1. How to assess productive skills?
2. Assessment Criteria
3. Rating scales
4. Scale development and construction
5. Alternative approach – paired comparisons
Counting, quantitative

- Counting of “objective” features
- Points for features shown
- Deductions for missing features
- Errors in focus

Advantages – Disadvantages?

Judging, qualitative

• Evaluating performances
• Judging qualitative features
• Basis: scale or checklist which describes features in a positive way

Advantages / Disadvantages?

## Assessment strategies

<table>
<thead>
<tr>
<th></th>
<th>Counting “negative”</th>
<th>Judging “positive”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Procedure</strong></td>
<td>add up score; counting strategies</td>
<td>rate a performance; judging strategies</td>
</tr>
<tr>
<td><strong>Criteria</strong></td>
<td>quantity of errors</td>
<td>quality of acceptable performance</td>
</tr>
<tr>
<td><strong>Focus</strong></td>
<td>what is missing</td>
<td>what students already can do</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>‘objective’ assessment</td>
<td>qualitative description of observed performance</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>ignoring qualitative side of performance;</td>
<td>subjective judgement; ignoring quantity and missing features</td>
</tr>
</tbody>
</table>

Remember the ‘Psycholinguistic approach’ (Van Moere, 2012)
Assessment criteria
Assessment Criteria/Approach

• should be based on construct and learning objectives

• influenced by assessment culture and traditions – what criteria do you use?

• depending on context and assessment purpose: one holistic judgement, several analytic criteria, or a complementary approach

• if analytic approach: criteria should be defined independently from each other
Assessment criteria for writing

• Task Fulfillment
  – content
  – communicative effect
  – audience
  – register, style
  – genre

• Organisation
  – macrostructure
  – line of argument
  – coherence
  – cohesion
  – paragraphs

• Language
  – vocabulary
  – grammar
  – range / accuracy
  – orthography

• Other?
  – handwriting?
  – length?
  – ...


Assessment criteria Speaking I (Fulcher, webinar)

- **ACCURACY** - Language Competence
  - Pronunciation
  - Stress
  - Intonation
  - Syntax
  - Vocabulary
  - Cohesion

- **FLUENCY**
  - Hesitation
  - Repetition
  - False starts
  - Self-correction
  - Re-selecting lexical items
  - Restructuring sentences

- **COMMUNICATION STRATEGIES**
  - Overgeneralization
  - Paraphrase
  - Word coinage
  - Restructuring
  - Cooperative strategies
  - Code switching
  - Non-linguistic strategies
Assessment criteria Speaking II (Fulcher, webinar)

- DISCOURSE COMPETENCE
  - Turn taking
  - Openings and closings

- INTERACTIONAL COMPETENCE
  - Managing co-constructed speech

- TASK COMPLETION
  - Is the outcome successful?

- PRAGMATIC & SOCIOLINGUISTIC COMPETENCE
  - Appropriateness
  - Situational sensitivity
  - Topical knowledge
  - Cultural knowledge
Rating Scales
Rating Scales

For judgements, we need guidance:

• Assessment criteria are usually described in rating scales (holistic or analytic)
• Rating scales guide the assessment, help improve reliability and validity
• Contain one or several assessment criteria, described on several ascending levels or bands
• Form the basis for judgements
• Laborious to construct
• Rater training necessary
Different kinds of Rating Scales


- **Holistic** - one global impression ‘top-down’
  + quick, strengths can be acknowledged, higher inter-rater reliabilities
  - no diagnostic information, no profiles, halo effects, subjective criteria and weighting, validity?
- **Analytic** - several (defined) criteria ‘bottom-up’
  + guidance, higher intra-rater reliabilities, more meaningful feedback to learners, diagnostic information
  - time intensive, halo and holistic effects may still occur
- **Task-specific** - scale for one specific task
  + adequately addressing task demands
  - laborious in construction, can only be used for one task
RATING SCALES are not necessarily a description of performances, but...

“...a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance“.

(Lumley, 2002: 286)
Benchmarks

• Prototypical performance examples
• Illustrating certain criteria / levels of a scale
• Contain representative features relevant for a criterion / level
• Ideally chosen in team, using the rating scale / task in question
• Justifications critically important – description of what features and criteria the benchmark illustrates
Rating scale development and construction
Scale Construction - methods

- Intuitive, qualitative or/and quantitative methods
- Starting with existing descriptors (measurement-driven)
  starting with performance samples (evidence-based)
  starting with theories

Fulcher, Davidson, Kemp (2011)

- Intuitive methods: expert advice; teacher intuition; no data collection involved.
- Qualitative: formulate key concepts or analyse performances for salient features; use comparative judgements or sort performances; sort descriptor drafts (into criteria / levels); etc.
- Quantitative: discriminant analysis (qualitative key features - regression analysis for most significant ones); IRT-scaling of descriptor ratings; etc.
Scale Construction – the circular conundrum

- Expert view
- Scale criteria
- Performances
- Analyses of performances
- Scores
Scale Construction – principles

• The descriptions of the levels are meaningful on their own.
• They enable yes/no decisions.
• They describe abilities or knowledge in a positive way.
• They are concrete, clear and short.
• They contain not much technical terminology.
• They describe rather narrow bands.

Schneider & North 2000
Scale Construction – categories and levels

- Decide on construct-valid categories
- Decide on relevant levels/bands (how many do you need, how many are feasible?)
- Decide on holistic or analytic approaches – suitable for your assessment purposes!
- Decide on how to validate your scale (resources)
Paired Comparisons (Pollitt, 2004, 2009)

- Take “naïve” judges, experts in the field, judging effectiveness of performance
- No rating scale needed, no training needed
- Compare performances pairwise – which one is better, given the task?
- You need a large pool of judges
- Every performance needs to be judged by at least 12 judges (i.e. compared to 12 other performances)
- Results in calibrated judges and ranked performances, highly reliable and valid
The TERU work has shown that a simple comprehensive statement expressing what is *important* in the subject being assessed is a fully satisfactory definition of the task; a short discussion to make sure every judge understands what is and isn’t important is the only training needed – and this does not need to be repeated for each task.

(Pollitt 2009:5)


Fulcher, G. Webinar: https://larc.sdsu.edu/glennfulcher/assessingspeaking4-12.pdf


References


