**Norming speaking assessment**

*Inga Farte-Zalite, Latvian NAF Language School*

*Marju Laurits, Estonian Military Academy*

*Birgitte Grande, Norwegian Defence University College*

*Gabriela Repova, The Language Institute of Slovak Armed Forces*

The presentation gives an overview of the speaking rating norming activity created by an international team of BILC experts and the Latvian STANAG Testing team, which was carried out remotely over the summer involving the volunteers of the BILC testing community. The presentation is divided into 3 parts - the 1st part explains the rationale behind using extracts rather than full BAT2 interviews as well as the process of creating a set of extracts used for the norming activity (slides 5-8), the 2nd reports and analyses the results of it (slides 9-18), and the 3rd focuses on some extracts that caused most disagreement in ratings (slides 19-29).

The norming was based on 11 extracts from BAT2 interviews ranging from Level 1 to 3 plus levels included. It was organised remotely via a Google form where the raters were asked to listen to the extracts and a) determine the task and the targeted STANAG level; b) determine the level of performance; and c) provide their comments on them if they wished.

It is important to ensure that the raters of productive skills have consensus and are consistent in their ratings. The raters should follow the rating protocol and be fair to the test takers. Knoch, Fairbairn and Yan (2021) discuss the following rater effects that may impact the ratings: leniency and harshness, consistency, halo effect, central tendency or restriction of range, bias and use of scoring criteria. Knoch et al (2021) present a number of indices and methods that can be used to investigate rater effects (see *Table 2.5. Rater quality measures - ability to detect rater effect* in Knoch et al, 2021).

In total, 39 raters from NATO and partner countries took part in the study. The analysis of the ratings submitted for this activity showed that in some cases there was larger than desirable disagreement between the participants, and disagreement between the experts’ and participant’ ratings. The results are presented in Table 1 (slide 12 of the presentation), where the blue cells indicate the level assigned to the extract by the expert raters. When discussing the results, it is important to keep in mind that the number of performances the participants rated was rather small (11 extracts) and several of the performances were at level 2. Both of these aspects influenced the results.

Data analysis was carried out using free programs ReCal OIR (Freelon, 2013), Minifac (Linacre, 2023), jamovi (The jamovi project, 2023) and modules on the website [www.langtest.jp](http://www.langtest.jp/) (Mizumoto, 2015).

The following consensus indices were calculated: pairwise agreement between the experts’ rating and the participants’ rating. The level of agreement ranged from 45.5% to 100%. Knoch et al suggest that the expected level of agreement should be around 80%. 15 raters out of 39 met that criteria.

Also, the pairwise Cohen’s kappa was calculated between the experts’ rating and the participants’ rating. This index was used because it takes into consideration agreements that might happen by chance. The results ranged from 0.29 to 1.00. This means that in the case of some of the participants there was minimal or weak agreement. For more information about the kappa results, please look at slide 15.

Consistency of the ratings was investigated by looking at the Spearman correlation between the experts’ rating and the participants’ rating (slide 16). The correlations ranged from 0.84 to 1.00. This means that the raters quite consistently ranked the performances in the same order.

Finally, Many-Facet Rasch Measurement model was used to look at the possible rater effects. The analysis was carried using Minifac, the free version of FACETS (Linacre, 2023). The results show that the test takers mostly grouped together based on the levels assigned to them by the expert group. There was some difference in the leniency the raters employed. The infit values identified the raters who were misfitting and they are marked with an orange box on the Wright map (slide 18). The experts’ rating is marked with a red box. The raters are somewhat spread out on the logit scale in terms of their leniency or harshness of ratings. However, the rater severity is 1.56 which seems to indicate that there is not too much difference between the raters when it comes to leniency and harshness.

Finally, two speaking-test extracts that had caused disagreement between raters were played and commented on, and a couple of general points are worth drawing attention to. First, deciding the level of the test task revealed some difference of opinion between raters. For example, a task that asked the candidate to report on current events (level 2) was understood by some raters to be a “state and support opinion”-task (level 3). Uncertainty about what level (in other words, which linguistic functions) the task is targeting, could have implications for rating. Second, despite having correctly determined the level of a task, e.g. giving a past narration (level 2), there were testers who went on to rate the corresponding speech extract L2+. Since level 2+ is an unsustained  level 3 performance, it is not really possible to be rated L2+ at a level 2 task. In that context, eliciting L3 language has not been targeted, hence not tested. What interlocutors might do if they sense a test taker is probably more proficient than L2, is to probe for level 3 language production sooner.

References:

Eckes, T. (2011). Introduction to Many-Facet Rasch Measurement. Bern, Switzerland: Peter Lang D.

Freelon, D. (2010). ReCal: Intercoder reliability calculation as a web service. International Journal of Internet Science, 5(1), 20-33. Website:<http://dfreelon.org/utils/recalfront/>

Knoch, U., Fairbairn, J. & Yan, J. (2021). Scoring Second Language Spoken and Written Performance. Issues, Options and Directions. Equinox Publishing.

Linacre, J. M. (2023) Minifac: Evaluation, Student and Demonstration (Demo) Version of FACETS. Beaverton, Oregon: Winsteps.com. Website:<https://www.winsteps.com/minifac.htm>

McNamara, T., Knoch, U., & Fan, J. (2019). Fairness, Justice, and Language Assessment: The Role of Measurement. Oxford University Press.

Mizumoto, A. (2015). Langtest (Version 1.0) [Web application]. Retrieved from [http://langtest.jp](http://langtest.jp/)

Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 1. Journal of Applied Measurement, 4(4), 386–422.

The jamovi project (2023). jamovi (Version 2.3) [Computer Software]. Retrieved from [https://www.jamovi.org](https://www.jamovi.org/)