Sundqvist \& Sandlund (2024): Testing Talk: Ways to Assess Second Language Oral Proficiency. Bloomsbury Academic

## RATER FEEDBACK - TEMPLATE

## [Logo]

## [Project name]

Department name DATE

## Dear [rater name],

[Introductory paragraph referring to the rater training that has taken place]
[Explanation about the type of feedback offered, both at the group and individual level]
[Closing comments, next steps etc.]
Researcher name(s)
[Project name]
[Affiliation, contact details etc.]

## RATER FEEDBACK - TEMPLATE

## Individual feedback

Rater code: $\qquad$

## Introduction

[Comment on the results at group level here]
[Message about the first summary of results here; we provide an English translation (in italics below) of what we used ourselves in the EquA Project, to give an idea of what it can look like. For the measurements used in our project, see Sundqvist, Skar, Sandlund, \& Tengberg, 2020]

Here is your first summary of results. In Table 1, you can see your own assessments per student and recording, both on the sub-aspects of Content and Language and Expression, and on the entire test. In the rightmost column, you can see the benchmark grades for all students from the Swedish National Agency for Education.

## RATER FEEDBACK - TEMPLATE

## MEASURE 1:

Table 1. Assessments in comparison with benchmarks for Rater $\qquad$

| Rater | Pupil | Recording | Session during rater training day | Pupil in test | Assessment factors |  | Test grade | Benchmark <br> Swedish <br> National <br> Agency for Education |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  | Content | Language and expression |  |  |
| (Rater code) | 1 | Test 1 | 1 | Pupil 1 | D | D | D | C |
| (Rater code) | 2 |  |  | Pupil 2 | C | B | C | B |
| (Rater code) | 3 | Test 2 | 2 | Pupil 1 | C | C | C | D |
| (Rater code) | 4 |  |  | Pupil 2 | E | E | E | E |
| (Rater code) | 5 | Test 3 |  | Pupil 1 | C | B | C | B |
| (Rater code) | 6 |  |  | Pupil 2 | A | B | B | A |
| (Rater code) | 7 | Test 4 | 3 | Pupil 1 | E | E | E | E |
| (Rater code) | 8 |  |  | Pupil 2 | E | D | D | D |
| (Rater code) | 9 | Test 5 |  | Pupil 1 | B | C | C | B |
| (Rater code) | 10 |  |  | Pupil 2 | B | B | B | C |

Note. The assessment data are authentic (data used in Sundqvist et al., 2020).
[Write a message to the rater about what to reflect upon, based upon the data provided in a table or a chart, etc. Again, we provide an English translation (in italics below) of what we used ourselves in the EquA Project, to give an idea of what the message can look like.]

In relation to your Table 1, we suggest that you reflect on the following:

- What do you see when you compare your own assessments to the benchmarks from the Swedish National Agency for Education? Does the picture you get align with your own perception/experiences of the material from the agency and your own assessment of student tests?


## RATER FEEDBACK - TEMPLATE

- If the goal is equity in the assessment of oral English, what (if anything) should you consider moving forward?

The next step is a measure we call relative generosity. We arrive at the relative generosity measure by looking at your mean score of the ten student performances (the test grade) and compare it with the group's mean score. It is a 6 -grade scale ( $F=1, E=2, D=3, C=4, B=5$, and $A=6$ ). In the chart that follows below all raters are sorted from "the most severe" (the leftmost bar) to "the most lenient" (or "the most generous") (the rightmost bar). Your own bar is colored blue.


Reflect on how you view your results. How do you compare yourself to the average (the mean)? Does this align with how lenient/strict you thought you were? What do you think could explain the results?

## RATER FEEDBACK - TEMPLATE

In the two charts that follow, you can compare your own scale usage with that of the group. In the group's chart, you can see the number of grades assigned by your teacher group per grade level ( $A-F)$. You can then compare your own distribution across the grading scale with that of the group and reflect on how it looks. Think again about how you view your results/scores. Is there a scale step that you clearly use to a greater extent than the group? And is there a scale step that you clearly use to a lesser extent than your colleagues in the group? What do you think can explain the results?


Below you can take notes about your reflections. Please bring your notes to our Training Day on [date].

