EALTA Summer School, Innsbruck, 2016

Rating the construct reliably

Carol Spoettl, Claudia Harsch,
Nivja de Jong, and Jayanti Banerjee

Rater training

Jay and Carol

Session Aims

- To enable participants to
  - Consider the impact of decisions taken in rater training sessions on the construct to be measured
  - Provide informed advice
    - Colleagues at institutional level
    - Political players at national or local level
  - Run rater training sessions

The issues

- Ethics
  - Obtain permission
    - to gather performances
      - Heads of Unit, School Inspectorate
    - to use sample performances
      - in advance (i.e. at the data gathering stage)
      - design a consent form including the local legal requirements
- Changing mind sets
  - The old and the new

What do you need 1

- Introduction ppp on the construct to be measured
  - Establish the theoretical underpinnings first
  - Discussions will always need to be resolved by referring to this
    - It should never appear as your personal opinion
    - Your task is to deepen their understanding of the standard to be measured
- Familiarization materials
  - CEFR based exam? see Manual 2009
  - Commercial materials?
  - Locally designed materials?

What do you need 2

- A timetable and your idea of timing
  - How many written performances will you need?
  - How many speaking performances will you need?
  - How expert is your audience? (The more expert the group often the more discussion time is needed, the more novice the group the more explanation time is needed)
  - How much discussion time will you need to allow?
    - Training sessions often require more time allocated to understanding the procedures and the standard to be measured
    - Benchmarking sessions often require more time for discussion of the performances
- Performances
- A set of rater numbers
- Laminated copies of rating scales
- Rating sheets
  - Reflecting the skill and scale
  - With room to provide justifications
- An excel table to record judgements

What do you need 3

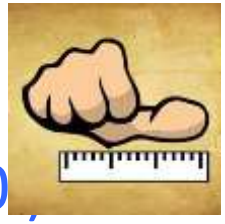
Performances

- Speaking rating is faster
  - you need enough sample performances to rate
- Speaking is easier to plan for
  - Everyone finishes at the same time
- Speaking involves technology
  - which can go wrong if it's not your institution and you don't know the equipment
- Writing takes time
  - Some rate very fast and slower raters feel under pressure
  - Slower raters must be given the time they need
  - What will you do with those who are finished fast?

How do you do it?

- Group size crucial

  Rule of thumb (min 12 max 20)



  - Big enough to allow productive discussion
  - Small enough to avoid chaos
- Data entry
  - Who is going to do it?
  - Writing is a possible one woman job
  - Speaking … not advisable. Plan for a second person

The rating sheet

- How will you recognize the rater?
  - Design a space to enter the rater number
- How will you know which performance the judgment is for?
  - Design a space to enter the performance number
- How do you want raters to record their justifications?
  - Reference to numbered descriptors on the rating scale? The CEFR? Or what standard?

The excel table

- Why use one?
  - To allow a more anonymous and objective discussion of rater reliability
  - To identify inter-rater issues
    - Criterion based?
    - Descriptor based?
    - Standard based
  - Identify intra-rater issues
    - Predictable rater vs unprediciable
    - Lenient vs severe
- How?
  - Setting up an excel template
- What?
  - Mean, mode, median