**0**
8. I am flexible and I do accept authorities' comments in rating even if they are against mine.	Strongly agree	**4**
	Agree	**6**
	Disagree	**2**
	Strongly disagree	**0**
9. It is difficult for me to notice my mistakes in rating.	Strongly agree	**0**
	Agree	**2**
	Disagree	**9**
	Strongly disagree	**1**

face-to-face training program in order to ascertain that they had understood what they had been told during the briefing session. Table 3 indicates the raters' responses to all nine fixed-choice items of the pre-training questionnaire.

Six raters had experiences in rating compositions but the other six had little or no experience in this regard.

I rate their compositions intuitively, and sometimes it depends on my mood. If I feel good their scores are better if not vice versa. (Rater NEW4)

The raters felt comfortable being engaged with the training program.

I have always welcomed training as it helped me to learn more and grow. (Rater OLD6)

However, rater NEW5 expressed some apprehension and stated

What if I don't get very close to zero or if I start close and keep on moving away?

All the raters supported the notion that they need to know to rate students' writings.

One of the productive skills that students need to master is writing. In order to help students to activate this and get rid of the common problems, we need to know how to rate their writings. (Rater OLD1)

Most of the raters opposed the belief that there is no need for formal assessment and training process for ensuring the comparability of standards.

It is mandatory to know the standards of marking the writings. (Rater OLD3) However, rater OLD5 believed that there is no need for any formal assessment and training process,

Formal assessment and training process make writing mechanical and not realistic to students and contaminates this important skill, rating must be creative.

One rater had no idea about this (rater NEW6). No problem was anticipated at the start of training. Just some raters were concerned about not being able to catch up with other raters in scoring compositions in the norming session (Raters NEW3,

NEW4, and NEW6). Some others were worried about the limitation of time allotted to raters to score compositions in that time in the norming session (Raters NEW2 and OLD1).

Ten of the raters felt optimistic about the anticipated effects of the training (one did not respond to this question). A summary of their comments indicated that they believed it would be effective because they could absorb as much knowledge in a face-to-face training program as needed. They could also practice what was taught in the norming session as much and as often as needed since they could review the CDs at home for several times. However, one rater was not so much optimistic about the effectiveness of the training program.

Training will be a little effective because rating is quite subjective (Rater OLD 5).

Regarding flexibility in accepting experts' comments, also all raters were positive. Answering this question, rater OLD1 said that

I would disagree, not if they are against mine, but if they are against the standards. And other raters said

I would accept those commands provided that they make sense to me. (Rater OLD4)

Just two raters were against accepting the experts' comments but they did not have any reason for it.

Most of the raters, except two, believed that is quite reasonable for them to notice their mistakes. Rater OLD6 noted that

One has to face realities.

And another rater believed that

Having mistakes in a course in which I am a student is much better than having them in a class where I'm a teacher. (Rater NEW4)

However, two raters, both experienced, believed that after these years of experience in rating, it is difficult for them to notice their mistakes and change their rating method; perhaps, they would prefer to follow their own method in rating (Rater OLD4, and OLD5).

The responses to Question 10 (How much do you think your way of rating has changed as a result of face-to-face training program?) indicated that all raters appreciated the main purpose of the training program in enhancing the accuracy of ratings and giving them hints and techniques in rating as systematically and precisely as possible.

Table 4.
Fixed-Choice Responses to Questions on Post-Training Questionnaire

Questions	Responses	Number
1. Altogether how effective did you find the face-to-face training program?	Very	**8**
	To some extent	**2**
	A little	**2**
	Not at all	**0**
2. Overall, how friendly (trainer-trainee) did you find the face-to-face training program?	Very	**9**
	To some extent	**3**
	A little	**0**
	Not at all	**0**
3. How much did you enjoy your face-to-face training experience?	Very	**10**
	To some extent	**2**
	A little	**0**
	Not at all	**0**
4. How much do you think the face-to-face training program achieved its purpose?	Very	**8**
	To some extent	**2**
	A little	**2**
	Not at all	**0**
5. How much material descriptors, (e.g. scripts, notes, etc.) was covered in the program?	Too much	**0**
	Just right	**9**
	A little	**3**
6. How much do you think your way of rating has changed as a result of face-to-face training program?	Very	**7**
	To some extent	**3**
	A little	**2**
	Not at all	**0**

To answer the first question, two raters said that

The post-training questionnaire consisted of 11 questions. The first six questions required a fixed-choice response. Questions 7 to 11 were open-ended questions and asked the raters to remark the advantages and disadvantages of the training program, what they liked the most or least and their suggestions for future workshops. Table 4 indicates the responses for all the 6 fixed-choice items.

It was good to get instant feedback in the norming session after rating each paper and receiving appropriate hints in rating. (Rater NEW3)

I found the trainers comments very useful compared to my own. (Rater NEW5) However, one rater questioned the effectiveness of this program.

I don't think my rating has improved enough as a result of this training because rating is not teachable and it's quite subjective. (Rater OLD4)

The second and third questions asked the raters if they enjoyed the face-to-face training program and whether it was friendly or not. All the raters were positive and they expressed that they all enjoyed a lot and had close and friendly relationship with the trainer.

The training program was exactly as I expected, the trainer was kind and friendly and I didn't feel any trainer-trainee distance. (Rater OLD6)

The answer to the forth question was similar to the first question . All the raters except two believed that the training program achieved its goals.

I myself have changed quite a lot and my perspective to rating is now different.(Rater NEW1)

However, two raters, for the same reason mentioned in the first question 1, stated that the training did not achieve its goals.

I think still I'm going to use my own way in rating students compositions. The scripts which were rated by the trainer couldn't exemplify the huge flood of different compositions we face during rating. There are many cases which do not match those examples. (Rater OLD3)

The fifth question asked the raters to state their opinions about the materials covered in the training program. The majority believed that the materials were just right.

It was very much beneficial to give the norming session CDs to the raters so that they can review the hints and techniques taught by the trainer for several times. (Rater OLD5)

However, two raters expressed their disagreement in this respect.

A complete pamphlet including all types of mistakes and errors students make in writing is needed so that rating would be easier and more precise. (Rater OLD2).

The sixth question six asked the raters to indicate if their way of scoring has changed or not. The majority of raters were positive and they stated that their rating has greatly improved.

I think after this training program I'm a new rater and I see rating completely different. (Rater NEW1)

Questions 7 to 11 were open-ended questions which asked the raters to comment on the benefits and drawbacks of training, what they liked the least and the most and what they thought could be improved regarding the training sessions.

Regarding the advantages of the training program, rater NEW5 mentioned that *having interaction with other raters and hearing what others say as well as being sociable is remarkable.*(Rater OLD4)

However, regarding the disadvantages rater NEW 2 believed that

Being obliged to do the ratings during the norming session in a limited time gave me much stress". "I was worried a lot because I thought my ratings might be much more different from the benchmark than those of others.(Rater NEW3).

The last question asked the raters to give their suggestions for further improvements of the training program.

I think increasing the number of sessions would increase the efficacy of this program. (Rater OLD4)

Including some videotapes of the ratings of real IELTS raters into the program would help raters to a high extent. (Rater NEW6)

5.2. The Analysis of the Second Question

RQ2: Do experienced and inexperienced raters differ in their scoring of compositions before and after training?

In order to understand whether the training program was more beneficial for NEW raters or OLD raters, the raters' consistency for both groups should be measured before and after training.

Raters' consistency before training

The average correlation coefficients for NEW raters, measured before and after the training program, were **0.1** and **0.4** respectively. This shows that the rating reliability of NEW raters prior to the training program was low and NEW raters were just a little consistent among each other before training. The average correlation coefficients for OLD raters, before and after the training program, were also measured to be **0.3** and **0.72** respectively. This shows that the rating reliability of OLD raters prior to the training program was higher than NEW raters; however, it was not a very high reliability.

The standard deviation was also measured to find out to what extent NEW and OLD raters have dispersion prior to training. Therefore, the mean standard deviation at the pre-training phase measured to be **1.98** for the NEW raters and **2.31** for OLD raters. This suggests that the mean dispersion among raters in giving consistent scores to compositions prior to the training program was **1.98** for NEW raters and **2.31** for OLD raters. This difference in giving consistent scores is quite high.

Raters' consistency after training

For the second step, the post-training data were analyzed to get the interrater consistency following training. The average correlation for NEW raters following the training program measured to be 0.78 and the final interrater reliability among NEW raters following training measured to be 0.95. This shows that the rating reliability of NEW raters after training was high and NEW raters became highly consistent among each other after training.

The average correlation for OLD raters after training measured to be 0.83, and the final interrater liability among them after training was 0.80. This shows that the rating reliability among OLD had a very little improvement after training. Through comparing the interrater reliabilities among NEW and OLD raters before and after training (see Table 5), it is clear that training was much more effective for NEW raters than OLD ones.

Table 5.
Interrater Reliability among New and Old Raters Before and After Training

Raters	Pre-training	Post-training
NEW raters	0.40	0.95
OLD raters	0.72	0.80

Although interrater reliability for NEW raters was low before training, after training, they became highly consistent. For OLD raters, despite being experienced and having a better reliability before training, the reliability of their scoring did not improve much surprisingly after training. The reason for this could be that OLD raters are not very much flexible in accepting experts' comments because of the high self-confidence or even arrogance they may have. However, NEW raters are very much flexible and willing to learn from experts and that is why after training, their interrater reliability was developed considerably.

The standard deviation was also measured for NEW and OLD raters to get their extent of dispersion after training. The mean standard deviation for NEW raters in the post training phase measured to be **0.86** for NEW raters and **0.92** for OLD raters. Table 6 summarizes NEW and OLD raters' dispersion before and after training.

Table 6.
NEW and OLD Raters' Dispersion Before and After Training

Raters	Pre-training	Post-training
NEW raters	1.98	0.86
OLD raters	2.31	0.92

A brief look at Table 6 reveals that training was effective for both groups in reducing dispersion among scores given by the raters; however, the result of reliability analysis shows that training was more effective for NEW raters than OLD raters.

Moreover, the sum of point differences for pre and post training were calculated. Table 7 shows the sum of point differences in the two phases of the study for NEW and OLD raters. The sum of point differences indicates the sum of average scoring differences for each group of raters before and after the norming session.

Table 7.
Sum of Point Differences

Raters	Pre-training	Post-training
NEW raters	13.12	3.92
OLD raters	11.89	5.42

The results show that although OLD raters had less point differences than NEW raters prior to the training program, they tended to have higher point differences after training. This again shows that training was more effective for NEW raters than OLD raters.

Moreover, in order to make sure whether the mean differences between NEW and OLD groups is significant at the post training stage or not, a paired t-test was run. The t-value obtained was t (NEW-OLD) = **3.75**, which was significant at **0.01**, df=**10**. The result shows that the difference between the means of NEW and OLD raters is significant and it is not due to random error. Moreover, it provides support for the success of the treatment.

Discussion and Conclusion

The major findings from the comparison across rater types were as follows. NEW raters were less consistent than OLD raters before training, as expected. More surprisingly, this was not the

case after training; by contrast, they became even more consistent than OLD raters after training. This finding is in-line with that of Weigle (1994b), and Cumming (1990), where they found NEW raters were more greatly affected by training; however, Knoch, Read, and Randow (2007) found the reverse. They found that OLD raters became more consistent after training.

In terms of the raters' attitude to the training program, those raters whose training behavior got a little better after training than those whose ratings were greatly developed tended to be somewhat less positive in their attitude toward the training program. For example, rater OLD5 did not have a high positive attitude about training and its effectiveness, and thus, his improvement after training was just a little. Although causal connections between attitudes and outcomes cannot be assumed, it is said that if any training is done in a friendlier atmosphere, it would be more effective (Hamilton et al., 2001). On the other hand, those raters who accepted experts' comments tended to move more closely toward the benchmark (as suggested by Reed & Cohen, 2001). Most raters had very positive attitudes toward 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. The raters also appeared to have good awareness of their own rating, especially to certain categories (similar to findings of Wigglesworth, 1993).

Implications of the Study

One very important implication of this study is that the findings suggest that all raters are capable of rating reliably regardless of their background and training. So decision makers, in selecting raters, should not be concerned about the raters' background because the variable of experience does not increase reliability and therefore, they should put their emphasis on training sessions for raters.

One related implication of this study is that since the findings reiterated better ratings by NEW raters, decision-makers should select economical groups of raters for evaluating essays (NEW

raters because they ask for low amount of money to do the job). In many cases, decision-makers tend to select only professional (OLD) raters; however, the findings provided evidence for better rating achievements by NEW compared to OLD raters.

Acknowledgements

I am indebted to Dr. Fahimeh Marefat, Dr. Cushing Weigle and Dr. Elder for their helpful comments on this project. I 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.

The Authors

Houman Bijani is a Ph.D. candidate in TEFL from Islamic Azad University (Science and Research Branch). He got his M.A. in TEFL from AllamehTabataba'i University in 2009 as a top student. He was awarded the TKT (Teaching Knowledge Test) certificate from Cambridge University in 2009. He is a faculty member at Islamic Azad University, Zanjan Branch, Iran. He has published several articles in national and international journals. He has also presented some papers in language related conferences. His areas of interest include testing, writing assessment and teacher education.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Breland, H. M., & Jones, R. J. (1984).Perceptions of writing skills.*Written Communication, 1*(1), 101-119.

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

Cumming, A. (1990). Expertise in evaluating second language compositions.*Language Testing, 7*, 31-51.

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.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of online rater training and monitoring .*System, 29*, 505-520.

Huot, B. (1990). In reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*, 201-213.

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.

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.

Ruth, L. & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex Publishing Corp.

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.

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.

بررسی تاثیر میزان درک وتجربه مصححان برسنجش مهارت نگارش در زبان دوم

هومن بیژنی
دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران

آموزش مصححان در اعتبار سازه ی آزمون های نگارش بسیار حائز اهمیت است. دلیل این اهمیت این است که از طریق آموزش، مصححان به استفاده از توانایی دانش آموزان بجای معیار های شخصی در ارزیابی مهارت نگارش تغییر رویه می دهند. با وجود آنکه مباحث زیادی در رابطه با سنجش مهارت نگارش مطرح شده، اما تحقیقات بسیار کمی به بررسی نحوه ی نگرش مصححان و شیوه ی برخورد ایشان با برنامه ی آموزشی پرداخته اند. همچنین، مطالعات بسیار کمی تفاوتهای بین مصححان آموزش دیده و مصححان آموزش ندیده در ارزیابی مهارت نگارش را مورد بررسی قرار داده اند. علاوه بر آن، مطالعات کمی در حوزه ی سنجش مهارت نگارش، از متد قبل و بعد آموزش استفاده کرده اند. در این تحقیق به بررسی میزان تاثیر و نحوه ی عملکرد برنامه ی آموزشی بر روی مصححان با تجربه و مصححان بی تجربه پرداخته شد. 12 مصحح شرکت کننده در این مطالعه تعداد 45 برگه نگارش را در سه مرحله ی قبل، بعد و در خلال برنامه ی آموزشی تصحیح کردند. همچنین، این برگه ها توسط مربی دوره ی آموزشی به منظور استفاده به عنوان معیار سنجش نیز تصحیح شدند. یافته های این تحقیق نمایانگر وجود ناهماهنگی گسترده تری در میان مصححان بی تجربه نسبت به مصححان با تجربه در مرحله ی قبل از اجرای برنامه ی آموزشی بود. اما پس از اجرای برنامه ی آموزشی، نتایج از وجود هماهنگی بسیار بالاتری در میان مصححان بی تجربه در مقایسه با مصححان با تجربه در این مرحله از برنامه ی آموزشی حکایت دارند.

کلید واژه ها: ارزشیابی، پایایی، پایایی ارزیابها، اعتبار