Comparing the Rating Effectiveness of Personalized vs. Non-personalized Feedback to On-line Raters of English Speaking and Writing Assessment

Alex Volkov [1] Kristina Chang [1] Jake E. Stone [1] Michelle Y. Chen [1, 2] Amery D. Wu [2]

1. Paragon Testing Enterprises 2. University of British Columbia

A UBC Subsidiary

Executive Summary

- Rater training and calibration in operational settings are actively studied in an on-line rating context.
- Ongoing personalized feedback presented here is a tool to keep active raters more in line with rating benchmarks. The automated method can be considerably less costly than more conventional techniques.
- The results show that the proposed method is effective, though not uniformly stable.

Introduction

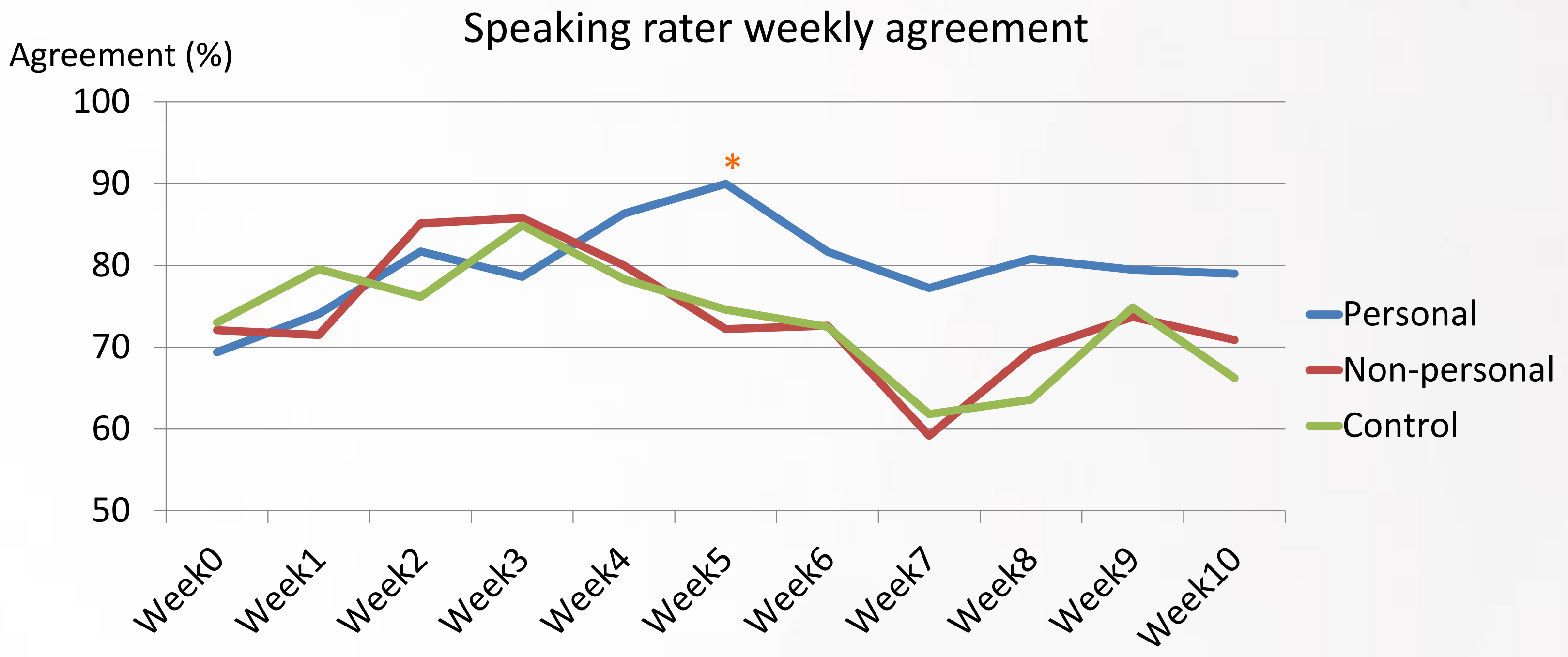
- Elder et al. (2007) state "there has thus far been little research into rater attitudes to [online] training" (p.37). Even though there is plenty of research on initial rater training, methods for ongoing feedback and calibration are not sufficiently studied.
- Cash et al. (2012) suggest that "the level of training required to establish acceptable inter-rater reliability … can require intensive resources in terms of time and money" (p.530). The ever increasing popularity of on-line rating can open opportunities for automated and therefore cheaper mechanisms.
- After examining a number of rater training methods, Woehr and Huffcutt (1994) state that "raters trained to evaluate performance using the same standards as 'expert raters' will produce ratings more like the 'expert ratings'" (p.200).
- The current study suggests an on-line feedback method that allows raters to compare their scoring with the expert raters' scores.
- The Canadian English Language Proficiency Index Program – General (CELPIP-G) Test measures functional English language proficiency in four domains (listening, reading, speaking, and writing) using a computer-administered format.
- The CELPIP-G is high-stakes, as its scores are used to demonstrate proficiency in English for Canadian citizenship and immigration applications.

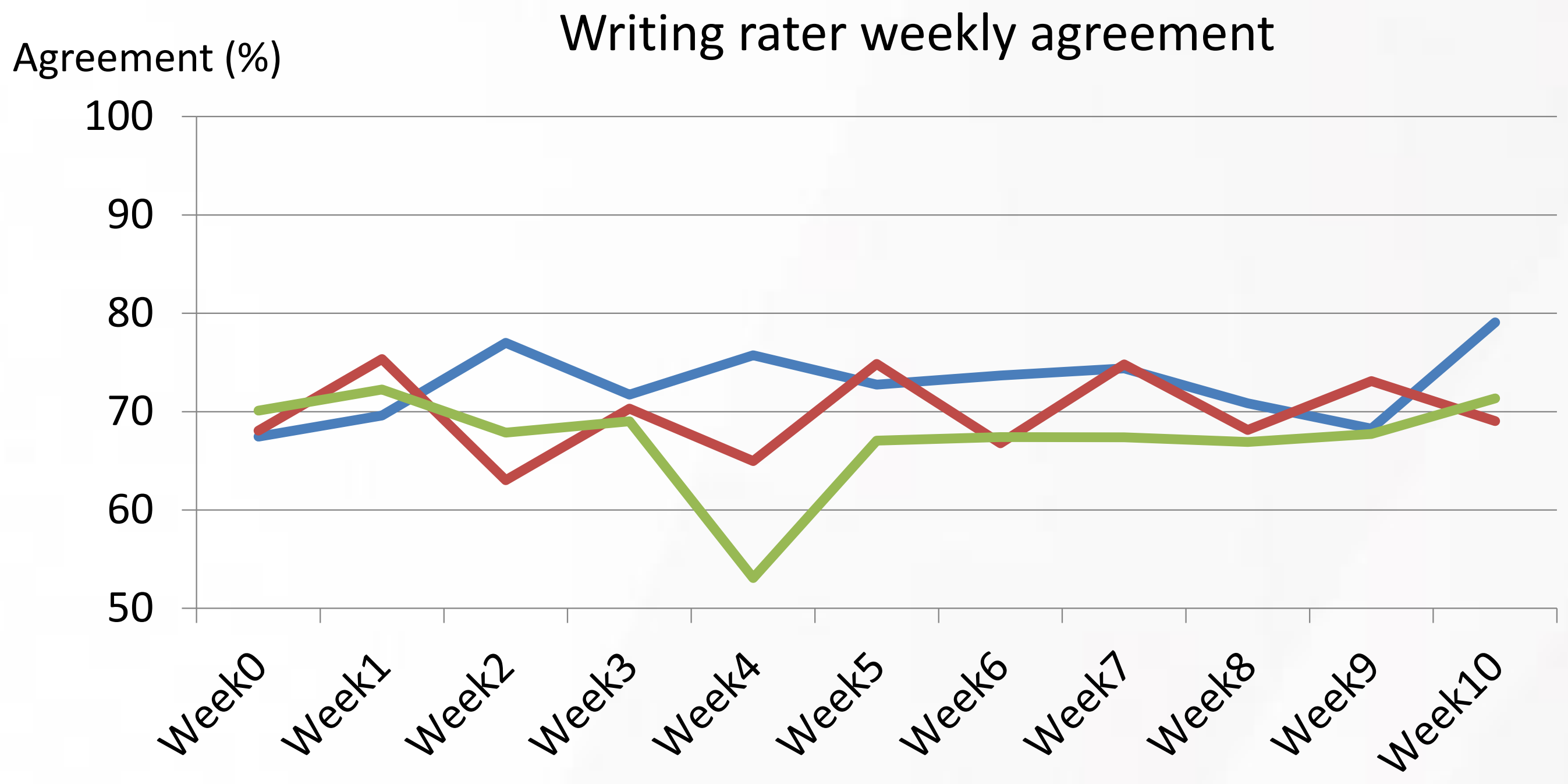
Method

- Based on all the assignment completed by the raters during 1 month, we selected 15 speaking and 15 writing raters who were underperforming.
- The raters were identified based on exact agreement, and exact and adjacent agreement on a 12-point scale.
- All the selected raters had rated at least 200 responses across 8 sessions for CELPIP-G to ensure that the change in rater performance is not largely attributed to the growing exposure.
- These underperforming raters were randomly divided into three groups. Each group had 5 speaking raters and 5 writing raters.
- Three different kinds of "treatment" were randomly assigned to each of the groups: (1) One group received personalized feedback; (2) one group received non-personalized feedback; (3) and the last group was used as control, and didn't receive any feedback.
- Personalized feedback: Speaking raters were given 20 short responses (40 seconds each) and writing raters were given four 200-word tasks they had personally rated. The raters could see the responses and directly compare their judgment with the assessment provided by the benchmark raters.
- Non-personalized feedback: Same amount of feedback was given to this group. However, raters only received benchmark raters' ratings, and they did not see any comparison with their own scoring.
- Control (no feedback): Did not receive any feedback.
- The feedback was given to raters for 8 consequent weeks. Rater agreement statistics were collected for each week.

Results

- Exact and adjacent agreement (1.5 points) on a 12-point scale was collected over 11 weeks: 1 pre-intervention week, 8 intervention weeks, 2 post-intervention weeks.





Note: * indicates the statistical significance at p < .05 level

Figure 1. Change of Raters' Weekly Agreement Over 11 weeks

The figure shows that

- 1) raters' performance in terms of their agreement with other raters is not stable across time, which suggests that rater performance needs to be continuously monitored;
- 2) for speaking raters, the agreement level of the personalized feedback group has improved, but it takes time (4-5 weeks after starting providing feedback) to be observed;
- 3) the effectiveness of the personalized feedback for speaking raters doesn't last very long;
- 4) Interestingly, neither the personalized feedback nor the non-personalized feedback improved the writing raters' agreement. One possible explanation is that speaking test fosters a stronger engagement with the responses.

Conclusions

- The feedback shows limited gains in agreement, though the results are not consistent across skills (speaking and writing rating).
- To improve the rater agreement of underperforming speaking raters, the proposed feedback method can be used as a low cost ongoing calibration.

Limitations

- The main limitation of this study is that the sample size is small, which largely limited the power of hypothesis testing and the generalization of the results.
- The number of assignments each rater got fluctuated from week to week, so the accuracy of the rater agreement estimates may differ across time.
- We did not control how much time each rater actually spent analyzing the samples. The level of personal engagement with the feedback can be an important factor in the effectiveness of feedback.

References

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: does it work? *Language Assessment Quarterly, 2*(3), 175–196.

Cash, A.H., Hamre, B.K., Pianta, R.C., & Myers, S.S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542.

Woehr, D.J., Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology ,67*(3), 189-205.