EALTA Summer School, Innsbruck, 2016

Rating the construct reliably

Jayanti Banerjee and Claudia Harsch

Session Outline

- What is the rating process?
- Why do we need rater training?
- Rater training research
- The effect of rater feedback
- Good practice in rater training

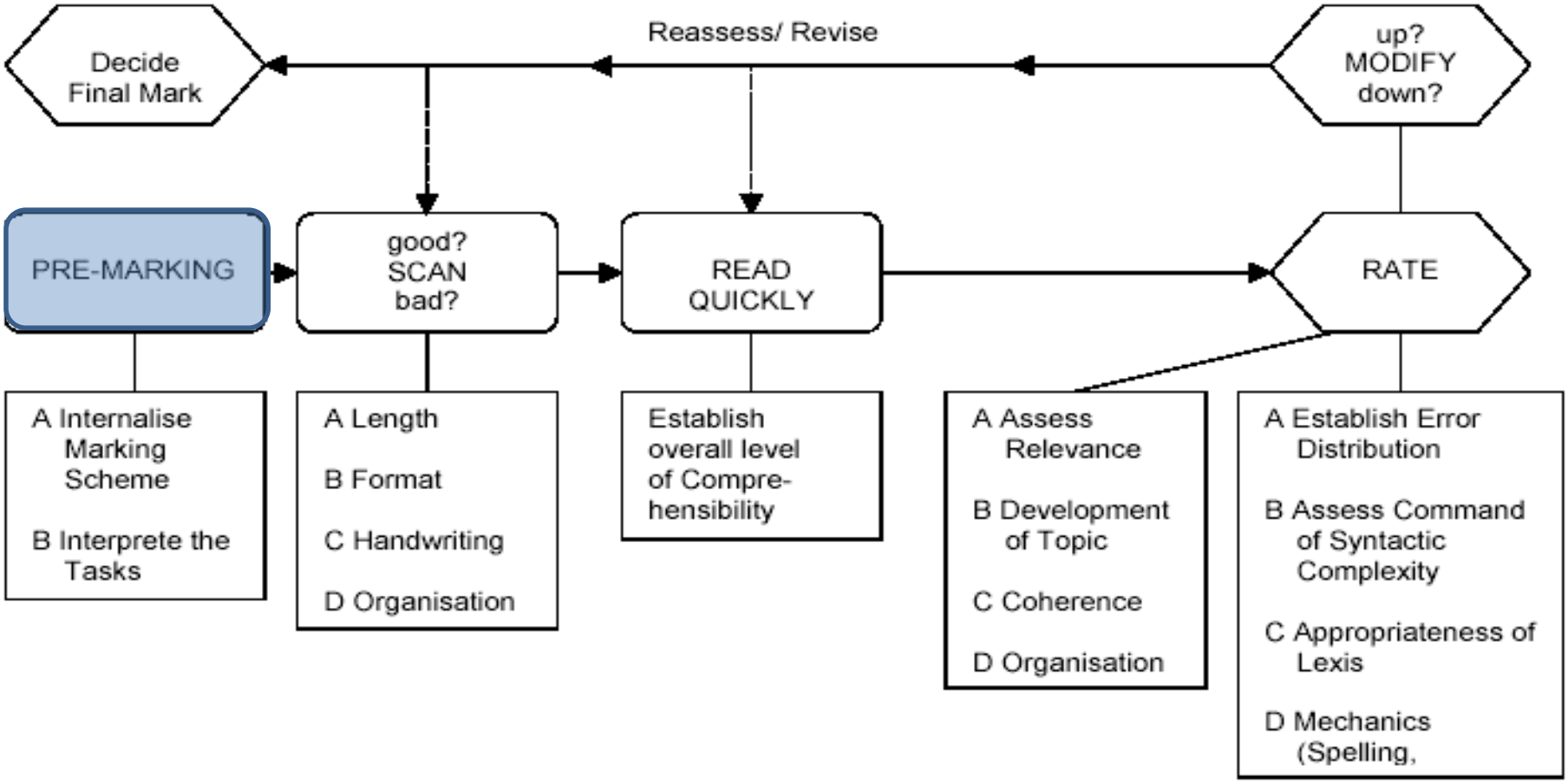
What is the rating process?

This has been investigated using:

- Think-aloud protocols
- Retrospective verbal reports
- Retrospective written reports
- Interviews

Model of the rating process – writing assessment

Milanovic, Saville & Shuhong (1996, p. 95)



Rating stages – writing assessment

Lumley (2002)

Stage	**Rater's focus**	**Observable behaviours**
1. First reading (pre-scoring)	Overall impression of text: global and local features	• Read task and text • Comment on salient features
2. Rate all categories in turn	Scales and text (and benchmarks)	Articulate and justify scores • Refer to scale descriptors • Reread text
3. Consider scores given	Scale and text: consistency of scores given	• Compare with bench-mark texts or other rated texts • Confirm or revise existing score

Rating exercise

In groups, discuss the rating exercise that you completed overnight:

- How much did you agree in your overall judgements?
- How much did you agree at the descriptor level?

IQB Study Method (Harsch & Martin 2013)

6 raters, training and familiarisation with tasks and scales, rating approaches in 2 separate studies

1. Give top-down scores (overall and criteria)
2. Give bottom-up descriptors-scores, and complement them with top-down scores

=> Compare 1. and 2., using the index *percent agreement with mode* (developed to monitor scoring consistency)

1. Results ‘top-down’

Criterion	**Agreement B1-task** (n=28 scripts)	**Agreement A2-task** (n=28 scripts)
Task Fulfilment	0.76	0.84
Organisation	0.85	0.86
Grammar	0.85	0.85
Vocabulary	0.87	0.87
Overall	0.86	0.83

Top-down rater agreement

Criterion	**R01**	**R02**	***R03***	**R04**	***R05***	**R06**
Task Fulfilment	0.81	0.78	*0.70*	0.74	*0.70*	0.81
Organisation	0.89	0.96	*0.70*	0.85	*0.78*	0.93
Grammar	0.78	0.96	*0.74*	0.81	*0.85*	0.93
Vocabulary	0.81	0.96	*0.70*	0.85	*0.93*	0.96
Overall	0.85	1.00	*0.74*	0.89	*0.70*	0.96

2. Results ‘bottom-up’

Criterion	Agreement B1 task (n=55 scripts)
Task Fulfilment	0.81
Organisation	**0.83**
Grammar	0.85
Vocabulary	0.84
Overall	0.87

Results ‘bottom-up’

Descriptors for *Organisation*	**Agreement**
Descriptor O1	**0.75**
Descriptor O2	**0.56**
Descriptor O3	**0.73**
Descriptor O4	0.82
Descriptor O5	**0.54**
Descriptor O6	0.83
Descriptor O7	0.84
Descriptor O8	**0.66**
Descriptor O9	**0.63**

Bottom-up rater agreement

Descriptors for *Organisation*	**R01**	**R02**	**R03**	**R04**	**R05**	**R06**
Descriptor O1	0.84	0.80	0.76	**0.62**	**0.53**	0.93
Descriptor O2	**0.62**	**0.33**	**0.25**	**0.73**	0.78	**0.64**
Descriptor O3	0.80	0.91	0.80	0.78	**0.71**	**0.38**
Descriptor O4	0.93	0.93	0.82	0.75	0.87	**0.65**
Descriptor O5	0.78	**0.47**	0.76	**0.15**	0.84	**0.25**
Descriptor O6	0.80	0.84	0.78	0.80	0.91	0.87
Descriptor O7	0.76	0.89	0.84	0.82	0.84	0.87
Descriptor O8	0.98	0.93	0.76	**0.36**	0.89	**0.05**
Descriptor O9	**0.02**	0.78	**0.20**	0.91	0.93	0.96
AVG	0.73	0.76	0.66	0.66	0.81	0.62

Interpretation

- Fairly high agreement on criterion-layer ratings is NOT the result of uniform interpretation of descriptors …
- BUT rather results from cancellation of deviations on the descriptor-layer during the formation of the criterion judgments

➢ Compromises rating validity *if we do not monitor* what raters are actually doing with the descriptors

Why is rater agreement important?

Reliability:

Ensure that the same/a very similar score is awarded to a performance by different raters (or by the same rater on different occasions).

Validity:

Confidence that when two raters give the same score, it is for the same reasons. In other words, the score has the same meaning regardless of who scores the performance.

"Rating is certainly possible without training, but in order to obtain reliable ratings, both training and reorientation are essential in order to allow raters to learn or (re)develop a sense of what the institutionally sanctioned interpretations are of task requirements and scale features, and how others relate personal impressions of text quality to the rating scale provided."

Lumley, 2002: 267

Rater training research

Weigle (1998)

- Interested in the effect of rater training upon experienced and novice raters.
- Study included 8 experienced raters and 8 novice raters.

Rater training research

Weigle (1998)

- Procedure:
  - Rate a batch of essays
  - Participate in a rater calibration session
  - Rate a second batch of essays
- Analysis of ratings to establish rater severity/leniency and consistency:
  - Before training
  - After training

Rater training research

Weigle (1998)

- Before training results
  - The raters differed in their levels of severity.
  - The novice raters tended to be much more severe than the experienced raters.
  - The novice raters tended to be less consistent in their judgements than the experienced raters.

Rater training research

Weigle (1998)

- After training results
  - The raters still differed in their levels of severity; novice raters were still significantly more severe than experienced raters.
  - But the magnitude of the difference was smaller.
  - The novice raters were generally more consistent in their rating than previously.

Rater training research

Weigle (1994)

- Conducted verbal protocols with four inexperienced raters:
  - Before training
  - After training
- Analysis showed that, after training they:
  - Understood the rating criteria better.
  - Had modified their expectations of student writing.
  - Had a reference group of other raters to compare themselves against.

Rater training research

"… it is not enough to be able to assign a more accurate number to examinee performances unless we can be sure that the number represents a more accurate definition of the ability being tested." (Weigle, 1998, p. 281)

Effect of rater feedback

Wigglesworth (1993)

- Two versions of the same speaking test – a face-to-face version and a tape-mediated version – with a number of tasks.
- Analysis to uncover rating patterns by task (leniency / severity)
- Assessment ‘map’ for each rater
- Feedback presented to each rater individually.

Effect of rater feedback

Wigglesworth (1993)

- Subsequent rating showed evidence that the feedback was useful – differences between raters reduced.
- How long does the effect of feedback last?
- No control group – so we don't know if the improved results were due to the feedback.
- This approach results in more comparable scores but does not guarantee that the raters are interpreting the scale in the same way.

Effect of rater feedback

Knoch (2011)

- Looked at longitudinal effects of rater feedback for both speaking and writing.
  - Are some rating behaviours more amenable to feedback than others?
  - Do speaking and writing raters differ in their receptiveness to feedback?
  - What are the raters perceptions of the feedback?
  - Is there a relationship between their perceptions of the feedback and improvements in rating behavior.

Effect of rater feedback

Knoch (2011)

- Feedback areas
  - Severity
  - Bias (severity/leniency with respect to a particular rating criterion)
  - Consistency in their use of rating scale categories
  - Overall evaluation

Effect of rater feedback

Knoch (2011)

- Findings were disappointing (but useful):
  - Feedback did not appear to be any more effective than random variation.
  - Raters were generally positive about the feedback but did not use it at all or did not use it long term.
  - There was little relationship between the raters' perception of the feedback and their subsequent ratings.

Effect of rater feedback

Knoch (2011)

- Findings were useful. The problem might not be with the concept of giving feedback but with:
  - The type of feedback given.
  - How it was delivered.
  - When it was delivered.

How might raters still differ?

Raters may be consistently more lenient/harsh:

- Regardless of who they are examining (perhaps because of their level of experience and/or training)
- Depending on a specific variable:
  - Type of task
  - The criterion being assessed
  - Certain groups of test takers

How might raters still differ?

Eckes (2008)

- Studied the rating behaviour of trained and experienced raters of the TestDaF (a test of German as a second language)
- All the participants met the quality assurance criteria for being a TestDaF rater.

How might raters still differ?

Eckes (2008)

- Method:
  - Raters first indicated the importance of each of the TestDaF marking criteria (from less important to extremely important)
  - Based on these results, raters could be clustered according to their ratings of the criteria.

How might raters still differ?

Eckes (2008)

- Results:
  - Six rater types emerged
    - One type (N = 1) privileged 'correctness' above all the other criteria.
    - Two types (N = 5 and N = 6) did not perceive *any* of the criteria to be extremely important.

"… high interrater reliability could simply be due to these raters' type-specific points of view regarding the weight of scoring criteria, which actually may capture only a small part of the construct being assessed." (p. 179)

SUMMARY

Rater training is an important step in assuring the validity of the interpretation of your test scores. BUT… raters are human beings and there will always be variation in their decision-making processes.

This means that rater training and rater conferences must be on-going.

References

- Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing*, 25(2), 155-185.
- Elder, C., Knoch, U., Barkhuizen, G., and von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2(3), 175-196.
- Harsch, C. & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education* 20(3), 281-307.
- Knoch, U. (2011). Investigating the effectiveness of individualised feedback to rating behaviour – a longitudinal study. *Language Testing*, 28(2), 179-200.
- 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.
- Milanovic, M., Saville, N. & Shuhong, S. (1996). A Study of the Decision-Making Behaviour of Composition Markers. In: Milanovic, M. & Saville, N. (eds): *Language Testing 3 – Performance, Testing, Cognition and Assessment*. Cambridge: CUP, 92-114
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-319.
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197-223.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 253-287.