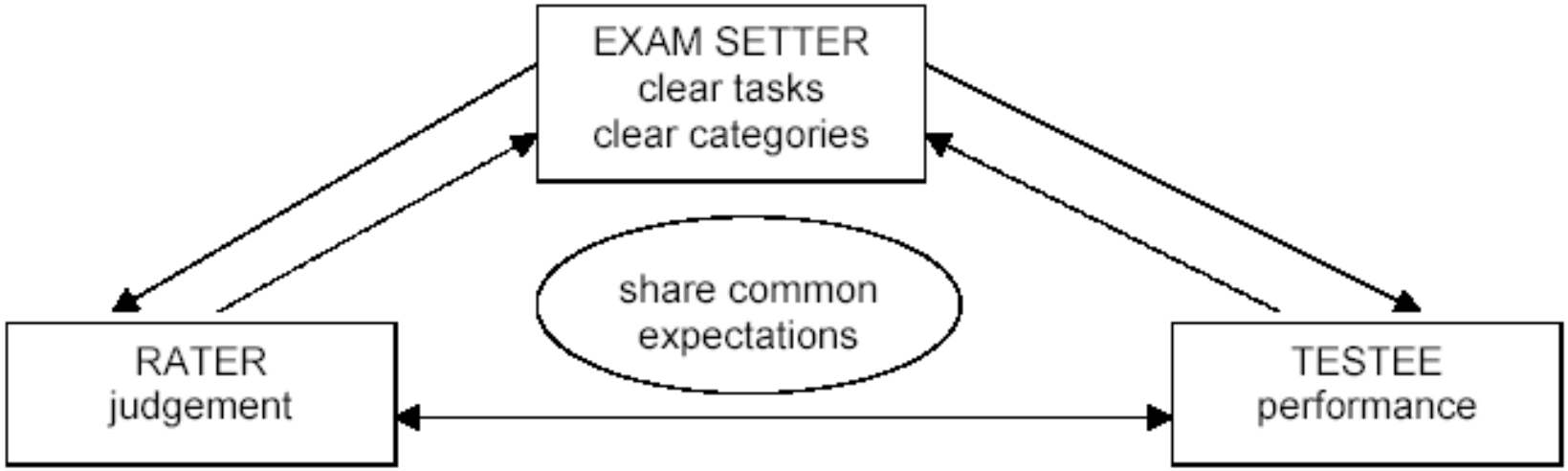
Rating the Constructs – Assessment Criteria Rating Scales

Nivja De Jong and Claudia Harsch

Overview

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



Cohen 1994: 307

Counting, quantitative

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

Advantages – Disadvantages?

Pollitt (1991) and Pollitt & Murray (1996)

Judging, qualitative

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

Advantages / Disadvantages?

Pollitt (1991) and Pollitt & Murray (1996)

Assessment strategies

	Counting“negative”	**Judging “positive”**
Procedure	add up score counting strategies	rate a performance judging strategies
Criteria	quantity of errors	quality of acceptable performance
Focus	what is missing	what students already can do
Advantages	‘objective’ assessment	qualitative description of observed performance
Dis-advantages	ignoring qualitative side of performance;	subjective judgement; ignoring quantity and missing features

Pollitt (1991) and Pollitt & Murray (1996)

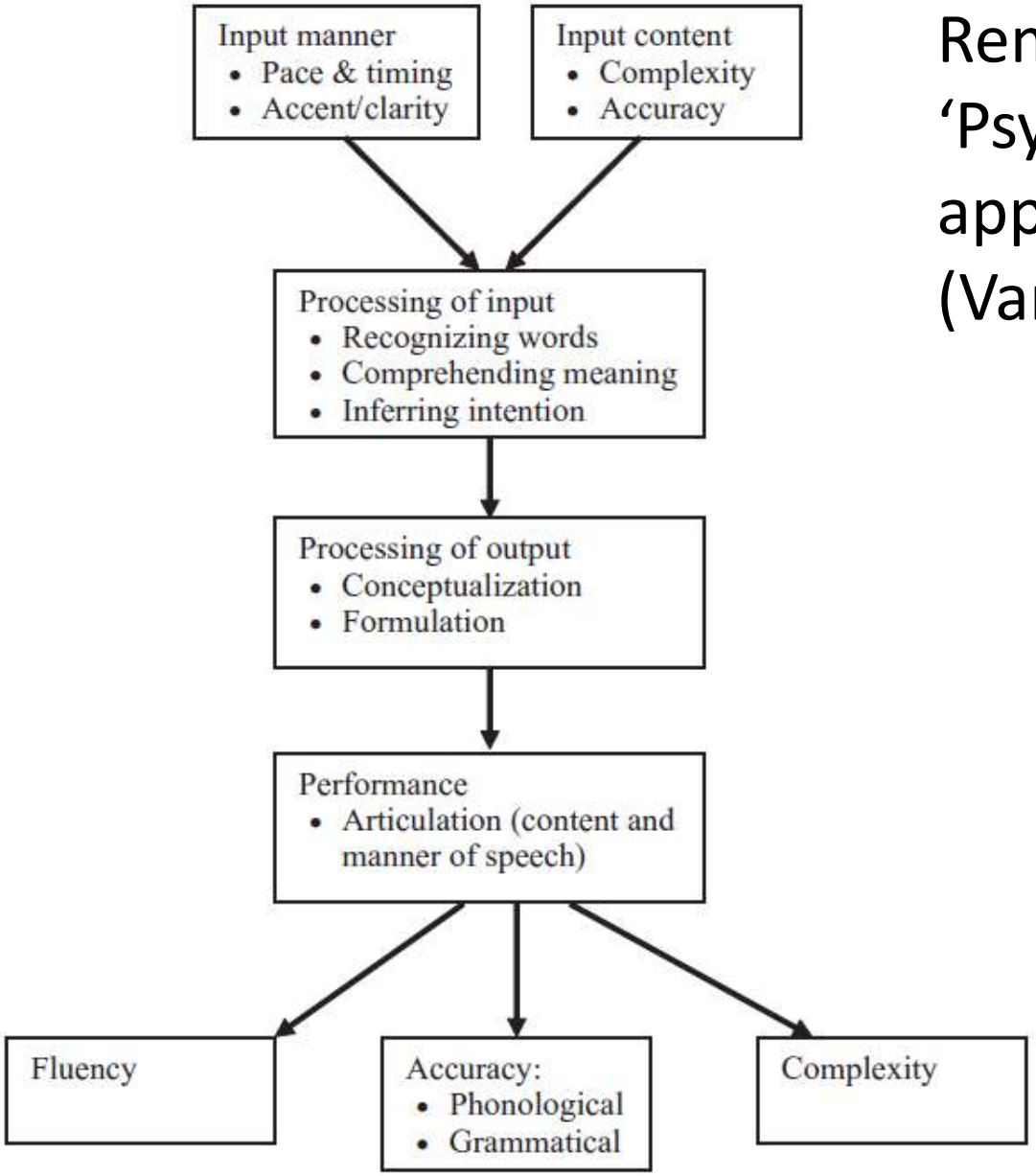


Figure 1. A 'three way model' for measuring processing competence in real-time oral tasks

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
- Other?
  - handwriting?
  - length?
  - …
- Organisation
  - macrostructure
  - line of argument
  - coherence
  - cohesion
  - paragraphs
- Language
  - vocabulary
  - grammar
  - range / accuracy
  - orthography

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
- PRAGMATIC & SOCIOLINGUISTIC COMPETENCE
  - Appropriateness
  - Situational sensitivity
  - Topical knowledge
  - Cultural knowledge
- TASK COMPLETION
  - Is the outcome successful?

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

e.g. Barkaoui, 2011; Hamp-Lyons 1995, Harsch & Martin, 2012, 2013

- *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 demads
  - 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

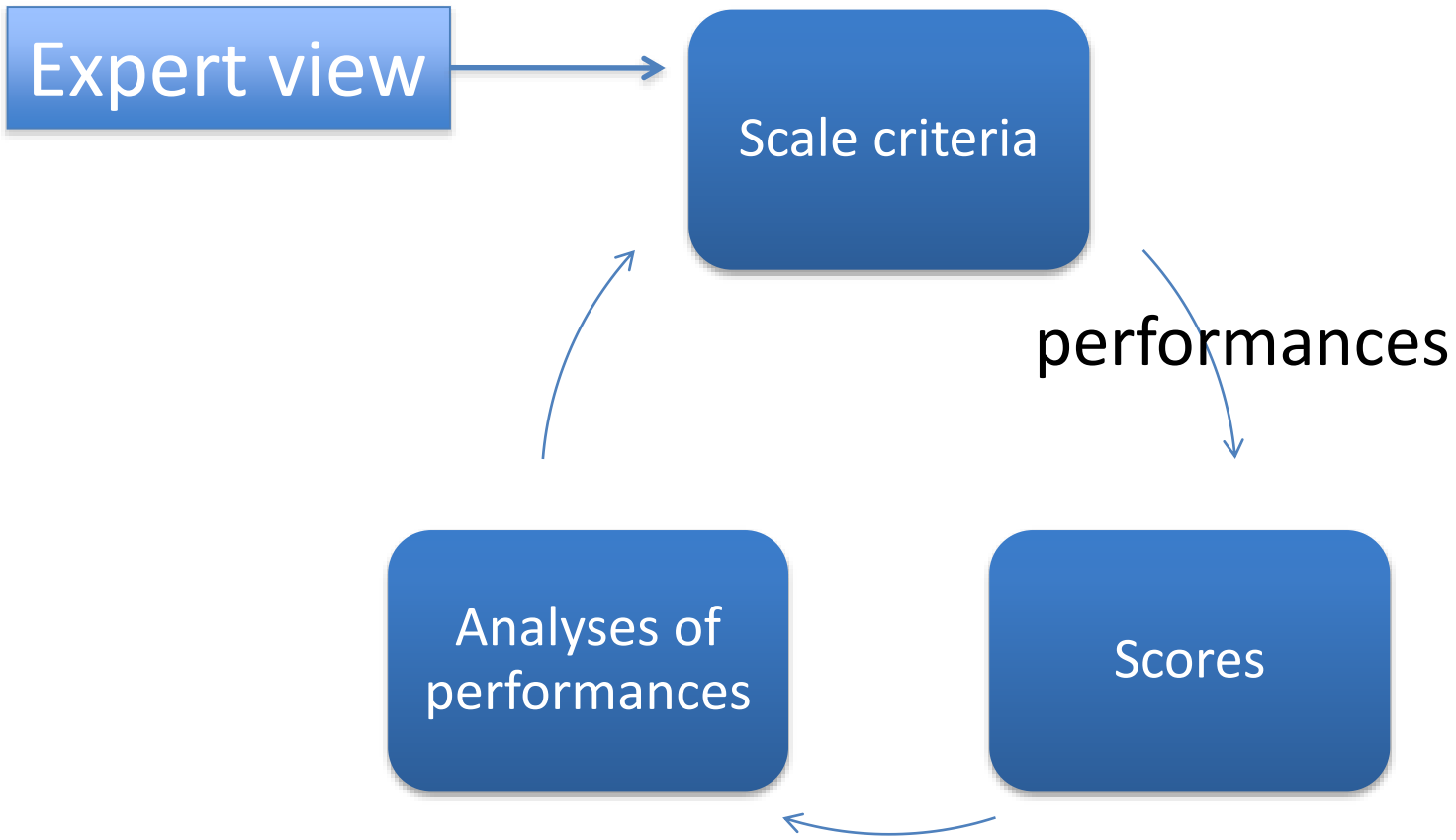
- Prototypical performance examples
- Illustrating certain criteria / levels of a scale
- Contain representative features relevant for a criterion / level
- Ideally chosen in team, using the rating scale / task in question
- Justifications critically important – description of what features and criteria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

Rating scale development and construction

Scale Construction - methods CEFR, Appendix A

- 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



Scale Construction – principles Schneider & North 2000

- 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.

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)

Alternative approach

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

Paired Comparisons (Pollitt, 2004, 2009)

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)

References

Alderson, C. 1991. Bands and Scores. In: Alderson, C. & North, B. (eds): *Language Testing in the 1990s: The Communicative Legacy*. London: Macmillan, 71-86.

Barkaoui, K. 2011. Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293.

Cohen, A. 1994². *Assessing Language Abilities in the Classroom*. Boston: Heinle & Heinle.

Council of Europe. 2001. The Common European Framework of Reference for Languages: Learning, Teaching and assessment. Strasbourg.

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

Fulcher, G., Davidson, F., Kemp, J. 2011. Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing* 28 (1), 5-29.

Hamp-Lyons, L. 1995. Rating Nonnative Writing: The Trouble with Holistic Scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. 1996. The Challenges of Second-Language Writing Assessment. In E. White, W. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Politics, Policies, Practices* (pp. 226-240). New York: MLAA.

Hamp-Lyons, L., & Kroll, B. 1996. Issues in ESL Writing Assessment: An Overview. *College ESL, 6*(1), 52-72.

Harsch, C. & Martin, G. 2013. Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice 20(3)*, 281-307.

References

Harsch, C. & Martin, G. 2012. Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17, 228–250.

Lumley, T. 2002. Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19 (3), 246-276.

McNamara, T. 2000. *Language Testing*. Oxford: Oxford University Press.

Pollitt, A. 1991. Giving Students a Sporting Chance: Assessment by Counting and by Judging. In: Alderson, C. & North, B. (eds): *Language Testing in the 1990s: The Communicative Legacy*. London: Macmillan, 46-59.

Pollitt, A. 2004. Let's stop marking exams. IAEA, Philadelphia, June 2004. Available at http://www.camexam.co.uk/ Our publications.

Pollitt, A. 2009. Abolishing marksism and rescuing validity. Paper presented at the 35th Annual Conference of the International Association for Educational Assessment, September, Brisbane, Australia. Available at: http://www.iaea.info/documents/paper_4d527d4e.pdf

Pollitt, A. & Murray, N. 1996. What raters really pay attention to. In: Milanovic, M. & Saville, N. (eds): *Language Testing 3 – Performance, Testing, Cognition and Assessment*. Cambridge: CUP, 74-91.

Schneider, G. & North, B. 2000. *Fremdsprachen können – was heißt das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Zürich: Ruegger.