IRR Training Protocol for Course-Embedded Assessments

Instructional faculty in data collection point courses attend assessment-specific IRR training. IRR training sessions are videotaped and archived. The training is reviewed by new faculty as part of the onboarding process; those instructors submit scores for the final step of the training, which are tracked for reliability purposes. Faculty who do not meet minimum competency retrain through the online module. In addition, faculty have the opportunity to review training modules as refreshers.

The following is the basic set of guidelines for IRR training sessions used to promote objective and consistent evaluation of candidate performance.

1. **Provide a process overview to give the observers the big picture.** Participants are introduced to reliability, what it means to have a reliable measure and why it is important. Associated assignments are reviewed. In addition, participants discuss how rubric results are used to provide feedback to students and instructors. For students, these ratings serve as indicators of performance. For instructors, the results from the training sessions are used as part of the continuous improvement process to actively track the reliability of their ratings.
2. **Explain the rating dimensions (standards of performance & rubrics).** This is to help the raters become (more) familiar with the rubrics. Each dimension on the rubric is defined, and what each rating represents and means is explained. Participants discuss how they interpret the ratings; it is emphasized that if each participant has distinct understandings and interpretations, it becomes difficult to evaluate how consistent ratings are between participants.
3. **Help raters put aside biases.** Common biases are identified, and strategies to support objective ratings are discussed.
4. **Explain common rater errors to avoid.** We discuss common rater errors such as leniency rating, confirmation bias, similarity ratings, and halo, and participants discuss strategies to work through them.
5. **Describe the process for decision-making.** Raters develop objective strategies for making ratings, including note taking, sticking to the rating scale, and rating on each dimension separately.
6. **Participants practice observing and recording evidence/connecting evidence to performance dimensions, with feedback provided.** Raters pull out evidence from a provided example, linking their observations to performance. The raters discuss their findings and receive feedback.
7. **Participants practice interpreting the rubrics.** For this step, raters work through the language presented in the rubric rating scales. Words will be explored to better understand how different raters interpret the words selected in the ratings. Prototype examples are shared, which raters use for practice. Discussion follows this step.
8. **Rater training will be concluded with independent ratings on common examples, which are documented for reliability via adjacent agreement**.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Retrieved from http://www.tifcommunity.org

Halgren, KA (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(10), 23-34.