Iranian Journal of Applied Linguistics (*IJAL*)

Vol. 20, No. 1, March 2017, 113-150

Investigating the Effect of the Training Program on Raters' Oral Performance Assessment: A Mixed-Methods Study on Raters' Think-Aloud Verbal Protocols

Houman Bijani, *Islamic Azad University, Science and Research Branch, Tehran, Iran*
Mona Khabiri*, *Islamic Azad University, Central Branch, Tehran, Iran*

Abstract

Although the use of verbal protocols is growing in oral assessment, research on the use of raters' verbal protocols is rather rare. Moreover, those few studies did not use a mixed-methods design. Therefore, this study investigated the possible impacts of rater training on novice and experienced raters' application of a specified set of standards in rating. To meet this objective, the study made use of verbal protocols produced by 20 raters who scored 300 test takers' oral performances and analyzed the data both qualitatively and quantitatively. The outcomes demonstrated that through applying the training program, the raters were able to concentrate more on linguistic, discourse, and phonological features; therefore, the extent of their agreement increased specifically among the inexperienced raters. The analysis of verbal protocols also revealed that training how to apply a well-defined rating scale can foster its use for raters both validly and reliably. Various groups of raters approach the task of rating in different ways, which cannot be explored through pure statistical analysis. Thus, think-aloud verbal protocols can shed light on the vague sides of the issue and add to the validity of oral language assessment. Moreover, since the results of this study showed that inexperienced raters can produce protocols of higher quality and quantity in the use of macro and micro strategies to evaluate test takers' performances, there is no evidence based on which decision makers should exclude inexperienced raters solely because of their lack of adequate experience.

Keywords: Bias; Oral performance assessment; Rater training; Think-aloud verbal protocols

Article Information:

Received: **12 November 2016** **Revised: 15 February 2017** **Accepted: 21 February 2017**

Corresponding author: Department of English Language Teaching, Central Tehran Branch, Islamic Azad University, Tehran, Iran
Email address: Mona.khabiri@iauctb.ac.ir

[DOI: 10.29252/ijal.20.1.113]

1. Introduction

It is well documented that there is a need to test oral ability in language syllabuses. An important characteristic of oral ability assessment is that test-takers are needed to produce language verbally, and that their real performance is assessed on the basis of predetermined rating criteria (Green, 1998). Test-takers' performances, derived from performance-based tasks, are scored by raters and their language ability is inferred from their test scores. However, it is well understood that raters do not always reach consensus upon oral performances scores. One very important reason for this lack of consensus, no matter how carefully the test is constructed, is the behavior of the rater or the interviewer which can directly influence the outcome of performance assessment. Some previous research on rater behavior has demonstrated a considerable amount of rater variability which is mostly related to raters' characteristics and not the test takers' performance (e.g., Carey, Mannell, & Dunn, 2011; Knoch, 2011).

Rater training is commonly used as a means for compensating variability due to factors such as raters' backgrounds and thus adjusting raters' expectations. Training, along with the use of a scoring rubric is said to clarify the expected criteria and to have raters judge performance based on those expected criteria rather than their own; to reduce differences regarding different backgrounds of raters; to allow raters to focus on the suitable criteria; and to modify expectations of good speaking by clarifying for the raters the requirements of the tasks and the characteristics of the speakers (Knoch, 2009). Research on how raters use descriptors is typically done through instruments like questionnaires, interviews, or think-aloud protocols (Barkaoui, 2011; Knoch, 2009; Sawaki, 2007). In think-aloud protocols, raters verbalize their thinking process while doing the rating. Through analyzing the verbalized data, researchers find out how raters interpret the descriptors of the rating scale, thus coming to a particular given score.

[DOI: 10.29252/ijal.20.1.113]

2. Review of the Related Literature

2.1. Verbal protocols in performance assessment

The analysis of verbal protocols has a long history in psychological research, but it was only with the work of Ericsson and Simon (1993) that the theory of verbal reports and methodology for collecting and analyzing protocol data became systematized. A majority of the work in protocol analysis deals with problem solving, mathematics, or decision making. However, it has been during the last two decades that protocol analysis has also been applied to the study of language related academic tasks such as composition writing (Trace, Janssen & Meier, 2017), test taking (Nakatsuhara, 2011), and oral speaking assessment (Kuiken & Vedder, 2014). Kuiken and Vedder (2014) also advocate the use of verbal protocols as a source of evidence in construct validation of tests. Wolfe (2004) suggested the use of think-aloud protocols in selecting, training, and observing raters. He suggested monitoring raters as they think aloud so as to identify the problematic aspects of scoring. Verbal protocols have been used in studies of oral performance ratings (e.g., Attali, 2016; Barkaoui, 2011; Cumming, Kantor, & Powers, 2002; Kim, 2011; 2015; Sasaki, 2014; Weigle, 1999; Wolfe, 2004). According to Davis (2016), the use of think-aloud protocols can unfold raters' thoughts in order to identify why and how a rater chooses a certain score. The advantage of think-aloud protocols over questionnaires and interviews is that think-aloud protocols are immediate. Besides, unlike questionnaires and interviews, think-aloud protocols reflect what raters actually do when rating rather than what they just believe in as in questionnaires and interviews.

2.2. Merits and demerits of the use of verbal protocols

Although verbal protocols can reveal the picture of what happens in raters' minds which will enable researchers to specify what happens during rating regarding the raters' interpretations in the use of rating scale categories, the use of verbal protocols has its own limitations. Firstly, they are difficult to administer because participants typically are not used to verbalizing their

[DOI: 10.29252/ijal.20.1.113]

thoughts when concentrating all their attention on rating a performance. Secondly, the process of collecting the protocol data and transcribing and coding them later on is time-consuming and hard to administer (Ling, Mollaun, & Xi, 2014). The biggest criticism was mentioned by Bowles (2010) who expressed his concern on the use of verbal protocols in two issues, *Veridicality* and *Reactivity*. According to Bowles, veridicality is concerned with whether think-aloud protocols truly report the raters' real thinking and rating process, whereas reactivity concerns whether the procedure to produce verbal protocols can affect the outcome of scoring. He also claimed that think-aloud protocols are incomplete because during their production, long-term memory is inaccessible for verbalization. He added that although participants have got access only to their short-term memory, this does not reduce the value of the collected protocol data.

In spite of all these criticisms with respect to the subjectivity, inaccuracy, and inconclusiveness of the nature of verbal protocols in the provision of data, Smagorinsky (2001) strongly recommends that protocol analyses can provide valuable information if they are collected and analyzed systematically. Several studies have investigated the application of think-aloud protocols in rating. Kim (2011) analyzed the protocols of nine experienced raters scoring six oral performances on a holistic scale. He recommended that while raters can agree on many performances based on the guidelines for holistic assessment, they may disagree on their own rating style for performances which do not clearly fit the descriptors of the holistic scale. Moreover, Kim (2015) studied the think-aloud protocols of eight raters scoring 42 students' oral performances on a holistic scale. Four of the raters were trained, experienced raters, and four had no training or experience. Although he found no differences in the rating criteria for the two groups of raters, he found that the more experienced raters benefitted from the more efficient strategies and a more extensive range of responses to the performances than did the inexperienced raters.

Another comparison between expert and novice raters using think-aloud protocol analysis is the study conducted by Attali (2016). Attali analyzed the protocols of eight expert and novice raters rating 12 ESL students' essays which differed on the basis of the dimensions of language

[DOI: 10.29252/ijal.20.1.113]

proficiency (intermediate vs. advanced) and writing expertise (professional vs. average). Essays were evaluated on language use, content, and organization. Attali, then, found that both groups of raters were able to distinguish between writing ability and language proficiency in their evaluations of the essays, although novice raters were significantly more lenient in their judgments of content and organization than were expert raters. While both groups of raters made approximately the same number of decision-making behaviors in their evaluations, the types of behaviors differed between the two groups. For example, expert raters reported more self-reflexive behaviors such as reflecting on how they were distinguishing between the rating categories, whereas a majority of the novice raters did substantially more editing of errors while evaluating compositions.

In another study Davis (2016), in assessing oral performance, found that experienced raters made more comments after listening to the oral performance than did inexperienced ones who made more comments while listening to the oral performance. Weigle (1999), in a study of raters' verbal protocols in writing assessment, found that some raters announced difficulty expressing their thoughts out-loud and that some raters provided much more protocols than the others. Barkaoui (2011) used the Multifaceted Rasch Measurement (MFRM) and verbal protocols in a writing test and could identify misfitting raters. In his study of 25 raters and 150 samples, he found significant differences among raters. He further found that the use of think aloud influenced raters' severity level, but did not provide much information about the rating process. He also found that grammar was the most severely scored category. It should be noted that Barkaoui benefited only from articulate participants in his study; therefore, this can limit the generalizability of the study outcomes. Table 1 lists some of the published studies using think-aloud protocols for performance assessment.

[DOI: 10.29252/ijal.20.1.113]

Table 1
Summary of think-aloud studies on oral and written test performance

Study	**Research type and purpose**	**Raters**	**Rating scale**
Attali (2016)	Descriptive; what strategies the raters use	8 experienced and inexperienced raters	Analytic
Barkaoui (2011)	Comparative; experienced and inexperienced raters	14 experienced 11 inexperienced	Analytic
Cumming, Kantor and Powers (2002)	Descriptive and comparative	4 raters with different backgrounds	Holistic (6 levels)
Davis (2016)	Descriptive	4 experienced raters	Analytic
Erdosy (2004)	Descriptive and comparative; effect of raters background on their ratings	4 raters with different backgrounds	Holistic (6 levels)
Kim (2011)	Descriptive; how a scale affects their scoring	9 experienced raters	Holistic
Kim (2015)	Comparative; how experienced and inexperienced raters treat the test	4 experienced and 4 inexperienced raters	Holistic
Sasaki (2014)	Descriptive; how raters deal with tasks	3 experienced raters	Analytic (5 categories and 4 levels)
Weigle (1999)	Comparative (rating before and after training)	4 experienced raters	Analytic (3 categories and 10 levels) Holistic (6 levels)
Wolfe (2004)	Comparative; raters of different proficiency levels	12 raters	Holistic (6 levels)

Although the use of verbal protocols is fairly extensive in oral assessment research through the provision of rich data, researchers who use

[DOI: 10.29252/ijal.20.1.113]

it must be aware of concerns attributed to it. Ericsson and Simon (1993) argue that not mentioning a particular feature or features by a rater does not indicate that those features do not exist. That is, raters frequently have thoughts passed through their minds that they were not able to articulate. Cohen (1994) cautions that people forget salient aspects as soon as the thoughts have passed their minds. He further appeals for training program before verbal report production.

It has been impossible to explain why raters demonstrate differences in rating behavior. Besides, research on the use of raters' verbal protocols, although highly essential, is rather rare. Even those very few ones on the application of verbal protocols (e.g., Barkaoui, 2011; Kim, 2011, 2015; Sasaki, 2014; Weigle, 1999; Wolfe, 2004) did not use both qualitative and quantitative analysis models together. The reason for such importance is that it is only through verbalization during the rating process that researchers understand how raters make judgments about the quality of oral discourse, or whether raters display interference when providing protocols. Moreover, there is little research investigating how experienced and inexperienced raters approach the rating task for oral language assessment. The significance is that such finding will clarify the differences between the two rater groups and will help raters tackle with even the smallest obstacles which cause inconsistency in rating among them. Moreover, there is evidence that raters deal with direct and semi-direct oral assessment tests differently; however, there is little justification explaining why such differences occur although the tasks are commonly the same. This is something that only the analysis of raters' collected verbal protocols will shed light on. Finally, there is still paucity of research investigating the extent to which a rater training program can contribute to raters' consistency and reduce the measures of bias through the analysis of their verbal thoughts and how long the effectiveness of the training program will last (i.e., whether there is any reduction of the effectiveness of the training program on raters when rating test takers' oral performances as reflected in their verbal protocols). Therefore, this study investigated the validity of the current procedures for assessing EFL speaking ability in a specific setting, particularly with regard to the training of raters to apply a specified set of

[DOI: 10.29252/ijal.20.1.113]

standards in ratings. This is to find what features the experienced and inexperienced raters mostly focus on when scoring test takers' oral performances and to what extent the training program can bring about systematicity in this regard. This study investigated the use of raters' collected verbal protocols and analyzed the obtained data both qualitatively and quantitatively to have a deeper and more precise understanding of raters' decision making behaviors. Based on the above-mentioned issues, the following research questions can be formed:

RQ1: How do raters' verbal protocols affect the scoring procedure? And what do protocols reveal about the raters' scoring patterns when assessing oral performance?

RQ2: Is there any significant difference between experienced and inexperienced raters' verbal protocols before and after the training program?

3. Method

3.1. Participants

Three hundred Iranian adult students of English as a Foreign Language (EFL), including 150 males and 150 females, ranging in age from 17 to 44 participated in the study as test takers. The students were selected from intermediate, upper-intermediate, and advanced levels studying at the Iran Language Institute (ILI).

Twenty Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 participated in this study as raters. These raters all graduated in English language- related fields of study. In order to search for rater participants for the present study, a background questionnaire, adapted from McNamara and Lumley (1997), eliciting the following information including (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training* and (5) *relevant courses passed* was given to the raters. Based on the above-mentioned method of rater

[DOI: 10.29252/ijal.20.1.113]

classification, raters were divided into two levels of experienced and inexperienced raters as outlined below.

A. Raters who had no or less than two years of experience in rating and had not received rater training, and had no or less than five years of experience in teaching and passed less than four core courses related to ELT major (i.e., pedagogical English grammar, phonetics and phonology, second language acquisition and second language assessment). Hereinafter we call these raters as NEW.

B. Experienced raters who had over two years of experience in rating and had received rater training, and had over five years of experience in teaching and passed all four core courses plus at least two selective courses related to ELT major. Hereinafter we call these raters as OLD.

3.2. Instruments

3.2.1. The scoring rubric (analytic)

The purpose of using an analytic rating scale was to assess test takers' oral performance in order to determine to what extent the evaluation of test takers' oral proficiency is done in a more valid and reliable way and to identify how well the raters use the rating scale categories, based on the given descriptors, systematically and without bias. Test takers' task performance was assessed using the ETS (2001) analytic rating scale using criteria including *fluency*, *grammar*, *vocabulary*, *intelligibility*, *cohesion* and *comprehension*.

3.2.2. Oral tasks

The elicitation of test takers' oral proficiency was done through the use of five different tasks including description, narration, summarizing, role-play and exposition tasks. Task 1 (*Description Task*) is an independent-skill task which elicits test takers' personal experience or background knowledge to respond in a way that no input is provided for it (McNamara, 1996). On the other hand, tasks 3 (*Summarizing Task*) and 4 (*Role-play Task*) elicits test

[DOI: 10.29252/ijal.20.1.113]

takers' use of their listening skills to respond orally. In other words, the content for the response was provided for the test takers through listening – short or long. For tasks 2 (*Narration Task*) and 5 (*Exposition Task*), the test takers are required to respond to pictorial prompts including sequences of pictures, graphs, figures, and tables. The tasks were obtained from Luoma (2004) and all test takers were required to take all the tasks.

3.3. Procedure

3.3.1. Pre-training phase

Prior to collecting any data from the test takers, the background questionnaire was given to the raters to fill out. The aim of having the raters fill out the raters' background questionnaire sheets was to enable the researcher to classify them into the two groups of rating expertise. Then, they were randomly divided into three groups each containing 100 individuals. Since the study was done in three phases (pre-training, immediate post-training, and delayed post-training), each group of test takers participated in one phase. The reason for conducting the study in three phases was to evaluate the ongoing effectiveness of the training program in short and long terms. Although the raters participating in this phase of the study had not been instructed how to provide think-aloud protocols yet, they were asked to tape-record their verbal reports of thoughts while scoring the oral performances for further analysis. The purpose was to make comparisons among the raters' of the three research phases.

3.3.2. Rater training procedure

The steps in the operational training program were taken precisely to ensure that grading was done fairly and consistently. The training program was done in two sessions, each lasting for about six hours. The four components of rater training (*rater norming, training for verbal reports, rating with verbal protocol reports, and feedback on previous rating*) are discussed below.

[DOI: 10.29252/ijal.20.1.113]

3.3.2.1. Rater norming

All the raters participated in a training (norming) session in which the speaking tasks and the rating scale were introduced and they were given some time to practice the instructed material with some sample responses. Moreover, the raters discussed the differences in their scores and reviewed their decision making processes with the instructor. A norming packet was used in the norming session including the tasks, representative samples of oral performances from previous ratings representing various scoring bands to better provide raters with awareness of the scoring principles, and the analytic scoring rubric. The training was done by an authorized ETS rater trainer.

3.3.2.2. Training for verbal reports

The raters (NEW and OLD) were also instructed how to verbally report their thoughts while they listened to a speaking response and made a score decision. To better enhance the impact of training raters for verbal report production, they were provided with video-recordings of previously-performed verbal protocols conducted by the researcher. Meanwhile the raters were asked to provide their own reasons, logic and comments on anything significant they saw on the basis of the observed rating videos.

The raters (NEW and OLD) were reminded to (1) rate the tasks in the way they would if they were not supposed to think aloud; (2) verbalize all their thoughts during rating; (3) be thoroughly natural, and without bias in rating. They were also told to feel free in rating and producing as many protocols as they wished. The raters were videotaped all throughout the study to make sure that all the requirements were met. The trainer frequently asked the raters to verbalize their thoughts since according to Wagner (2006), those who are not familiar with verbal reports are likely to forget to constantly talk aloud. Although group rater training session was the main part of the rater training program, it was, however, accompanied with rating norming practice, group discussion, and score negotiation. These procedures were continued until they reached consensus and all raters

[DOI: 10.29252/ijal.20.1.113]

were confident with determining test takers' scores across the descriptors of the scoring rubrics.

3.3.2.3. Rating with verbal protocol reports

In this study, unlike the previous ones, the raters were required to perform verbal reports, which is abnormal in most actual rating sessions. Although according to Weigle (1999), a request for verbal reports may affect the raters' scoring, such elicitation method is necessary to observe the raters' decision-making process. The researcher transcribed the verbally-recorded reports based on Shohamy's (1994) discourse features framework to analyze the produced verbal protocols based on lexical density, rhetorical functions and structures, genre, speech moves, communicative properties, discourse strategies, content and topic of discourse, prosodic/paralinguistic features and contextualizations, type of speech functions, discourse markers, and register for qualitative data analysis to achieve further certainty of raters' change of behavior in various rating groups among pre-, post- and delayed post-training stages. The think-aloud protocols were rather extensive, ranging in length from 8 to 21 typed pages per rater in the whole study.

3.3.2.4. Feedback on previous ratings

In addition to the training sessions, feedback on previous ratings was provided to each rater individually in the second norming session. As Wallace (1991) argues, repeated practices do not guarantee the development of professional competence. Thus, for him, prior rating performance would give raters an opportunity to reflect on their rating behavior. Since each rater had a different rating ability and exhibited various rating behavior, feedback was provided to each rater individually. The feedback also included each rater's use of rating scales examined through the qualitative analysis of each rater's verbal reports. The following qualitative analyses were thus included:

[DOI: 10.29252/ijal.20.1.113]

1. Whether the raters were able to distinguish different components/criteria given in the rating scale accurately.
2. Whether they gave explicit attention to all descriptors in the rating scale.
3. Whether they could match the features in the responses to appropriate descriptors while assigning scores.

3.3.3. Immediate post-training phase

The data were collected from the second group of test takers (including 100 test takers) through having them perform the oral tasks. Meanwhile, the raters (NEW and OLD) were asked to precisely follow the instructed techniques and principles of how to report their thoughts verbally and to tape-record them for further analysis. It is reiterated that the purpose for this step was to make comparisons among the scoring behaviors of different rater groups for the three research phases of this study.

3.3.4. Delayed post-training phase

Exactly two months (as suggested by McNamara, 1996) after the immediate post-training data collection, the last third of the test takers (including 100 test takers) were used from whom to elicit data. The collected data were given to both raters (NEW and OLD) to rate. Also, the raters were again asked to record their think-aloud protocols on the basis of the techniques, strategies, and principles they had already been instructed to observe and obtain evidence on their change of behavior or probable forgetfulness of the rating strategies and techniques they were instructed during the norming session throughout the lapse of time. They were also reminded to tape-record the protocols accordingly for further analysis. It is noteworthy to indicate that they were repeatedly observed and video-recorded in order to make sure they would follow the research requirements. The purpose of this phase of the study was to achieve further certainty about rating behavior changes among the raters in different groups compared to the previous two phases. Once again, the videotaped recordings of the interview performance

[DOI: 10.29252/ijal.20.1.113]

sessions were given to the raters so that with the help of which they would be able to better identify the extralinguistic clues to enhance their rating behaviors.

3.4. Data analysis

The collected data were analyzed using both a qualitative and quantitative research design. The qualitative analysis was done through systematic coding of the collected verbal protocol data based on the raters' viewpoints and the features they concentrated on, and the quantitative data were analyzed by measuring descriptive statistics of the collected data for each feature. An independent samples *t*-test was also run to measure any significant difference between NEW and OLD raters' produced data protocols to determine the hypothetical advantage of one group over the other.

4. Results

In order to demonstrate the raters' views about test takers' performance behaviors, raters' collected verbal protocols were analyzed. A review of the verbal protocols indicated that although raters focused on linguistic features, discourse features were also of great concern. It was evident from the protocol transcripts that some raters tended to produce more commentary than others. However, in general, through analyzing the verbal protocols, it was observed that OLD raters were likely to verbalize their thoughts more extensively in a way that they produced longer protocols with more details compared to NEW ones. This finding is in line with that of Davis (2016) and Kim (2015) who also found that OLD raters produced more comments and elaborated more on their judgments than NEW ones. On average, each rater produced between 7812 and 15577 words throughout the entire study.

Also the observations of the raters' verbal protocols revealed that NEW raters, in general, produced their verbal protocols mainly when they were finished with each test taker's assessment. That is, a majority of NEW raters tended to produce their verbal thoughts when a test taker was done

[DOI: 10.29252/ijal.20.1.113]

with his/her speech production on the oral tasks. In contrast, most OLD raters produced their verbal protocols while they were listening to test takers' speeches. They would halt in several intervals and produce their verbal thoughts while the interaction was in progress. In order to analyze the raters' produced-verbal protocols more systematically and to better document the data on the basis of a well-defined structural format, all collected verbal data were analyzed and classified according to a checklist demonstrating raters' decision making behaviors in all the three phases of the study. The checklist reports the mean frequencies and standard deviations for NEW and OLD raters as well as the results of the independent samples *t*-test to assess the differences between the two groups of expertise with respect to the quantity of their produced speeches.

The analysis of verbal protocols at the pre-training data collection phase, presented in Table 2, shows that OLD raters adopted more metacognitive strategies such as asking themselves what could be added or edited to better develop the content and compensate for the missing data provided by the test takers. Furthermore, data analysis at this stage revealed that OLD raters devoted more attention than NEW ones to *considering the situation of the examinee* ($\overline{X}$ = 12.55, sd. = 2.71 vs. $\overline{X}$ = 6.51, sd. = 1.29); *making more comparisons among various performances of different examinees* ($\overline{X}$ = 16.80, sd. = 3.64 vs. $\overline{X}$ = 8.37, sd. = 1.73); *summarizing their own judgments to finalize the outcome of their assessments* ($\overline{X}$ = 6.60, sd. = 1.43 vs. $\overline{X}$ = 2.40, sd. = 0.52); *identifying vague parts* ($\overline{X}$ = 6.63, sd. = 1.39 vs. $\overline{X}$ = 3.00, sd. = 0.65); *employing logic and reasoning* ($\overline{X}$ = 64.20, sd. = 5.34 vs. $\overline{X}$ = 29.40, sd. = 6.37); *having novelty, originality and creativity* ($\overline{X}$ = 13.20, sd. = 2.86 vs. $\overline{X}$ = 3.60, sd. = 0.78); *identifying information redundancies* ($\overline{X}$ = 12.13, sd. = 2.67 vs. $\overline{X}$ = 1.80, sd. = 0.39); *assessing the comprehensibility of spoken discourse* ($\overline{X}$ = 17.40, sd. = 3.77 vs. $\overline{X}$ = 12.54, sd. = 2.66); *focusing on pronunciation and accent* ($\overline{X}$ = 8.40 , sd. = 1.82 vs. $\overline{X}$ = 5.40, sd. = 1.17); as well as *concentrating on fluency* ($\overline{X}$ = 4.77 , sd. = 1.12 vs. $\overline{X}$ = 2.38, sd. = 0.44). This last finding (pronunciation and accent) is rather in contradiction with that of Sasaki (2014) who found that experienced raters were as careful as inexperienced raters about accent

[DOI: 10.29252/ijal.20.1.113]

and pronunciation. A deeper analysis of verbal protocols comments revealed that NEW raters, instead of concentrating on phonological features, focused their attention on the overall quality of their accent and pronunciation.

- He can provide his opinions well and has given good support. I liked his body language, too. His eye contact and gestures were very helpful to follow the conversation. *(Rater OLD3-Pre-training)*
- He can express his opinions very clearly. He also added some nonverbal language to his speaking. He also used humor at times. *(Rater OLD7-Pre-training)*
- She is incapable of comprehending the message and responding quickly. It takes a long time for her to answer. *(Rater OLD1-Pre-training)*

Nevertheless, NEW raters focused their attention more on *revising their ratings* ($\overline{X}$ = 62.40, sd. = 6.52 vs. $\overline{X}$ = 43.20, sd. = 6.36); *evaluating the quantity of spoken data* ($\overline{X}$ = 44.40, sd. = 4.28 vs. $\overline{X}$ = 31.80, sd. = 6.89); and *identifying the frequency of errors committed by the examinees* ($\overline{X}$ = 1.80, sd. = 2.34 vs. $\overline{X}$ = 6.60, sd. = 1.43). They also tended to concentrate more *on evaluating grammatical and syntactic structures in the examinees' performances* ($\overline{X}$ = 30.0, sd. = 3.50 vs. $\overline{X}$ = 12.0, sd. = 2.60). In this respect, the analysis of verbal protocol comments revealed that, for example, NEW raters produced more comments on the correct use of prepositions and verb tenses than OLD raters. The less attentive attitude of OLD raters regarding grammar rules and accuracy, as compared to NEW raters, might be due to the fact that either these raters had not been trained to evaluate test takers' language performance, in particular accuracy, without resorting to score numbers but through commenting critically on the shortcoming, or they considered grammar rules not too severe. This latter argument is quite similar to Kim's (2011) results revealed that inexperienced raters produced higher frequency of protocols compared to experienced ones.

NEW raters, unlike OLD ones, reverted more to their initial decisions on the same or previous performance(s) to better handle their scoring judgments. This might have been due to their lack of experience compared to OLD raters, or perhaps they did their ratings with more hesitation because

[DOI: 10.29252/ijal.20.1.113]

they were aware of the fact that their ratings were part of a research study. At this phase, although 76% of the NEW raters' comments were positive on the effectiveness of the training program, OLD raters commented both positively and negatively on the matter. This demonstrates that NEW raters tended to be more lenient raters at this phase. NEW raters tended to comment on test takers more individually with only 15.6% of their comments involving inter-examinee comparisons. However, for OLD raters, around 51% of their comments involved inter-examinee comparisons. NEW raters were mostly analytical, whereas OLD ones were mainly holistic in a way that they were more likely to make summary comments.

- This subject has quite interesting performance. However, his performance was boring and monotonous, but he was creative and his speech was meaningful. *(Rater NEW4-Pre-training)*
- She communicates readily and accurately. She is quite fluent. She is rather like a native speaker. She tends to talk a lot. Her vocabulary storage and grammar are okay. She can say her intentions reasonably. *(Rater NEW2-Pre-training)*
- He can follow the conversation. His speeches are interconnected. *(Rater NEW5-Pre-training)*
- I see minor errors in the use of articles; however, the meaning is still clear. There are some incomplete sentences, too. I think I will give him a 5 in grammar and cohesion. *(Rater NEW3-Pre-training)*

In order to make sure whether the mean differences among OLD and NEW raters regarding their decision making behaviors for each behavioral factor was significant or not, an independent samples *t*-test was run. The results showed a significant mean difference for all the above mentioned mean differences. However, for the rest of the behavioral factors including *articulation of general impression*, *evaluation of relevance*, *classification of errors* and *evaluation of semantics*, no significant mean difference among the raters was observed.

[DOI: 10.29252/ijal.20.1.113]

Table 2

Analysis of raters' verbal protocols for NEW and OLD raters (pre-training)

Decision making behaviors	NEW raters		OLD raters		Both groups		Sig.
	Mean	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**	
Consider the situation of the examinee	6.51	1.29	12.55	2.71	9.53	2.00	*
Compare with other performances	8.37	1.73	16.80	3.64	12.58	2.68	*
Summarize the judgments	2.40	0.52	6.60	1.43	4.50	0.97	*
Articulate general impression	22.20	4.81	19.20	4.16	20.70	4.48	
Revise rating	62.40	6.52	43.20	6.36	52.80	6.44	*
Identify vague parts	3.00	0.65	6.63	1.39	4.81	1.02	*
Evaluate logic and reasoning	29.40	6.37	64.20	5.34	46.80	5.85	*
Evaluate relevance	6.00	1.30	8.40	1.82	7.20	1.56	
Evaluate novelty, originality and creativity	3.60	0.78	13.20	2.86	8.40	1.82	*
Identify information redundancies	1.80	0.39	12.13	2.67	6.96	1.53	*
Classify errors	33.60	5.28	16.80	3.64	25.20	4.46	
Evaluate the quantity of spoken data	44.40	4.28	31.80	6.89	38.10	5.58	*
Evaluate comprehensibility	12.54	2.66	17.40	3.77	14.97	3.21	*
Indentify frequency of errors	6.60	1.43	10.80	2.34	8.70	1.88	*
Evaluate pronunciation and accent	5.40	1.17	8.40	1.82	6.90	1.49	*
Evaluate fluency	2.38	0.44	4.77	1.12	3.575	0.78	*
Evaluate semantics	15.06	3.20	18.00	3.90	16.53	3.55	
Evaluate grammar	30.00	3.50	12.00	2.60	21.02	3.05	*
TOTAL	**16.42**	**2.57**	**17.93**	**3.24**	**17.18**	**2.90**	

[DOI: 10.29252/ijal.20.1.113]

Figure 1 displays the graphical representation of the raters' decision making behaviors at the pre-training phase.

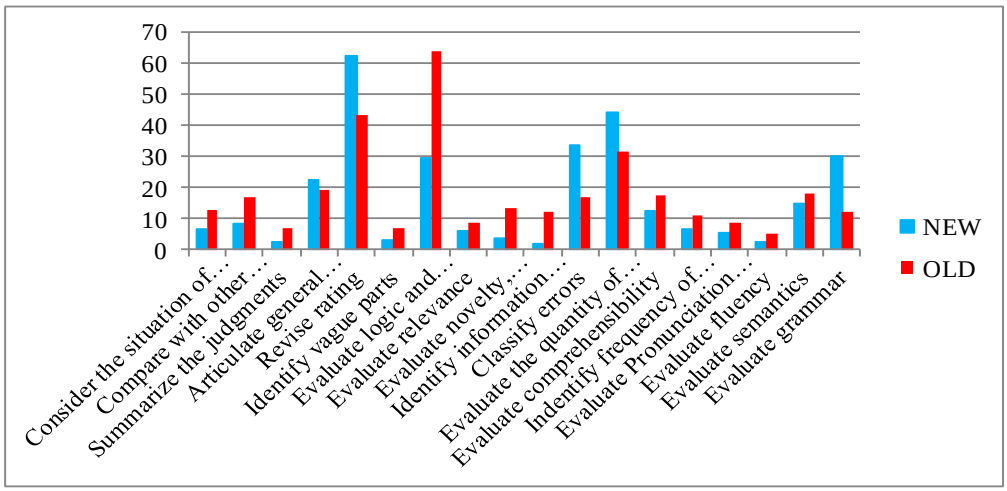

Figure 1. NEW and OLD raters' quantity of produced protocols of their decision making behavior (Pre-training)

The above figure, as indicated already, demonstrates that OLD raters, on average, provided more protocol comments when rating the test takers' oral performances. Among the scale categories, cohesion was perhaps the most challenging one even for OLD raters. The data protocols revealed that they could hardly make a conclusive decision over the issue.

- The use of cohesive devices doesn't seem to be good enough. So, maybe something between two or three. Well, it is not that much bad. Maybe a four even could be OK. *(Rater OLD3-Pre-training)*
- I notice the lack of cohesive devices here. It caused a big problem in organization, too. It doesn't seem effective. I give it a three. *(Rater OLD9-Pre-training)*

[DOI: 10.29252/ijal.20.1.113]

Although raters seemed to have understood the content of rating scale categories, the protocol analyses also revealed that raters had some difficulty understanding and applying the rating scale descriptors.

- The idea is OK, so I give it a four. However, there is something wrong with the meaning. This makes the task difficult to understand. So I give it a three instead. *(Rater OLD4-Pre-training)*
- The task asks them to describe their work-place. It is meaningful, but she didn't really describe it. She talked about her interest in it. So it is meaningful but inappropriate. The descriptor doesn't make a clear-cut difference between them. So, I consider it inappropriate. I give it a one. *(Rater OLD8-Pre-training)*

NEW raters, besides having the same problem, mostly compared their ratings across the other test takers, and the scores they gave, in many cases, rarely related to the scale descriptors but to the previous test takers.

- This one is almost related to the topic. I gave the other candidate a three, and since this guy spoke less I should mark him down, I give him a two. *(Rater NEW7-Pre-training)*
- The amount of speech is not enough. So, a three would be good. The meaning is also confusing. So, why not a two compared to the one before. But the vocabularies are good enough. So, maybe between three and four. *(Rater NEW4-Pre-training)*

Regarding the practicality and facility of verbal protocol production, NEW and OLD raters had different ideas. At the pre-training phase, 12 raters (8 NEW and 4 OLD) expressed difficulty in verbal production. These raters felt that thinking aloud reduced their rating speed to a high extent because sometimes they had to listen to the performance several times. Five NEW raters reported that thinking aloud lowered their self-confidence in rating because it made them doubt about their ratings and felt that they were being monitored and tested. However, three raters (2 OLD and 1 NEW)

[DOI: 10.29252/ijal.20.1.113]

reported that thinking aloud helped them better consider and rate the performances.

- I feel I am under control by an outsider. *(Rater NEW8-Pre-training)*
- Thinking aloud helped me be aware of the scoring process that I didn't use to think about in my previous ratings. *(Rater NEW9-Pre-training)*
- Thinking aloud enabled me look more carefully and pay more attention to various aspects of rating. *(Rater OLD1-Pre-training)*

A number of raters had contradictory reports about their ability to concentrate on the rating scale categories and other similar criteria when rating.

- I might have ignored and skipped some features that were important in rating. *(Rater NEW3-Pre-training)*
- Thinking loudly helped me focus my attention on the rating factors; something I wouldn't normally do when rating silently. *(Rater OLD7-Pre-training)*

Three raters (OLD) indicated that rating loudly influenced the scores they assigned.

- Thinking aloud has occasionally changed my mind in scoring the students. Several times, when I selected a score and talked about it loudly, suddenly I changed my mind and gave another score. *(Rater OLD2-Pre-training)*

Nevertheless, it is noteworthy to note that the remaining 17 raters (10 NEW and 7 OLD) reported that thinking aloud did not affect their scoring. Some raters (4 NEW and 1 OLD) expressed difficulty employing think-aloud technique because it distracted their attention from rating.

[DOI: 10.29252/ijal.20.1.113]

- Speaking while rating and hearing your own voice while making decisions on the scores interfere with each other and makes the rating process erroneous. *(Rater NEW5-Pre-training)*
- Instead of concentrating on scoring the tasks I should direct my attention to my speaking and the things I'm going to say. Then I want to focus on the task and I forget that I was supposed to speak. *(Rater OLD6-Pre-training)*

However, some others had other viewpoints.

- Hearing your own voice helps you rethink about the errors and better able to detect more of them. *(Rater NEW1-Pre-training)*

One of the raters, rater OLD3, suggested that a combination of both thinking and rating with the flexibility of being silent and loud at times not only would provide enough insight about the rating pattern, but also would not interfere with the rating process. Thinking aloud also seems to put the focus of the raters' attention to the shortcomings and inefficiencies of test takers' performance, thus making raters severer than they might be.

- I was a bit more lenient before rating silently. Thinking loudly made me assign lower scores than I would give before. *(Rater OLD4-Pre-training)*
- When rating aloud, automatically your attention is drawn to the mistakes. Errors appear more noticeable than before and errors become salient. *(Rater NEW6-Pre-training)*

In general, the analysis of the data protocols obviously demonstrated the outperformance of OLD raters over NEW ones regarding the better and more enhanced use of decision making behaviors when rating test takers' spoken performances.

Table 3 exhibits the raters' change of decision making behaviors at the immediate post-training phase. At first glance, the table shows a drastic shift in decision making behaviors to benefit NEW raters. Data analysis revealed

[DOI: 10.29252/ijal.20.1.113]

that OLD raters merely used more *logic and reasoning in their assessment* ($\overline{X}$ = 67.80, sd. = 9.80) and could better *identify redundancies* ($\overline{X}$ = 12.11, sd. = 1.73) compared to NEW ones ($\overline{X}$ = 43.80, sd. = 6.14, and $\overline{X}$ = 7.20, sd. = 1.01 respectively).

- This subject responded with short and separated answers. For every detail, I should ask her a new question. But the former one also provided me with more detailed information. *(Rater OLD4-Post-training)*
- He is like a beginner. He seems frustrated. He only gives very short answers. He is hesitant in his answers. His pauses are long, too. *(Rater OLD3-Post-training)*

The ongoing leading performance of OLD raters compared to NEW ones in the two above-mentioned protocol factors can be due to the fact that such qualities are experientially-oriented and for which short-term training programs may not be as useful as expected. In other words, training programs cannot enhance raters' logic and reasoning ability in judgment as the relevant experience can. Consequently, these factors have their roots in rating experience which is parallel with Bowles' (2010) finding that demonstrated experienced raters used more logic and reasoning while they were rating.

Nevertheless, NEW raters devoted more enhanced attention to *making comparisons with other examinees' performances* ($\overline{X}$ = 12.00, sd. = 1.68 vs. $\overline{X}$ = 6.04, sd. = 0.87), *summarizing* ($\overline{X}$ = 13.80, sd. = 1.94 vs. $\overline{X}$ = 9.10, sd. = 1.30), *revising their own judgments* ($\overline{X}$ = 72.13, sd. = 10.16 vs. $\overline{X}$ = 57.60, sd. = 8.32), *classifying* ($\overline{X}$ = 43.26, sd. = 6.06 vs. $\overline{X}$ = 28.80, sd. = 4.16), *identifying the frequency of errors* ($\overline{X}$ = 16.80, sd. = 2.36 vs. $\overline{X}$ = 13.80, sd. = 1.99), *evaluating spoken data comprehensibility* ($\overline{X}$ = 22.77, sd. = 3.26 vs. $\overline{X}$ = 17.40, sd. = 2.51), *focusing more on the evaluation of semantics* ($\overline{X}$ = 22.80, sd. = 3.20 vs. $\overline{X}$ = 18.60, sd. = 2.69), *and grammar of examinees' performances* ($\overline{X}$ = 48.12, sd. = 6.73 vs. $\overline{X}$ = 36.09, sd. = 5.20). The higher amount of NEW raters' attention to accuracy and grammatical aspect of test

[DOI: 10.29252/ijal.20.1.113]

takers' oral performances may most probably be due to the fact that prior to the training program, the raters were not reminded that they should make their comments as specific as possible, which might have resulted in fewer comments by these raters compared to the OLD ones. However, after the training program and the provision of feedback, they tended to produce more than OLD raters.

The leading position of NEW raters compared to OLD ones in the above mentioned features signify the effectiveness of the training in providing NEW raters with a more powerful judgmental tool in rating. Besides, it also reflected that unlike the previously-mentioned factors of *logic and reasoning in assessment* and *identification of redundancies* for which training was not shown to be effective enough and that experience was identified to be a more important influential factor, training was quite effective in establishing higher consensus among NEW raters than OLD ones in the production of verbal protocols and rating for the remaining factors. For instance, with regard to 'grammatical and semantic considerations', NEW raters were shown to be less attentive than OLD ones prior to the training; however, they developed more consideration after training.

At this phase, NEW raters were still more positive about the effectiveness of the training program in enhancing rating consistency and reducing levels of biasedness than OLD ones. That is, 82.6% of NEW raters were positive about this effect, whereas only 59.18% of OLD raters showed positive attitude. Similar to the pre-training phase, NEW raters tended to be more lenient than OLD ones. Although the percentage of inter-examinee comparison showed a higher increase for NEW raters compared to OLD ones after the training program, still OLD raters displayed more inter-examinee comments than NEW ones (62.33% for OLD raters vs. 44.07% for NEW ones). The independent samples *t*-test result vividly verified a significant mean difference between NEW and OLD raters for all the above-mentioned differences. It is noteworthy to note that although in many cases NEW raters had higher means than OLD ones on the remaining behavioral factors including *consideration of the examinees' situations*, *articulation of general impression, identification of vague parts, evaluation of relevance,*

[DOI: 10.29252/ijal.20.1.113]

novelty, originality and creativity, e*valuation of the quantity of spoken data* as well as *pronunciation, accent and fluency,* no significant mean difference was observed between NEW and OLD raters regarding the quantity of the verbal comments produced on these factors. This finding showed that training program could build more rapport between the raters of the two groups of expertise and thus have a constructive effectiveness.

Table 3

Analysis of raters' verbal protocols for NEW and OLD raters (immediate post-training)

Decision making behaviors	NEW raters		OLD raters		Both groups		Sig.
	Mean	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**	
Consider the situation of the examinee	23.40	3.28	21.03	3.03	22.21	3.15	
Compare with other performances	12.00	1.68	6.04	0.87	9.02	1.27	*
Summarize the judgments	13.80	1.94	9.10	1.30	11.45	1.62	*
Articulate general impression	30.06	4.21	25.20	3.64	27.63	3.92	
Revise rating	72.13	10.16	57.60	8.32	64.86	9.24	*
Identify vague parts	7.20	1.01	4.84	0.69	6.02	0.85	
Evaluate logic and reasoning	43.80	6.14	67.80	9.80	55.8	7.97	*
Evaluate relevance	9.60	1.35	10.20	1.47	9.90	1.41	
Evaluate novelty, originality and creativity	24.02	3.37	18.60	2.69	21.31	3.03	
Identify information redundancies	7.20	1.01	12.11	1.73	9.655	1.37	*
Classify errors	43.26	6.06	28.80	4.16	36.03	5.11	*
Evaluate the quantity of spoken data	46.80	6.56	36.60	5.29	41.70	5.92	
Evaluate comprehensibility	22.77	3.26	17.40	2.51	20.08	2.88	*
Indentify frequency of errors	16.80	2.36	13.80	1.99	15.30	2.17	*
Evaluate pronunciation and accent	8.40	1.18	6.60	0.95	7.50	1.06	
Evaluate fluency	12.60	1.77	10.80	1.56	11.70	1.66	
Evaluate semantics	22.80	3.20	18.60	2.69	20.70	2.94	*
Evaluate grammar	48.12	6.73	36.09	5.20	42.10	5.96	*
TOTAL	**25.82**	**3.62**	**22.28**	**3.21**	**24.05**	**3.41**	

[DOI: 10.29252/ijal.20.1.113]

Figure 2 displays the graphical representation of the raters' decision making behaviors at the immediate post-training phase.

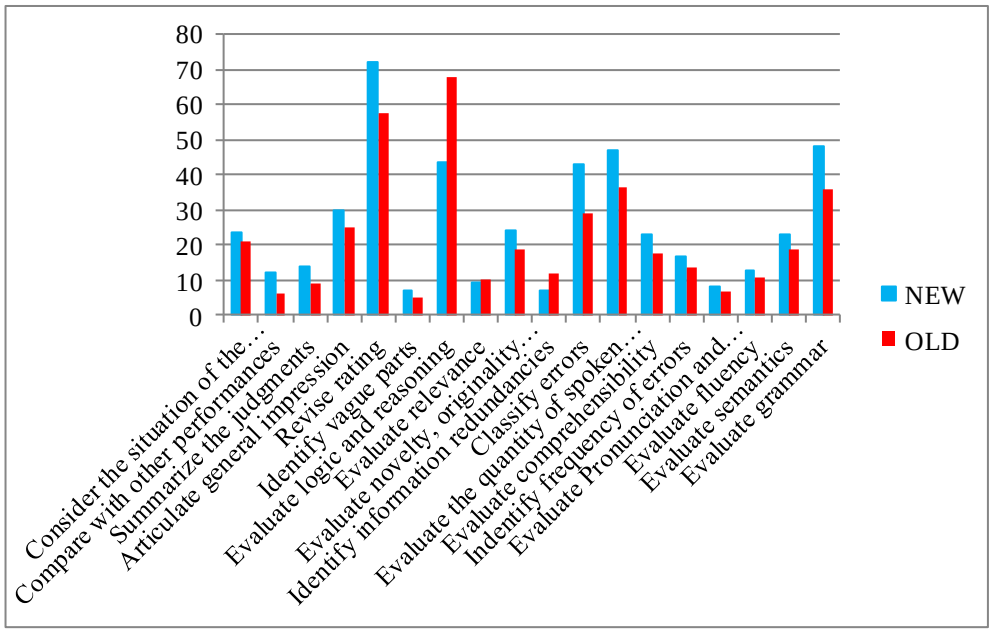

Figure 2. NEW and OLD raters' quantity of produced protocols on their decision-making behavior (Immediate post-training)

Unlike the pre-training phase, the above figure displays that NEW raters tended to produce more protocols than OLD ones after training showing that the training program could better motivate NEW raters than OLD ones in protocol production. Altogether, the analysis of data protocol at this phase vividly represented a drastic shift of behavior to benefit NEW raters over OLD ones compared to the pre-training phase. NEW raters obviously outperformed OLD ones with regard to the quality and quantity of the macro and micro strategies they used to evaluate test takers' performances.

[DOI: 10.29252/ijal.20.1.113]

- She has some pronunciation and intonation problems; however, she can infer her ideas well. I loved the way she justified her ideas. She relates what she has just said with what she is going to say very well. She builds up on her previous ideas. *(Rater NEW6-Post-training)*
- This candidate has native-like pronunciation and has made very few grammatical mistakes. However, her performance on the description task was not satisfactory. She still shows incorrect use of definite and indefinite articles. *(Rater OLD8-Post-training)*
- This student has strange facial and body expressions. She herself seems not to understand what she is saying. I suppose she is not interested in this topic. I really hardly understand what she says. *(Rater NEW9-Post-training)*

Table 4 represents the raters' change of behavior at the delayed post-training phase. On the first look, in spite of the reduction of raters' change of decision-making behaviors for both groups, the results show the superiority of NEW raters over OLD ones. In other words, despite the reduction in almost every factor of raters' protocols, NEW raters still outperformed OLD ones in applying decision-making behavior factors. Similar to the immediate post-training phase, the protocol analysis revealed that OLD raters dominated NEW ones in *logic and reasoning evaluation* ($\overline{X}$ = 60.65, sd. = 10.45 vs. $\overline{X}$ = 39.10, sd. = 6.72) and *redundancy identification* ($\overline{X}$ = 10.28, sd. = 1.76 vs. $\overline{X}$ = 6.17, sd. = 1.03).

- He can make connection between what he says and what he has said before. He also comments on whatever he says. He supports his utterances with reasons. *(Rater OLD3-Delayed post-training)*

As already mentioned, the reason why the above-mentioned factors tended to be constantly higher for OLD raters than NEW ones, even after the training program, is that such factors are rooted in raters' experience for which training does not seem to be that much effective. These are the concepts raters build as a result of constant interaction with the rating task rather than participating in short-term training programs.

[DOI: 10.29252/ijal.20.1.113]

In contrast, NEW raters, unlike OLD ones, dedicated more attention *to making comparisons with other examinees' performances* ($\overline{X}$ = 13.21, sd. = 2.28 vs. $\overline{X}$ = 6.44, sd. = 1.11), *revising their judgments* ($\overline{X}$ = 72.28, sd. = 12.41 vs. $\overline{X}$ = 57.60, sd. = 9.93), *identifying the frequency of errors* ($\overline{X}$ = 16.80, sd. = 2.91 vs. $\overline{X}$ = 12.19, sd. = 2.07) and *classifying them* ($\overline{X}$ = 41.40, sd. = 7.14 vs. $\overline{X}$ = 25.85, sd. = 4.45) as well as *concentrating on the evaluation of semantics* ($\overline{X}$ = 21.62, sd. = 3.72 vs. $\overline{X}$ = 14.47, sd. = 2.48) *and grammatical structure of examinees' performances* ($\overline{X}$ = 45.66, sd. = 7.86 vs. $\overline{X}$ = 32.42, sd. = 5.59).

- She repeats and rephrases her sentences several times. She keeps sighing. She has very long pauses and she cannot coherently connect her sentences. It is quite hard to follow her interactions. *(Rater NEW2-Delayed post-training)*
- He confirms what he says a lot. His over-use of body language shows that he is very confident. His body movements are too much. However, he beautifully changes his intonation depending on the context. *(Rater NEW7-Delayed post-training)*

Although the relative reduction of produced protocols in the above mentioned factors at the delayed post-training phase signify raters' tendency to forget their understandings of the training program instructions, both groups of raters still demonstrated higher frequency of production of protocols compared to the pre-training phase. It is noteworthy to indicate that NEW raters still took the lead proving to outperform OLD ones although their performance was fairly less than the immediate post-training phase. At this phase, still NEW raters showed a more positive attitude to the effectiveness of the training program in building more consistency and reduced levels of biasedness than OLD ones. In fact, 74.3% of NEW raters were positive in this respect, whereas only 62.27% of OLD raters showed such a positive attitude. Like the previous two phases, NEW raters tended to be more lenient than OLD ones. Once again the extent of raters' inter-examinee comparisons was measured to evaluate any possible change

[DOI: 10.29252/ijal.20.1.113]

throughout the study. The outcome showed that still OLD raters displayed more inter-examinee comments than NEW ones (54.45% for OLD raters vs. 48.24% for NEW ones).

The results of the independent samples *t*-test provided evidence on the significant mean difference between the two groups of raters for all the above-mentioned differences. It must be noted that although in many cases NEW raters had higher means than OLD ones on the remaining behavioral factors including *consideration of the examinees' situations*, *summarization of their judgments*, *articulation of general impression*, *identification of vague parts*, *evaluation of relevance, novelty, originality and creativity*, e*valuation of the quantity of spoken data*, *evaluation of comprehensibility* as well as *pronunciation, accent and fluency*, no significant mean difference was observed between them regarding the quantity of the verbal comments.

Although this finding reiterates the raters' tendency in both groups to forget the program's instructions as the result of time, it calls up on the fact that NEW raters benefitted more from the training program than OLD ones. In general, NEW raters still seemed to benefit more from the use of instructed macro and micro strategies in terms of both quantity and quality in the assessment of test takers' speaking performance.

[DOI: 10.29252/ijal.20.1.113]

Table 4
Analysis of raters' verbal protocols for NEW and OLD raters (delayed post-training)

Decision making behaviors	NEW raters		OLD raters		Both groups		Sig.
	Mean	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**	
Consider the situation of the examinee	16.83	2.77	13.80	2.38	15.31	2.57	
Compare with other performances	13.21	2.28	6.44	1.11	9.82	1.69	*
Summarize the judgments	14.44	2.48	9.33	1.62	11.88	2.05	
Articulate general impression	30.17	5.17	25.20	4.34	27.68	4.75	
Revise rating	72.28	12.41	57.60	9.93	64.94	11.17	*
Identify vague parts	6.20	1.07	4.20	0.72	5.2	0.89	
Evaluate logic and reasoning	39.10	6.72	60.65	10.45	49.87	8.58	*
Evaluate relevance	12.63	2.17	8.43	1.45	10.53	1.81	
Evaluate novelty, originality and creativity	21.66	3.72	17.45	3.00	19.55	3.36	
Identify information redundancies	6.17	1.03	10.28	1.76	8.22	1.39	*
Classify errors	41.40	7.14	25.85	4.45	33.62	5.79	*
Evaluate the quantity of spoken data	45.59	7.86	33.64	5.79	39.61	6.82	
Evaluate comprehensibility	18.63	3.21	15.38	2.59	17.00	2.90	
Indentify frequency of errors	16.80	2.91	12.19	2.07	14.49	2.49	*
Evaluate pronunciation and accent	6.57	1.14	4.84	0.83	5.70	0.98	
Evaluate fluency	10.23	1.76	9.12	1.57	9.67	1.66	
Evaluate semantics	21.62	3.72	14.47	2.48	18.04	3.10	*
Evaluate grammar	45.66	7.86	32.42	5.59	39.04	6.72	*
TOTAL	**24.39**	**4.19**	**20.07**	**3.45**	**22.23**	**3.81**	

[DOI: 10.29252/ijal.20.1.113]

Figure 3 displays the graphical representation of the raters' decision-making behaviors at the delayed post-training phase.

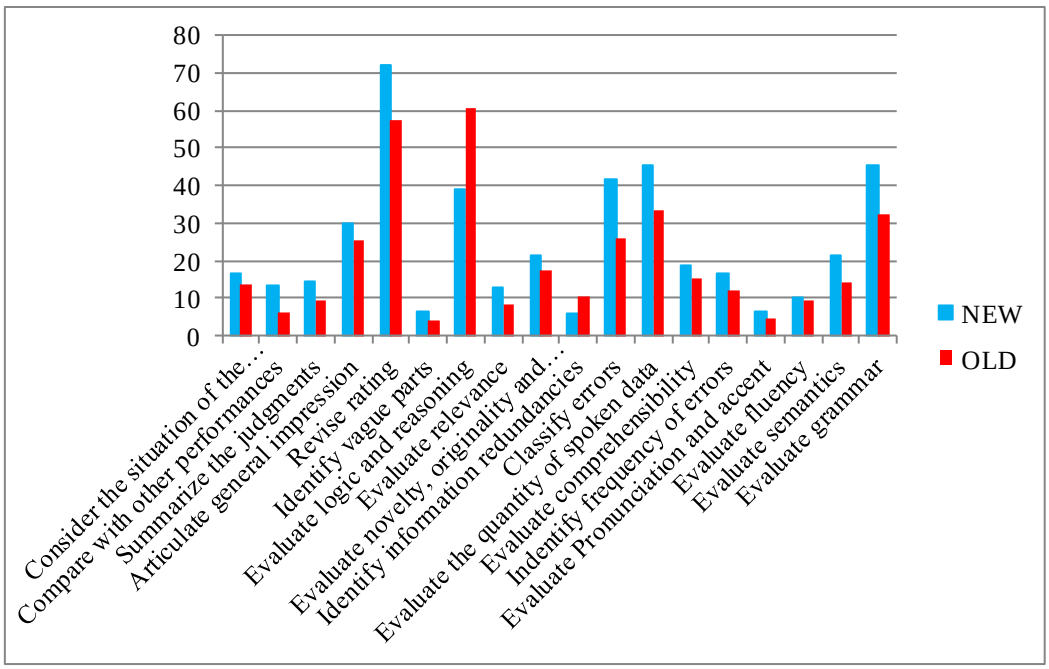

Figure 3. NEW and OLD raters' quantity of produced protocols of their decision making behavior (Delayed post-training)

Through observing the statistics given in the above figure, it is well understood that both NEW and OLD raters had fewer protocols produced at the delayed post-training phase. Therefore, the results demonstrated that the rater-training program had constructive effectiveness because raters' performances still demonstrated improvement at the delayed post-training phase compared to the pre-training phase. However, the results also demonstrated a gradual loss of program effectiveness as a result of time.

[DOI: 10.29252/ijal.20.1.113]

5. Discussion

The findings of the study provided enough evidence for the effectiveness of the think-aloud verbal protocols. The findings of the study showed that verbal protocols could shed more light on the vague parts of the rating task, which a mere use of statistical analysis does not reveal. Such finding is similar to that of some other researchers (e.g., Knoch, 2009; Sawaki, 2007; Wolfe, 2004). The findings of verbal protocol analysis also demonstrated that the use of body language and non-verbal behavior was a contributing element to the success of oral interaction and that various types of non-verbal language provided evidence on the ability to communicate in a second/foreign language. This finding is in line with that of Ducasse and Brown (2009) who found that body language has a key feature in interpersonal interaction.

The use of verbal protocols, similar to Kim's (2011) finding, was shown to benefit both groups of rater expertise through establishing higher degrees of consensus in the use of rating scale guidelines. Also, the analysis of verbal protocols demonstrated that raters were able to better match their ratings in accordance with the rating scales throughout the entire study. They tended to stick more to the descriptors of the scale rubric in the immediate and delayed post-training phases than the pre-training phase. This finding is in line with that of Attali (2016) who found that training helped raters classify test takers' errors based on the requirements of the rating scale. However, it is rather in contrast with Lumley's (2005) finding, who through the analysis of verbal protocols in writing, found no improvement in raters' use of rating scale to match their descriptors. However, unlike Lumley's (2005), who found *grammar* to be the severest rating category, the findings of this study indicated it as rather the least severe scale category, specifically for the inexperienced raters. The findings of the study on raters' attitude toward the production of verbal protocols during rating, is relatively parallel with those of Ling, Mollaun, and Xi (2014) who found verbal protocols a very difficult and demanding process.

The finding of this research, which showed that raters' attention was driven to other aspects of the rating criteria is similar to that of Kim (2015),

[DOI: 10.29252/ijal.20.1.113]

who found that raters tended to concentrate more on communicative aspect of speech rather than accuracy. Besides, OLD raters in this study were found to produce think-aloud protocols with more difficulty than NEW ones. This may be due to the fact that the verbalized thoughts produced by OLD raters tended to be so complex that made the parallel tasks of 'achieving homogeneity with other raters' and 'scoring the performances' quite challenging. This finding is in line with that of Barakoui (2011), Cumming, Kantor, and Powers (2002), and Davis (2016) who found that experienced raters had hardship verbalizing and articulating their thoughts aloud than inexperienced ones. However, Barakoui's study was on assessing writing performance.

In addition, the findings of the study indicated that think-aloud protocols, although quite useful, were incomplete because they could affect the rating process. Think-aloud protocols seemed to have affected the rating process in terms of performance comprehension, rating scale criteria, raters' self-confidence, the sense of privacy, and of course their decision making. This finding was rather consistent with some previous research (e.g., Barakoui, 2011; Kuiken & Vedder, 2014; Lumley, 2005; Sasaki, 2014) that found the effect of verbal protocol production on raters' scoring patterns. One important finding of the analysis of raters' verbal protocols was that raters seemed to have focused more on micro-level errors when rating aloud, although they also have paid considerable attention on overall comprehensibility of spoken discourse when rating silently as well. Therefore, employing a combination of both techniques, as suggested by one of the raters (OLD3), might be more effective. Attali (2016) found rather similar results in his study in which raters attended more to the organization of writing (macro-level errors) when rating silently, whereas they focused more on mechanics of writing (micro-level errors) when rating aloud.

This research also revealed that the use of verbal reports, as verified in the previous studies (e.g., Attali, 2016; Davis, 2016; Nakatsuhara, 2011), could be as effective on oral assessment as it has already been shown to be on writing assessment. Raters were able to verbalize their thoughts. Besides, they were able to focus their attention on the salient features; therefore, they

[DOI: 10.29252/ijal.20.1.113]

achieved a higher degree of rating agreement. Although there were some contradictory comments and some negativity due to the influence of verbal protocols on raters' scoring, verbal protocols were generally perceived to be useful, especially by inexperienced raters. The analysis of verbal protocols also demonstrated that having a well-defined and understandable rating scale could definitely foster a valid and reliable oral assessment. Moreover, it could also add up to the effectiveness of a training program.

6. Conclusion

Various groups of raters approach the task of rating in different ways. However, this is something to which a mere use of statistical analysis cannot be responsive. Therefore, the use of think-aloud verbal protocols can shed light on the probable vague sides of the issue and add to the validity of oral language performance assessment. Further research could be carried out to investigate the effect of individual differences and reflection on think-aloud protocols on raters' scoring performances.

Moreover, the study showed that NEW raters could rate as reliably as, or even more reliable in some aspects, than experienced raters based on the quality and quantity of the macro and micro strategies used to evaluate test takers' performances. The implication is that decision makers, when choosing raters, do not need to employ only experienced raters because the results of this study provided no evidence for the exclusion of inexperienced raters. Moreover, inexperienced raters are more economical than the experienced ones. They also showed to be more reliable after training or even without training – if standards are concerned. Although it is a general belief for decision makers to select experienced raters for achieving higher reliability, the findings showed the reverse. Therefore, instead of allocating a higher budget to the employment of experienced raters, decision makers allocate that budget to administering more efficient training programs.

One other way of investigating raters' scoring behavior, in addition to the use of verbal protocols, is the adoption of concept map (Papajohn, 2002) as a quick and convenient way to access the same information. Concept mapping graphically represents meaningful relationships among various

[DOI: 10.29252/ijal.20.1.113]

ideas by the use of charts and other depicting processes. Thus, further research could probe into the effect of using concept mapping to observe raters' decision-making behaviors. Finally, the findings of the study, which revealed the raters' viewpoints about verbalizing their thoughts, should be generalized with care due to the limited number of raters and their unfamiliarity with this type of scoring rubric. Obviously, more research with more raters would support the findings of this research and/or would shed light on aspects which this study might have failed to discover.

7. References

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study on their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

Cohen, A. D. (1994). Verbal reports on learning strategies. *TESOL Quarterly, 28*(4), 678-682.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

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

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423-443.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. Princeton, NJ: Educational Testing Service.

[DOI: 10.29252/ijal.20.1.113]

Ericsson, K. A., & Simon, H. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment.* Unpublished PhD thesis, University of Columbia.

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

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior –a longitudinal study. *Language Testing*, *28*(2), 179-200.

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing, 31*(3), 279-284.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt, Germany: Peter Lang.

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

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

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, *14*(2), 140-156.

Nakatsuhara, F. (2011). Effect of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483-508.

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

[DOI: 10.29252/ijal.20.1.113]

Sasaki, T. (2014). Recipient orientation in verbal report protocols: Methodological issues in concurrent think-aloud. *Second Language Studies, 22*(1), 1-54.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99-123.

Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. *Annual Review of Applied Linguistics, 21*(3), 233-245.

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22.

Wagner, M. J. (2006). *Utilizing the visual channel: An investigation of the use of video texts of second language listening ability.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Wallace, M. J. (1991). *Training foreign language teachers-A reflective approach.* Cambridge: Cambridge University Press.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35-51.

[DOI: 10.29252/ijal.20.1.113]

Notes on Contributors:

Houman Bijani is a Ph.D. candidate of TEFL at Islamic Azad University, Science and Research Branch, Tehran, Iran. He is also a faculty member of Zanjan Azad University. He got his MA in TEFL from Allameh Tabatabai University as a top student. He has published several research papers in national and international language teaching journals. His areas of research interest include quantitative assessment, teacher education, and language research.

Mona Khabiri is associate professor of Applied Linguistics at Islamic Azad University, Central Tehran Branch and the director of Journal of English Language Studies (JELS). She mainly teaches language testing, research methodology, seminar in TEFL issues, and teaching language skills at graduate level. Her main areas of research interest include teacher education, cooperative learning, and language testing and research. She has published papers in international and national academic journals.

[DOI: 10.29252/ijal.20.1.113]