THE EFFECTIVENESS OF VSTEP.3-5 SPEAKING RATER TRAINING

Nguyen Thi Ngoc Quynh, Nguyen Thi Quynh Yen, Tran Thi Thu Hien, Nguyen Thi Phuong Thao, Bui Thien Sao*, Nguyen Thi Chi, Nguyen Quynh Hoa

VNU University of Languages and International Studies, Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 09 May 2020
Revised 10 July 2020; Accepted 15 July 2020

Abstract: Playing a vital role in assuring reliability of language performance assessment, rater training has been a topic of interest in research on large-scale testing. Similarly, in the context of VSTEP, the effectiveness of the rater training program has been of great concern. Thus, this research was conducted to investigate the impact of the VSTEP speaking rating scale training session in the rater training program provided by University of Languages and International Studies - Vietnam National University, Hanoi. Data were collected from 37 rater trainees of the program. Their ratings before and after the training session on the VSTEP.3-5 speaking rating scales were then compared. Particularly, dimensions of score reliability, criterion difficulty, rater severity, rater fit, rater bias, and score band separation were analyzed. Positive results were detected when the post-training ratings were shown to be more reliable, consistent, and distinguishable. Improvements were more noticeable for the score band separation and slighter in other aspects. Meaningful implications in terms of both future practices of rater training and rater training research methodology could be drawn from the study.

Keywords: rater training, speaking rating, speaking assessment, VSTEP, G theory, many-facet Rasch

1. Introduction

Rater training has been widely recognized as a way to assure the score reliability in language performance assessment, especially in large-scale examination (Luoma, 2004; Weigle, 1998). A large body of literature has been spent on how to conduct an efficacious rater training program and to what extent rater training program had impact on raters' ratings. More specifically, documents have shown that in line with general education measurement, rater training procedures in language assessment were also framed into four main approaches namely rater error training (RET), performance dimension training (PDT), frame-of-reference training (FORT), and behavioral observation training (BOT). The effectiveness of rater training and these approaches were the topic of interest for numerous researchers either in educational measurement or language assessment such as Linacre (1989), Weigle (1998), Roch and O'Sullivan (2003), Luoma (2004), Roch, Woehr, Mishra, and Kieszczynska (2011).

The same concern arose for the developers of the Vietnamese Standardized Test of English Proficiency (VSTEP). Officially introduced

* Corresponding author. Tel.: 84-968261056
Email: sao.buithien@gmail.com

in 2015 as a national high-stake test by the government, VSTEP level 3 to 5 (VSTEP.3-5) has been considered to be a significant innovation in language testing and assessment in Vietnam, responding to the demands of "creating a product or service with a global perspective in mind, while customising it to fit 'perfectly' in a local market" (Weir, 2020). This launching then led to an urgent demand of quality assurance in all processes of test development, test administration, and test rating. As a result, a ministerial decision on VSTEP speaking and writing rater training was issued in the later year (including regulations on curriculum framework, capacity of training institutions, trainer qualification and minimum language proficiency and teaching experience requirements of trainees). Being assigned as a training institution, University of Languages and International Studies (ULIS) has implemented the training program from then on. Inevitably, the impact of the rater training program has drawn attention from many stakeholders.

As an attempt to examine the effectiveness of the ULIS rater training program and enrich the literature of this field in Vietnam, a study was conducted by the researchers – also the organizer team of the program. In the scope of this study, the session on speaking rating scales, the heart of the training program for raters of speaking skill, was selected to investigate.

2. Literature review

With regard to performance assessment, there is a likelihood of inconsistency within and between raters (Bachman & Palmer, 1996; McNamara, 1996; Eckes, 2008; Weigle, 2002; Weir, 2005). Eckes (2008) synthesized various ways in which raters may differ: (a) in the degree to which they comply with the scoring rubric, (b) in the way they interpret criteria employed in operational scoring sessions, (c) in the degree of severity or leniency exhibited when scoring examinee performance, (d) in the understanding and use of rating scale categories, or (e) in the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. (p.156). The attempt to minimize the divergence among raters was the rationale behind all the rater training programs of all fields.

Four rater training strategies or approaches have been described in many previous studies, namely rater error training (RET), performance dimension training (PDT), frame-of-reference training (FORT), and behavioral observation training (BOT). All of these strategies aim to enhance the rater quality, but each demonstrates different key features. While RET is used to caution raters of committing psychometric rating errors (e.g. leniency, central tendency, and halo effect), PDT and FORT focus on raters' cognitive processing of information by which the rating accuracy is guaranteed. Although PDT and FORT are similar in that they provide raters with the information about the performance dimensions being rated, the former just involves raters in co-creating and/ or reviewing the rating scales whereas the latter provides standard examples corresponding to the described dimensions (Woehr & Huffcutt, 1994, p.190-192). In other words, through PDT raters accustom themselves to the descriptors of each assessment criterion in the rating scale, and through FORT raters have chances to visualize the rating criteria by means of analyzing the sample performances corresponding to specific band scores. The last common training strategy, BOT, focuses on raters' observation of behaviors rather than their evaluation of behavior. To put it another way, BOT is used to train raters to become skilled observers who are able to recognize or recall the performance aspects consistent with the rating scale (Woehr & Huffcutt, 1994, p.192).

A substantial amount of research in the field of testing and assessment has put an emphasis on rater training (Pulakos, 1986; Woehr & Huffcutt, 1994; Roch & O'Sullivan, 2003; Roch, Woehr, Mishra, & Kieszczynska, 2011; to name but a few) in an attempt for improving the rating, yet the findings about its efficiency seem to be inconsistently documented. Many researchers and scholars posited that RET reduced halo and leniency errors (Latham, Wexley, & Pursell, 1975; Smith, 1986; Hedge & Kavanagh, 1988; Rosales Sánchez, Díaz-Cabrera, & Hernández-Fernaud, 2019). These authors assumed that when raters are more aware of the rating errors they may commit, their ratings are likely to be more accurate. Nonetheless, the findings of Bernardin's and Pence's (1980) research showed that rater error training is an inappropriate approach to rater training and that this approach is likely to result in decreased rating accuracy. Hakel (1980) clarified that it would be more appropriate to term this approach as training about rating effects and that the rating effects represent not only errors but also true score variance. It means that "if these rating effects contain both error variance and true variance, training that reduces these effects not only reduces error variance, but affects true variance as well (cited in Hedge & Kavanagh, 1988, p.68).

In the meantime, certain evidence for the efficacy of rater training has been recorded for the other rating strategies, PDT (e.g. Hedge & Kavanagh, 1988; Woehr & Huffcutt, 1994), FORT (e.g. Hedge & Kavanagh, 1988; Noonan & Sulsky, 2001; Roch et al., 2011; Woehr & Huffcutt, 1994), and BOT (e.g. Bernardin & Walter, 1977; Latham, Wexley & Pursell, 1975; Thornton & Zorich, 1980, Noonan & Sulsky, 2001); particularly, FORT has been preferable for improving rater accuracy. However, Hedge and Kavanagh (1988) cautioned about the limited generalizability of the results in FORT. Specifically, in this training approach, the trainees are provided with the standard frame of reference as well as observation training on the correct behaviors. In other words, the results are dependent on the samples, which can hardly be generalized in all circumstances. Moreover, Noonan and Sulsky (2001) highlighted that FORT revealed weakness in that it did not facilitate raters in remembering specific test takers' behaviors, which might lead raters to false assessment in comparison to the described criteria.

In consideration of strengths and weaknesses of each training approach, an increasing number of researchers and scholars have had an idea of combining different approaches to enhance the effectiveness of rater training. For example, RET was combined with PDT or FORT (McIntyre, Smith, & Hassett, 1984; Pulakos, 1984), or FORT was combined with BOT (Noonan & Sulsky, 2001; Roch & O'Sullivan, 2003). Noticeably, no significant increase in rating accuracy has been reported. Nonetheless, the number of studies on the combination of different approaches was modest, which makes conclusion on its efficacy yet to be reached.

With a hope to enhance the impact on rating quality in the context of VSTEP, a combination of all four approaches was employed during the course of rater training program. However, similar to the general context with limited research on integrated approach in rater training, research in Vietnam has recorded to date few papers on language rater training and no papers on the program for VSTEP speaking raters, not to mention intensive training on rating scales. Therefore, it is significant to undertake the present study to examine whether the combination of multiple training strategies has an impact on performance ratings and what aspects of the ratings are impacted.

3. Research questions

Overall, this study was implemented to, firstly, shed light on the improvement (if any) of the reliability of the scores given by speaking raters after they received training on the VSTEP.3-5 speaking rating scales. Secondly, the study expanded to scrutinize the impact of the training session on other aspects namely criterion difficulty, rater severity, rater fit, rater bias, and score band separation. Accordingly, two research questions were formulated as follow.

1. How is the reliability of the VSTEP.3-5 speaking scores impacted after rater training session on rating scales?
2. How are the aspects of criterion difficulty, rater severity, rater fit, rater bias, and score band separation impacted after rater training session on rating scales?

4. Methodology

4.1. Participants

The research participants were 37 rater trainees of the rater training program delivered by ULIS. They worked as teachers of English carefully selected by their home institutions. Some prerequisite requirements for them to enroll in this course include C1 English proficiency level based on the Common European Framework of Reference (CEFR) or level 5 according to the CEFR – VN and at least 3 years of teaching experience. Additionally, good background on assessment is preferable. Some of them had certain experience with VSTEP as well as VSTEP rating, while the majority had the very first-hand experience to the test in the training course. With such a pool of participants, the study was expected to evaluate the rating accuracy of novice VSTEP trainee raters. It can be said that they were all motivated to take the intensive training program since they were commissioned to their study as the representatives of their home institutions, and some were financially bonded with their institutions. When being invited to participate in the study, all participants were truly devoted as they considered it a chance for them to see their progress in a short duration.

4.2. The speaking rater training program

A typical training program for speaking raters at ULIS lasts for 180 hours, consisting of both 75 hour online and 105 hour on-site training. The program is described in brief in this table below.

Table 1: Summary of rater training modules for speaking raters

	Theories of Testing and Assessment
Module 2	Rater Quality Assurance
Module 3	Theories of Speaking Assessment
Module 4	The CEFR
Module 5	CEFR Descriptors for Grammar & Vocabulary
Module 6	VSTEP Speaking Test Procedure
Module 7	VSTEP Speaking Rating Scales
Module 8	Rating practices with audio clips
Module 9	Rating practices with real test takers
Module 10	Assessment

As can be seen from the table, the training provided raters-to-be with both theoretical background and practical knowledge on VSTEP speaking rating. Even though trainees were experienced in their teaching and highly qualified in terms of English proficiency, testing and assessment appeared to be a gap in their knowledge. Therefore, the program firstly focused on an overview of language testing and assessment, then the assurance to maintain the quality of rating activity, followed by theories of speaking assessment as the key goal of this course. Due to the fact that VSTEP.3-5 is based on the CEFR, there was no doubt that there should contain some modules about this framework with an attention to three levels namely B1, B2, C1 as these levels are assessed by VSTEP.3-5. Moving on VSTEP's part, trainees were introduced to the speaking test format and test procedure. The rating scales would be analyzed in great detail together with sample audios for analysis and practice. The emphasis of the training program in this phase was for rating scale analysis and audio clip practice. The last practice activity was with real test takers before trainees were assessed with both audio clip rating and real test taker rating.

A spotlight in this training program is that it is designed as a combination of the four training approaches mentioned in the Literature review. To be more specific, in module 2, rater quality assurance, rater trainees were familiarized with rating errors that are generally frequent to rater, which demonstrated for the RET approach. Regarding module 4 and 5, when the CEFR was put into a detailed discussion, the FORT and PDT approach were applied. That is to say, the trainees' judgment on VSTEP's test takers was guided to align with the CEFR as a standardized framework to assess language users' levels of proficiency. From distinguishing "can-do" statements across levels in the CEFR, especially CEFR descriptors for Grammar and Vocabulary, trainees were expected to make some initial judgments of their future test takers using the CEFR as a framework of reference. In module 7 and 8, the application of all four approaches was clearly seen. At the beginning of the rating activity, rater trainees focused on the rating scales as the standard descriptions for three assessed levels known as B1, B2 and C1. Based on the level description of all criteria, trainees did their marking on the real audio clips of previous tests. Thus, this is a combination of both accustoming the raters-to-be to the descriptors of each assessment criterion as a signal of applying PDT and helping the trainees visualize the rating criteria by analyzing the sample performances with agreed scores from the expert rater committee as a signal of applying FORT. At the same time, RET was also used when trainees had a chance to reflect their rating after each activity to see if they make any frequent errors. Besides, BOT aiming at training raters to become skilled observer who are able to recognize or recall the performance aspects consistent with the rating scales was emphasized during all modules related to VSTEP rating activity. To illustrate, trainees were reminded to take notes during their rating, hence the notes help them link the test taker's performance with the description in the rubric. In this case, observation and note-taking did play a substantial role in the VSTEP speaking rating. The integration of mixed approaches in rater training, therefore, has been proved in this program.

4.3. Data collection

The data collection was conducted based on a pre- and post- training comparison. 37 trainees were asked to rate 5 audio clips of speaking performance before Module 7 where an in-depth analysis of the rating scales was performed. At this stage, they knew about

the VSTEP.3-5 speaking test format and test procedure. They were also allowed to approach the rubric and work on the rubric on their own for a while. The 5-clip rating activity was conducted based on trainees' first understanding of the rating scales and their personal experience in speaking assessment. After a total of 20 hour on-site training in Module 7 and 8, the trainees involved in marking 10 clips including those 5 clips in random order. The reason why the initial 5 clips were embedded in the 10 later clips is that the participants were expected not to recognize the clips they had rated, which maintains the objectivity of the study. Rating the 10 clips is part of the practice session. The trainees' rating results were compared to those of an expert committee to check their accuracy. It is noteworthy to be aware that the clips used as research data were the recordings selected from practice interviews in previous training courses in which trainees were required to examine voluntary test takers. Both examiners and test takers in the interviews were anonymous, which guarantees the test security.

4.4. Data analysis

Multiple ways of analysis were exploited to examine the effectiveness of the rater training session. First of all, descriptive statistics of every rating criterion and of total scores were run. After that, traditional reliability analyses of exact and adjacent agreement, correlations, and Cronbach alpha were implemented. In order to further scrutinize the reliability, Generalizability theory was applied with the help of mGENOVA software. The approach of G theory utilized G study and D study to estimate variance component and dependability as well as generalizability of the speaking scores respectively. Finally, patterns of changes in rating quality were devolved into with many-facet Rasch analyses (FACETS software), in which how the criterion difficulty, rater severity, rater fit, rater bias, and score band separation was impacted after rater training session was unveiled.

5. Results

5.1. Descriptive statistics

Table 2: Descriptive statistics of speaking ratings before and after rater training (37 trainee raters, 5 test-takers, 1 test)

	Grammar		Vocabulary		Pronunciation		Fluency		Discourse management		Total score	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pre-training	4.93	2.038	4.84	2.084	5.04	2.028	4.79	2.180	4.89	2.169	24.49	10.222
Post-training	4.98	2.040	4.94	2.240	4.96	2.055	4.91	2.198	4.97	2.308	24.76	10.599
Committee score	5.00	2.121	5.00	2.449	5.20	1.924	5.20	2.775	4.80	2.387	25.20	11.563

In the first place, mean ratings for each speaking criterion were presented in Table 2, which showed that the raters' scores were lower than committee scores in all criteria except for discourse management. Although the differences were modest the post-training scores of most criteria and the total score were closer to the committee scores than the pre-training ones. To investigate further into changes in ratings after training, analyses of traditional reliability were conducted.

5.2. Traditional reliability analyses

First of all, with the acceptable score difference set at 4 (out of 50 in total), the exact and adjacent agreement between raters and the committee was calculated. The result revealed that there were 148 scores within the acceptable range in total of 185 scores prior to the training session, which means the agreement rate was at 80%. The rate increased to 86% for the post-training ratings (159 out of 185 scores ± 4 points apart from the committee ones).

Besides exact and adjacent agreement, the inter-rater correlations were also computed. There were 666 significant inter-rater Pearson correlations resulted from 37 raters in total ($p<.05$). The average inter-rater correlation was high at .962 in the pre-training session and higher at .966 in the post-training session.

Finally, regarding Cronbach alpha index, the reliability level rose slightly from .986 to .988 after the raters received the training.

It can be seen that raters are already consistent before the training but there still existed improvement. As the changes were slight and seemingly negligible, more robust analysis methods were in need to scrutinize the patterns of improvement in aspects other than traditional reliability. This was the reason why G-theory and many-facet Rasch model were utilized.

5.3. Generalizability theory analyses

With the help of G study, variance components of the speaking scores were revealed in Table 3.

Table 3: Variance components of the speaking scores ($p^\bullet$ x $r^\bullet$ model, 5 test-takers, 37 raters)

	Variance source	Grammar		Vocabulary		Pronunciation		Fluency		Discourse management	
Pre	p (test-takers)	4.31029	86.25%	4.7006	89.28%	4.24264	85.80%	5.34775	92.20%	5.15488	90.17%
	r (raters)	0.07875	1.58%	0.10841	2.06%	0.1045	2.11%	0.12162	2.10%	0.13056	2.28%
	pr, error	0.60863	12.18%	0.45616	8.66%	0.5979	12.09%	0.33063	5.70%	0.43161	7.55%
	Total	4.99767	100%	5.26517	100%	4.94504	100%	5.8	100%	5.71705	100%
Post	p (test-takers)	4.51607	89.44%	5.59369	91.50%	4.70495	91.49%	5.46486	92.63%	6.10105	93.58%
	r (raters)	0.13559	2.69%	0.07297	1.19%	0.02628	0.51%	0.04835	0.82%	0.03018	0.46%
	pr, error	0.39745	7.87%	0.44685	7.31%	0.41126	8.00%	0.38649	6.55%	0.38814	5.95%
	Total	5.04911	100%	6.11351	100%	5.14249	100%	5.8997	100%	6.51937	100%

As indicated in Table 3, there are totally 3 variance components in G study conducted with $p^\bullet$ x $r^\bullet$ design: test-takers (p), raters (r) and the interaction between test-takers and raters.

Observably, approximately 85% to more than 90% of the speaking score variance were substantially from the test-takers, that is, the difference in the speaking scores is mainly caused by the disparity in students' proficiency levels. In contrast, the small percentage of

variance coming from the main effect of rater variation source indicated that raters differed just slightly in their leniency/strictness and the difference was even narrowed from above 2% before training to negligible (1.19% or less) after training in four out of five criteria. In addition, it's noticeable that roughly 6% to 12% of the total variance in pre-training ratings was attributable to the variance component of the test-takers-raters interaction, which means the scores of test-takers varied to some extent across raters, especially for grammar and pronunciation. In the post-training round, this component explained less (6%-8%) of the total variance for all the criteria except for fluency. All these changes were the evidence for higher degree of consistency among raters after training.

Moreover, the D study also generated higher dependability and generalizability for the post-training ratings in all the criteria as well as the composite score (Table 4). Simply put, the ratings were more reliable after the training.

Table 4: Dependability and generalizability of the speaking scores ($p^{\bullet}$ x $r^{\bullet}$ model, 5 test-takers, 37 raters)

Criteria	Pre-training		Post-training	
	Φ (dependability)	$E\rho^2$ (generalizability)	Φ (dependability)	$E\rho^2$ (generalizability)
Grammar	0.99571	0.99620	0.99682	0.99763
Vocabulary	0.99676	0.99738	0.99749	0.99785
Pronunciation	0.99555	0.99621	0.99749	0.99764
Fluency	0.99772	0.99833	0.99785	0.99809
Discourse management	0.99706	0.99774	0.99815	0.99828
Composite (total score)	0.99778	0.99828	0.99858	0.99886

5.4. Many-facet Rasch analyses

Many-facet Rasch allowed the researchers to delve into the pattern of changes in different facets of the speaking test namely marking criteria, rater severity, rater misfit, rater bias, and band separation.

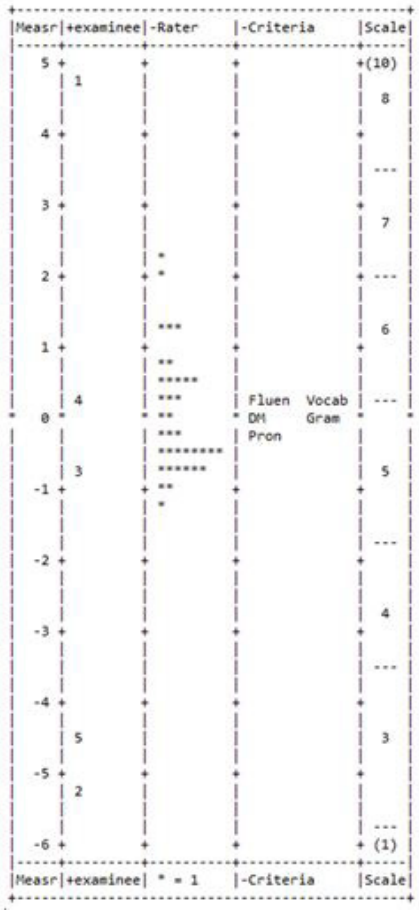
Regarding marking criteria, Figure 1 indicated that all five criteria of grammar, vocabulary, pronunciation, fluency, and discourse management gathered closely to each other on the difficulty scale, and even lined up after training. This enhancement emphasizes that all criteria were equally rated and no criterion was more difficult to fulfill than the others. The pronunciation criterion was shown to experience the most noticeable change when moving from the easiest position to the middle of the scale (Table 5).

When it comes to rater severity, Figure 1 showed that some raters became less severe and more raters clustered around the balanced point in the post-training ratings. This comes along with the decrease in the number of misfitting raters from eight to five out of 37 raters in total. These misfitting raters rated the test-takers' speaking performance differently from other raters, thus the infit mean square of their ratings was outside the desirable range from 0.5 to 1.5 (Table 6).

Table 5: Measures of the speaking criteria before and after training
(by ascending order of difficulty)

No.	Pre-training		Post-training	
	Criteria	Measure	Criteria	Measure
1 (easiest)	Pronunciation	-0.30	Grammar	-0.07
2	Grammar	-0.07	Discourse management	-0.04
3	Discourse management	0.01	Pronunciation	-0.03
4	Vocabulary	0.13	Vocabulary	0.04
5 (the most difficult)	Fluency	0.22	Fluency	0.10

Pre-training



Post-training

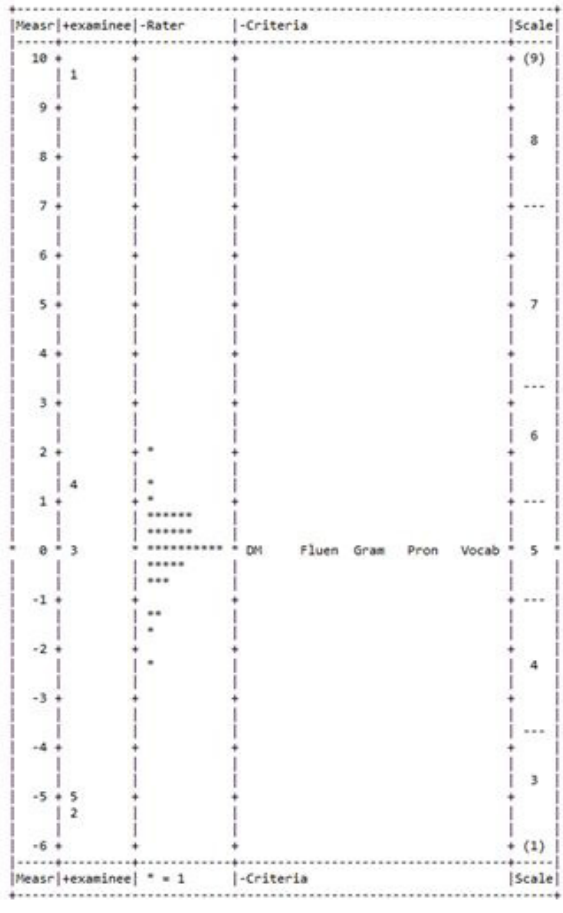


Figure 1: All facet summaries before and after training (5 test-takers, 37 raters)

Table 6: Rater fit indexes before and after training

Rater	Pre-training			Post-training		
	Measure	Infit		Measure	Infit	
		MnSq	ZStd		MnSq	ZStd
1	-1.16	0.67	-1.10	-0.03	0.70	-1.10
2	0.48	0.84	-0.40	1.11	0.68	-1.20
3	0.39	1.85	2.40	-0.13	1.43	1.50
4	-0.22	1.09	0.30	-0.03	0.93	-0.10
5	-0.65	0.92	-0.10	-0.13	1.06	0.20
6	1.17	0.53	-1.80	0.58	1.01	0.10
7	-0.57	1.10	0.40	-1.25	1.81	2.40
8	0.48	0.58	-1.50	0.07	1.05	0.20
9	-0.82	0.98	0.00	-0.33	0.89	-0.30
10	-0.22	1.03	0.10	-0.13	0.78	-0.70
11	-0.74	0.54	-1.70	-0.43	0.40	-2.80
12	-0.48	0.47	-2.10	-0.13	0.52	-2.00
13	-0.39	4.05	6.10	0.38	1.27	1.00
14	-0.39	0.64	-1.30	-0.23	0.94	-0.10
15	-0.57	0.40	-2.50	-0.03	0.63	-1.50
16	1.92	1.68	2.00	0.58	1.49	1.60
17	1.33	0.99	0.00	-0.33	0.71	-1.10
18	2.17	0.87	-0.30	-1.46	1.72	2.10
19	-0.48	1.34	1.10	-2.24	0.78	-0.60
20	0.65	0.65	-1.20	-1.67	0.63	-1.40
21	0.13	0.84	-0.40	2.17	0.69	-1.00
22	0.04	0.68	-1.10	0.58	0.63	-1.50
23	0.82	1.14	0.50	0.28	0.63	-1.50
24	0.56	0.65	-1.20	-0.53	0.99	0.00
25	1.33	1.42	1.30	-0.53	1.60	1.90
26	-0.99	0.54	-1.80	0.79	0.97	0.00
27	-0.82	0.97	0.00	-0.83	1.11	0.40
28	-0.13	1.43	1.30	-0.03	1.20	0.70
29	-0.57	1.24	0.80	0.38	0.65	-1.40
30	-0.57	0.46	-2.20	0.38	1.03	0.20
31	0.39	1.19	0.70	0.07	1.11	0.40
32	0.22	1.12	0.40	0.28	0.95	-0.10
33	-0.99	1.84	2.40	-0.23	1.08	0.30
34	-0.05	0.47	-2.10	0.18	1.82	2.50
35	0.30	0.96	0.00	0.69	0.70	-1.10
36	-0.82	0.55	-1.70	1.45	0.93	-0.10
37	-0.74	0.60	-1.40	0.69	0.58	-1.70

In addition, rater-occasion bias was also studied and 20 significant bias cases out of 74 bias terms were detected. Simply put, after training, ten out of 37 raters became significantly more lenient or severe than they had been in the pre-training phase.

Importantly, investigation in to score bands revealed better separation after training. Compared with Figure 2 which displayed that Band 4, 6, 7, and 8 were in overlap with other bands, Figure 3 showed a much more distinguishable separation for these central bands. Apparently, raters distinguish score bands considerably better after being trained. This is probably the biggest improvement in comparison with other aspects.

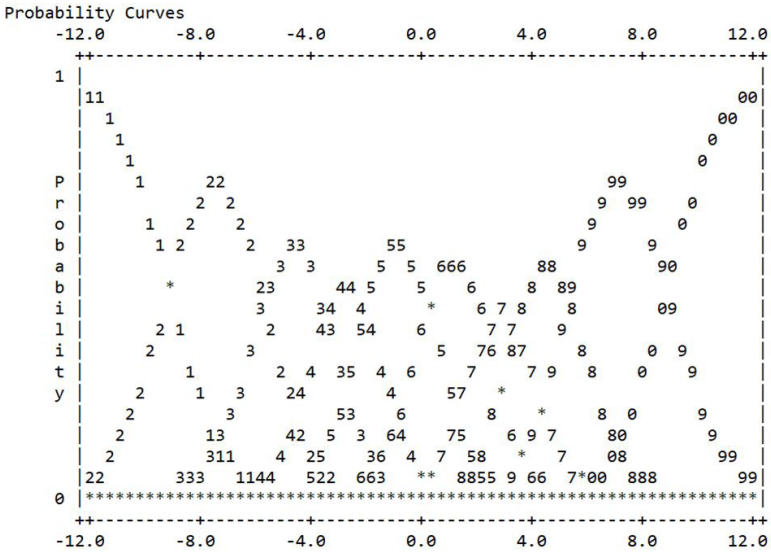


Figure 2: Score band separation before training

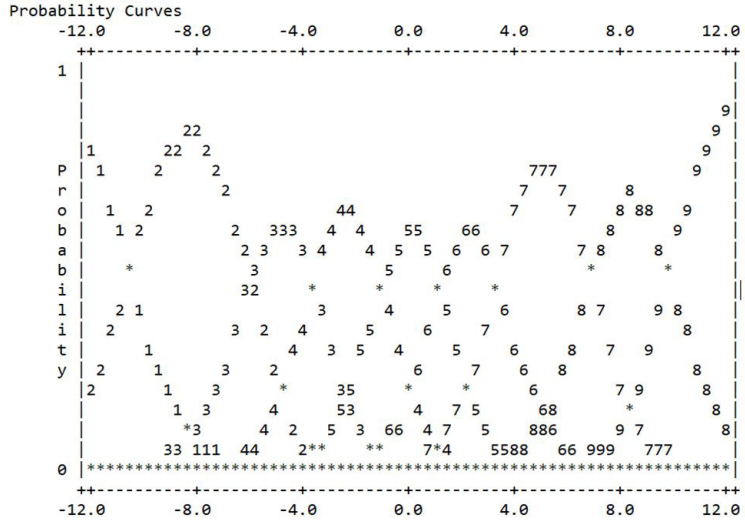


Figure 3: Score band separation afternoon training

6. Discussion

The study was implemented in order to find out the effectiveness of the rater training session on the VSTEP.3-5 rating scales. In general, the results have confirmed betterment in all the aspects examined.

With regard to the first research question, the reliability of the speaking scores analyzed both with traditional analyses and generalizability theory was showed to slightly increase after the training. In particular, higher values were recorded in exact and adjacent agreement among raters, inter-rater Pearson correlations, and Cronbach alpha reliability. The same is true for analysis with G-theory on the consistency among raters and the dependability and generalizability of the test scores.

Concerning the second research question, analyses with many-facet Rasch reported betterment in facets other than reliability: the balance in the difficulty of speaking criteria was enhanced, divergence in rater severity was lessened, rater fit was heightened, rater bias cases were fewer, and the score band separation was greater.

Generally, slight improvement was found for the majority of the aspects, and the most noticeable betterment was for the case of score band separation. Although the change was relatively small for some aspects of the speaking scores, it is still the evidence for the efficacy of the training session. Moreover, these positive changes can be considered important when taking other factors into consideration. Firstly, it is noteworthy in relation to the small number of training hours on the rating scales (20 hours). Furthermore, the pre-training rating session took place after Module 6, which means the raters already received a great amount of training on issues related to language assessment in general and speaking assessment and CEFR in particular. On top of that, research participants all had high-level qualifications and many years of experience in language teaching. All these factors likely helped the raters in shaping their ratings even before exposing to explicit guidance on the VSTEP.3-5 rating scale. Therefore, the impact is expected to be more visible and significant for either novice trainees or those yet to experience any training on standardized test scoring. This is also the researchers' suggestions for further research in the future.

Obviously, these above-presented results supported the point of researchers such as Smith (1986), Woehr and Huffcutt (1994), Noonan and Sulsky (2001), Roch et al. (2011), Rosales Sánchez, Díaz-Cabrera, and Hernández-Fernaud (2019), who had advocated and provided evidence for the enhancement of rating quality after rater training. Besides, the findings of small increase in score reliability was in line with several reports on slight improvement of rating accuracy by McIntyre, Smith, and Hassett (1984), Noonan and Sulsky (2001), Roch and O'Sullivan (2003). In addition to showing agreement with previous research, this study made meaningful contribution to literature in a way that it proved the effectiveness of a synthesized approach combining all four strategies of rater training and utilized various methods to statistically analyze the scores, both of which have not been widely documented in research so far. Moreover, it was this application of multiple statistical analyses that disclosed the noticeable enhancement in the score band separation of the speaking scores.

7. Conclusion

Overall, the study has rendered positive evidence for the efficacy of rater training, focusing on rating scale session with both guidance and practicing activities. After the training session, raters' ratings were found to be more reliable, consistent, and distinguishable.

Meaningful implications could be drawn from the study. Firstly, taking rater training administration into consideration, the combination of multiple/all training approaches is feasible and advisable. Although more research is needed to justify, the results suggest that if more combination is applied, greater impact is possible. This again not only restated the importance of rater training but also went beyond to emphasize the significance of how the rater training is conducted. Secondly, regarding methodological implication, the research showed that traditional statistical analyses through descriptives, Cronbach alpha, and correlations might not bring about sufficient information of the impact. In this case, the application of Generalizability theory and many-facet Rasch is recommended for better insights. Studies in the future should take this approach into consideration and expand to investigate the effectiveness of the whole rater training course for possible findings of significant changes.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*(1), 60–66. https://doi.org/10.1037/0021-9010.65.1.60

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric errors in ratings. *Journal of Applied Psychology*, *61*(1), 64-69.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Hakel, M. D. (1980). *An appraisal of performance appraisal: Sniping with a shotgun*. Paper presented at the First Annual Scientist-Practitioner Conference in Industrial-Organizational Psychology, Virginia Beach, VA.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology*, 73(1), 68-73.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, *60*(5), 550-555.

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

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

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*(1), 147–156.

McNamara, T. F. (1996). *Measuring second language performance*. Essex: Addison Wesley Longman.

Noonan, L. E., & Sulsky, L. M. (2009). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian Military Sample. *Human Performance*, *14*(1), 3-26.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*(4), 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational behavior and human decision processes*, *38*, 76-91.

Roch, S. G., O'Sullivan, B. J. (2003). Frame of reference rater training issues: recall, time, and behavior observation training. *International Journal of Training and Development*, *7*(2), 93-107.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2011). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*, 370-395.

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

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review*, *11*, 22-40.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology*, *65*(3), 351.

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

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Weir, C. J. (2020). Global, Local, or "Glocal": Alternative pathways in English language test provision. In L. I-W. Su, C. J. Weir, & J. R. W. Wu (Eds), *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*. New York: Routledge.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189-205.

HIỆU QUẢ CỦA HOẠT ĐỘNG TẬP HUẤN GIÁM KHẢO CHẤM NÓI VSTEP.3-5

Nguyễn Thị Ngọc Quỳnh, Nguyễn Thị Quỳnh Yến, Trần Thị Thu Hiền
Nguyễn Thị Phương Thảo, Bùi Thiện Sao, Nguyễn Thị Chi, Nguyễn Quỳnh Hoa

Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội
Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Tóm tắt: Giữ vai trò quan trọng trong việc đảm bảo độ tin cậy của hoạt động kiểm tra đánh giá các kỹ năng sản sinh ngôn ngữ, tập huấn giám khảo (rater training) là một chủ đề thu hút trong nghiên cứu về các bài thi quy mô lớn. Tương tự, với bài thi VSTEP, hiệu quả của chương trình tập huấn giám khảo cũng nhận được nhiều sự quan tâm. Do đó, một nghiên cứu đã được tiến hành nhằm tìm hiểu ảnh hưởng của phần tập huấn sử dụng thang chấm Nói VSTEP.3-5 với các giám khảo trong chương trình bồi dưỡng tổ chức bởi Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội. Dữ liệu được thu thập từ 37 học viên tham gia khóa tập huấn nhằm so sánh việc chấm điểm của các học viên trước và sau phần tập huấn sử dụng thang chấm Nói. Cụ thể, các khía cạnh về độ tin cậy của điểm số, độ khó của tiêu chí, độ khó tính, độ phù hợp, và độ thiên lệch của giám khảo cũng như mức phân tách của thang điểm đã được phân tích. Nghiên cứu đã thu được các kết quả tích cực khi điểm số của các giám khảo đưa ra sau phần tập huấn có độ tin cậy, thống nhất, và phân tách tốt hơn. Sự cải thiện rõ rệt nhất được tìm thấy ở khía cạnh độ phân biệt mức điểm trong thang chấm. Một số ý nghĩa về hoạt động tập huấn giám khảo cũng như phương pháp nghiên cứu hoạt động này đã được rút ra từ các kết quả nghiên cứu.

Từ khóa: tập huấn giám khảo, chấm Nói, kiểm tra đánh giá kỹ năng Nói, VSTEP, lý thuyết G (G theory), phân tích Rasch nhiều khía cạnh (many-facet Rasch)