PHILOLOGY

SPEAKING ASSESSMENT: IMPACT OF TRAINING SESSIONS

Mariana Burak, *PhD student*

Ukraine, Ternopil, Ternopil Volodymyr Hnatiuk National Pedagogical University

DOI: https://doi.org/ 10.31435/rsglobal_ws/30122018/6275

ARTICLE INFO

Received: 07 October 2018
Accepted: 08 December 2018
Published: 30 December 2018

KEYWORDS

Assessment,
speaking,
proficiency,
reliability,
accuracy,
efficiency.

ABSTRACT

The article focuses on the problem of examiner's objectivity in rating Speaking proficiency in a foreign language at standardized high-stakes tests. Since there are different factors which may impact the assessment reliability, special training sessions are widely usedby different testing centers. They are expected to eliminate examiners' subjectivity and lead to interrater agreement and intrarater consistency. The research described in the article was aimed at finding empirical evidence of the efficiency of such sessions. The outcomes of the study proved the sessions to be efficient in terms of rating accuracy of the examiners.

Citation: Mariana Burak. (2018) Speaking Assessment: Impact of Training Sessions. *World Science.* 12(40), Vol.2. doi: 10.31435/rsglobal_ws/30122018/6275

Introduction. Incorporation of communicative principles into language assessment made Speaking a mandatory subtest of most of the standardized high-stakes tests in English as a foreign language such as TOEFL, IELTS, Cambridge English tests etc. It typically has a form of an interview or a role play and is based predominantly on the examiner's judgment of the testees' performance, which may lack accuracy though.

Thus, raters' reliability in scoring Speaking tests has been a matter of deep concern in high-stakes language testing from the moment the method was introduced into language assessment. Since then several empirical studies have reported that the main factors of scoring reliability are intrarater and interrater agreement, which can be best achieved by training practices [1;2; 3; 5; 7; 8]. As Weigle states "rater variability cannot be eliminated, but extreme differences can be reduced" by means of training sessions [9, p. 269].

The aim of the present study was to conduct an independent research in order to find the empirical evidence of efficiency of the training sessions, as well as to answer the following questions under study:

1) In what way does the rater accuracy in scoring change with training?
2) In what way does the rater severity change with training?

Research results. The 40 participants of the research were language lecturers from Ternopil Volodymyr Hnatiuk National Pedagogical University and 3 other higher educational institutions. 5 of them were English native speakers (3 from the USA, 1 from the UK and 1 from Australia). In terms of the questions under study all the participants were considered inexperienced in standardized proficiency assessment, although most of them (34) had experience in teaching English and some were familiar with different testing formats of English proficiency.

FCE (a Cambridge English test for B2 CEFR level) was chosen as a format of a proficiency test to be practiced. "Cambridge English First. Handbook for teachers for exams from 2015" was used to introduce

sample papers, scoring scales and rubrics to the participants at the beginning of the training session. Cambridge Assessment FCE Oral Exam educational videos were used as initial stage samples for the training session. The training sessions were conducted by professional educators, assessors and teachers.

The 40 participants assessed 20 candidates each, 10 at a pre-training session and 10 at a post-training one. The responses were scored holistically by each rater individually in 5 domains (Grammar and Vocabulary, Discourse Management, Pronunciation, Interactive Communication and Global Achievement) and the scoring was adapted to a 10-point scale.

As a result, the 2-day event provided the study with 2 000 scores for the pre-training assessment session and 2 000 scores for the post-training session correspondingly. And this didn't include the experts' preliminary scoring, which was carried out according to the same specially created scoring scale (150 scores in total, 50 for the first session and 100 for the second). The expert scores later were used as reference scores and compared with the participants' ones to calculate the level of agreement prior to the training session and after it as well as to estimate the degree of a participant's severity or leniency and intrarater consistency.

On the first day of training the format of FCE and its assessment scale with rubrics were introduced to the participants so that they knew what to expect during the test and what to do. There was no further instruction on the methods of assessment of the exemplars at this stage. Thus, the scores received from the participants on the first day were later used as the pre-training results. On the second day, immediately after 8 hours of training on how to assess each of the criteria the participants assessed another group of 10 candidates and the obtained results were used as the post-training results.

The instruction consisted of 5 lectures on each criterion of the FCE scoring band (1) "English for Global Opportunities. Speaking Overview of the Most Widespread International Exams"; 2) "What's in a Tongue: Reflecting on Vocabulary and Grammar of Spoken English"; 3) "Being Ready for Everything. Exam Interviews"; 4) "Intelligible or is it? Teaching and Evaluating Pronunciation for B2 Level Exams"; 5) "We Need to Talk... Assessment Criteria for Interactive Communication"). The training sessions included discussions of the assessment of each criterion using the exemplars and were followed by a questions-answers session for further clarification of the rating procedure.

The obtained scores were brought together inspecial forms for each participant for both pre-training and post-training assessment.

Table 1. First day pre-training rater's scores in comparison with expert scores

		Rater's scores					Benchmark scores				
Candidate	**Rater**	Grammar& Vocabulary	Discourse Management	Pronunciation	Interactive Communication	Global Achievement	Grammar& Vocabulary	Discourse Management	Pronunciation	Interactive Communication	Global Achievement
Raphael	1	4.5	4.5	4	4	4.5	5	5	5	5	5
Maude	1	5	4.5	5	4	4.5	4.5	5	5	5	5
Victoria	1	3.5	4	5	3	3.5	4.5	4.5	5	5	4.5
Edward	1	3	3.5	3.5	3.5	3.5	3	3.5	3	3.5	3.5
Florine	1	3.5	3.5	4.5	5	4	5	5	5	5	5
Maria	1	2.5	3	4	2.5	3.5	3	3.5	3	3.5	3.5
Paolo	1	3	3.5	3	3.5	4	4	4.5	4.5	5	4.5
Natalie	1	3.5	4.5	3.5	3.5	3.5	4.5	4.5	4	5	4.5
Ottavia	1	3	3.5	3.5	3	3.5	3.5	3.5	3.5	4	3.5
Hannah	1	3	3	4	3.5	3.5	3.5	4	4	4	4

Later a discrepancy score for each criterion and each exemplar was calculated by comparing the scores of the raters with the benchmark.

Table 2. First day pre-training discrepancy scores

Candidate	Rater	Grammar& Vocabulary discrepancy score	Discourse Management discrepancy score	Pronunciation discrepancy score	Interactive Communication discrepancy score	Global Achievement discrepancy score	Total discrepancy score
Raphael	1	-0.5	-0.5	-1	-1	-0.5	-3.5
Maude	1	0.5	-0.5	0	-1	-0.5	-1.5
Victoria	1	-1	-0.5	0	-2	-1	-4.5
Edward	1	0	0	0.5	0	0	0.5
Florine	1	-1.5	-1.5	-0.5	0	-1	-4.5
Maria	1	-0.5	-0.5	1	-1	0	-1
Paolo	1	-1	-1	-1.5	-1.5	-0.5	-5.5
Natalie	1	-1	0	-0.5	-1.5	-1	-4
Ottavia	1	-0.5	0	0	-1	0	-1.5
Hannah	1	-0.5	-1	0	-0.5	-0.5	-2.5
				Total discrepancy score for all the assessments			**-28**

As a discrepancy score does not show the correlation between the scores of all the participants and their distribution around mean, z score for each participant was calculated to be further used in analysis. The positive scores show that the rater's score is more lenient than that of an expert and on the contrary a negative z score indicates the examiner's higher than expected severity.

Table 3. Discrepancy scores and their transformation into z scores

Rater	**Stage 1**		**Stage 2**	
	Discrepancy score	**Z score**	**Discrepancy score**	**Z score**
1	2,5	1,65	-29,5	0,46
2	-47	-0,39	-26	0,68
3	-53,5	-0,66	-41	-0,25
4	-66,5	-1,2	-58,5	-1,34
5	-25,5	0,49	-26	0,68
6	-8,5	1,19	-44	-0,44
7	-106,5	-2,85	-55,5	-1,16
8	-7,5	1,23	-55,5	-1,16
9	-30,5	0,29	-44	-0,44
10	-30,5	0,29	-13,5	1,46
11	-21	0,68	-19,5	1,08
12	-53	-0,64	-34,5	0,15
13	-64,5	-1,12	-22	0,93
14	-43	-0,23	-74	-2,31
15	-24	0,55	-21,5	0,96
16	-43	-0,23	-74,5	-2,34
17	-70,5	-1,36	-47	-0,63
18	-50	-0,52	-39,5	-0,16
19	-6,5	1,28	-27,5	0,59
20	-25	0,51	-23	0,87
21	-44	-0,27	-37	0
22	-67,5	-1,24	-64	-1,69
23	-35	0,1	-41,5	-0,28
24	0	1,54	-25,5	0,71
25	-35,5	0,08	-28,5	0,52
26	-45	-0,31	-56	-1,19
27	-14	0,97	-29	0,49
28	-32,5	0,2	-34,5	0,15
29	16,5	2,22	-29,5	0,46
30	-45	-0,31	-53	-1
31	-6	1,3	-18	1,18
32	-78,5	-1,69	-25	0,74
33	-34,5	0,12	-21,5	0,96
34	-36	0,06	-8,5	1,77
35	-48	-0,44	-47	-0,63
36	-56	-0,77	-18,5	1,15
37	-27	0,43	-32,5	0,28
38	-26,5	0,45	-40	-0,19
39	-43,5	-0,25	-59,5	-1,41
40	-65,5	-1,16	-31,5	0,34

The comparative analysis of the results before training and after it shows thatdiscrepancy z score of most of the raters improved. Strangely, however, the highest degree of leniency expressed by z score of 2.22 by rater No 29 though being the most "subjective" positive score at the first stage changed to 0,46 at the second stage, signifying the highest degree of efficiency of the training for the rater. The most severe score of rater No.7 at the first stage remained much different from the benchmark at the second stage but as a result it became twice more lenient and thus showed a significant positive dynamic. Some raters, such as No.4, 22, 26, 39 showed the tendency of becoming even harsher after the training and No.14 and No.16 strangely increased their severity by 10 times. As for No.6 and 8, on becoming harsher and having changed their discrepancy score from positive into negative, they still made a positive dynamic toward the benchmark. And on the contrary, examiners No.10, 11 and 34 developed even higher level of leniency and No. 36 changed the score from negative to positive so tremendously that the leniency score even acceded the previous severity score. However, No. 13, 32 and 40 having changed their results from negative to positive approached the benchmark in their progress.

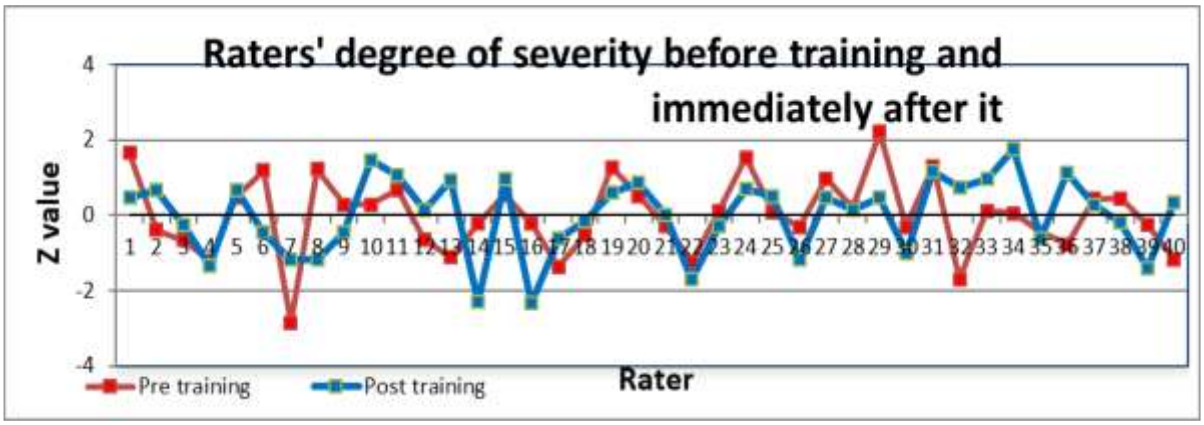

Fig. 1. Rater's severity and leniency before and after training

Table 4 illustrates the progress or regression of the raters whose score at either of the stages was above 1 or below 1.

Table 4. Rater's severity and leniency before and after training

	Stage 1	Progress	Stage 2
1	1,65	+	0,46
4	-1,2	-	-1,34
6	1,19	+	-0,44
7	-2,85	+	-1,16
8	1,23	+	-1,16
10	0,29	-	1,46
11	0,68	-	1,08
13	-1,12	+	0,93
14	-0,23	-	-2,31
16	-0,23	-	-2,34
17	-1,36	+	-0,63
19	1,28	+	0,59
22	-1,24	-	-1,69
24	1,54	+	0,71
26	-0,31	-	-1,19
29	2,22	+	0,46
31	1,3	+	1,18
32	-1,69	+	0,74
34	0,06	-	1,77
36	-0,77	-	1,15
39	-0,25	-	-1,41
40	-1,16	+	0,34

The analysis of the data leads us to the conclusion that for most of the raters the training session was efficient and resulted in the raters' score approaching the benchmark. Out of 40 raters only 10 showed negative changes in their assessment. Thus, the efficiency of the training session can be rated at 75%.

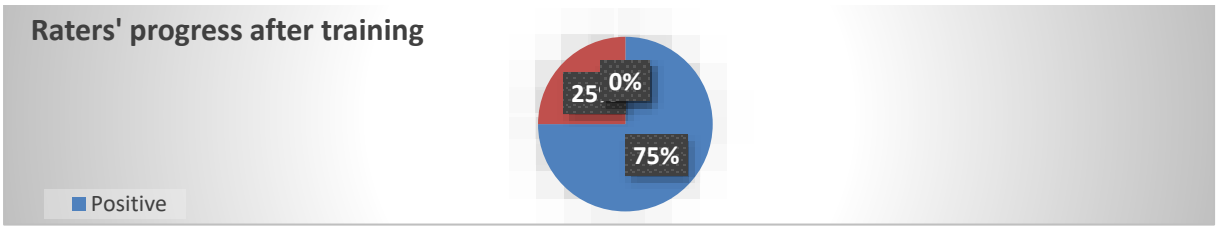

Fig. 2. Raters' progress after training

The study was also accompanied by 2 background Google Forms surveys. One was intended to collect personal data on participants and the other one was a follow-up survey used to collect the feedbacks of the participants about their possible progress after the training. Interestingly the "subjective" self-assessment of the participants'progress and the outcomes of statistical calculations appeared to be almost identical.

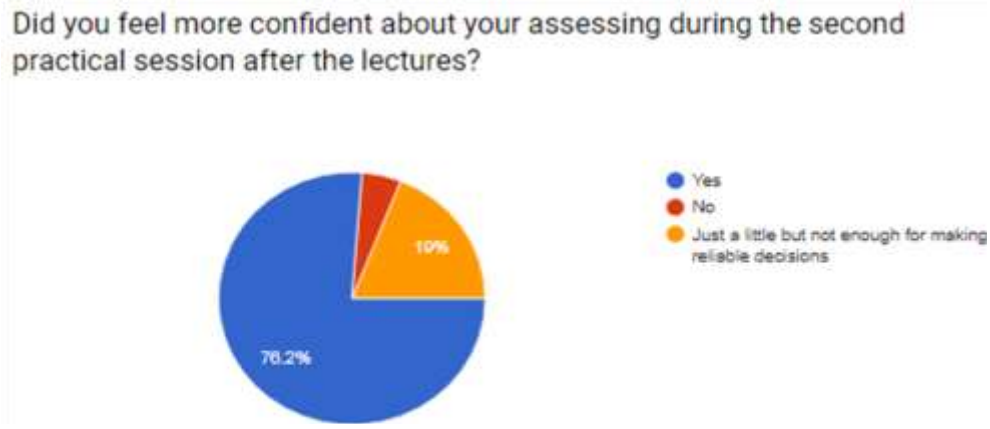

Fig. 3. Self-assessment data of the participants' progress

The findings of the study reported above show that though the examiners' biasedness in Speaking assessment was not absolutely eliminated, most of the raters became more accurate in their scoring. It proves the idea that although causal connections between attitudes and outcomes cannot be proved, it may be assumed that if training causes more user's satisfaction it is more effective, "leading to greater compliance with the benchmarks and hence increased conformity of rater behavior" [4; p.57].

The slight differences of the raters' severity, which cannot be eliminated, can thus 'be modelled in MFRA to some extent and the reduction of the variability in raters' severity should not be the main purpose of rater training" [6; p.4].

Conclusions. In terms of personal attitude to the possible progress in rating the participating students with little or no confidence in their progress showed less improvement than those who felt more confident about the benefit of the training. As for the level of the examiners' severity and leniency it changed in most cases but differently as some of the raters became harsher while others became more lenient. However, 75% of the participants of the training sessions made their progress toward the benchmark score which signifies anoticeable increase in their objectivity.

REFERENCES

1. Attali Y. A Comparison of Newly-Trained and Experienced Raters on a Standardized Writing Assessment. // Language Testing,33(1), 2016. Pp. 99–115.
2. Barrett S. The impact of training on rater variability. // International Education Journal,2(1), 2001. Pp. 49–58.
3. Davis L. The influence of training and experience on rater performance in scoring spoken language. // Language Testing,33(1), 2016. Pp. 117–135.
4. Elder C. Evaluating Rater Responses to An Online Training Program for L2 Writing Assessment. / Elder C., Barkhuizen G., Knoch U., von Randow J. // Language Testing, 24(1), 2007. Pp. 37-64.
5. Fahim M. The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment./ Fahim M.,Bijani H.// Iranian Journal of Language Testing. Vol. 1, No. 1, 2011. Pp.1-16.
6. Kondo Y. Examination of Rater Training Effect and Rater Eligibility in L2 Performance Assessment. // Pan-Pacific Association of Applied Linguistics 14 (2), 2010. Pp. 1-23.
7. Pufpaff L. A. The Effects of Rater Training on Inter-Rater Agreement / L.A.Pufpaff, L.Clarke, R.E.Jones // Mid-Western Educational Researcher.2015. Volume 27, Issue 2 117. Pp.117-141
8. Stansfield C.W., Kenyon D.M. Development and validation of the Hausa Speaking Test with the ACTFL Proficiency Guidelines. // Issues in Applied Linguistics, 4, 1993. Pp. 5-31.
9. Weigle S. C. Using FACETS to model rater training effects. // Language Testing, 15, 1998. Pp.263-268.