Some considerations on

STANAG 6001 Tester Training

Gerard Seinhorst
Netherlands Defence Language Centre

BILC STANAG 6001 Testing Workshop 2017
Skopje, Macedonia

Outline

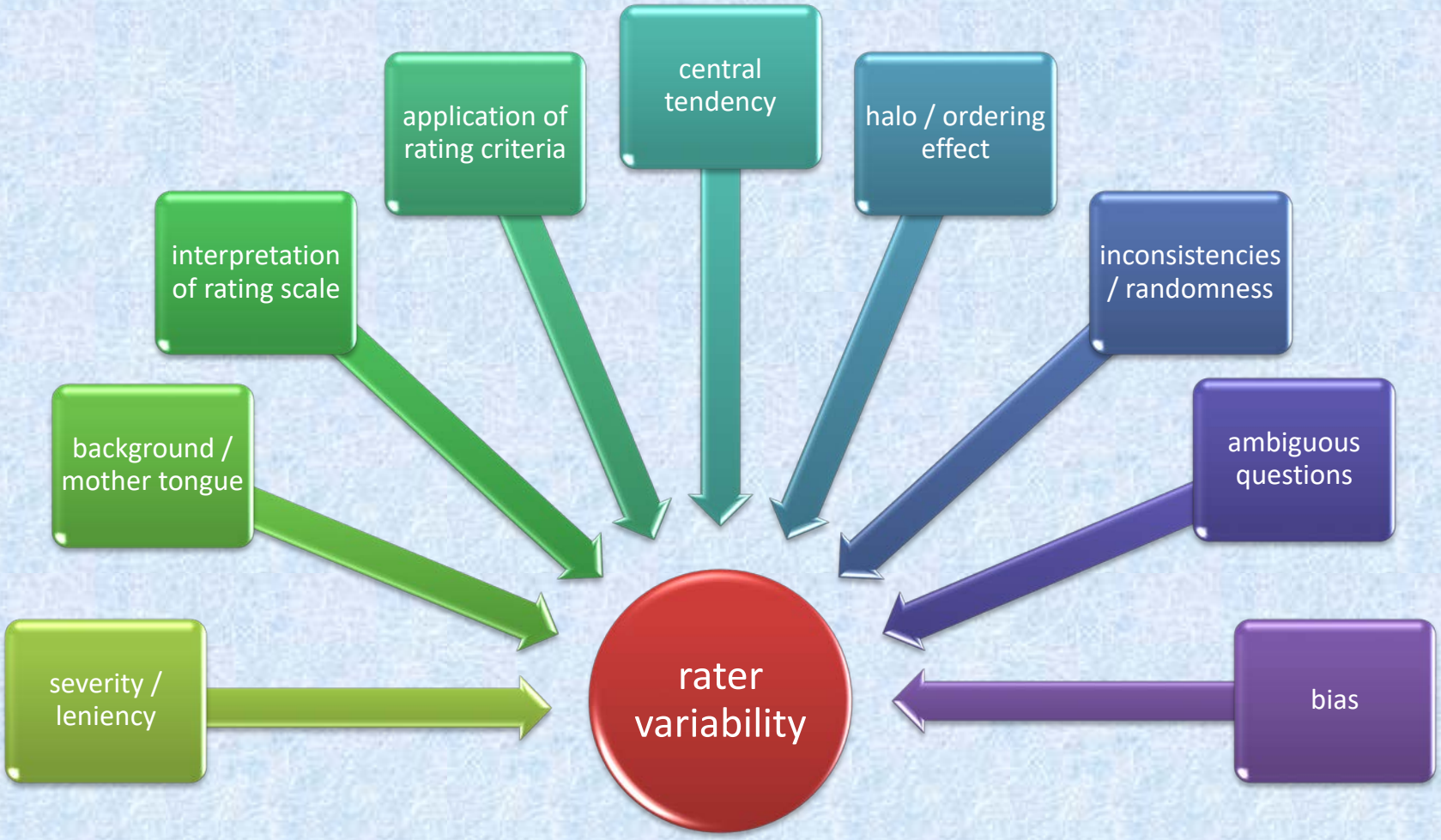
- The importance of tester training
- The effects of training - *Research findings*
- General and role-specific tester training
- References

Introduction

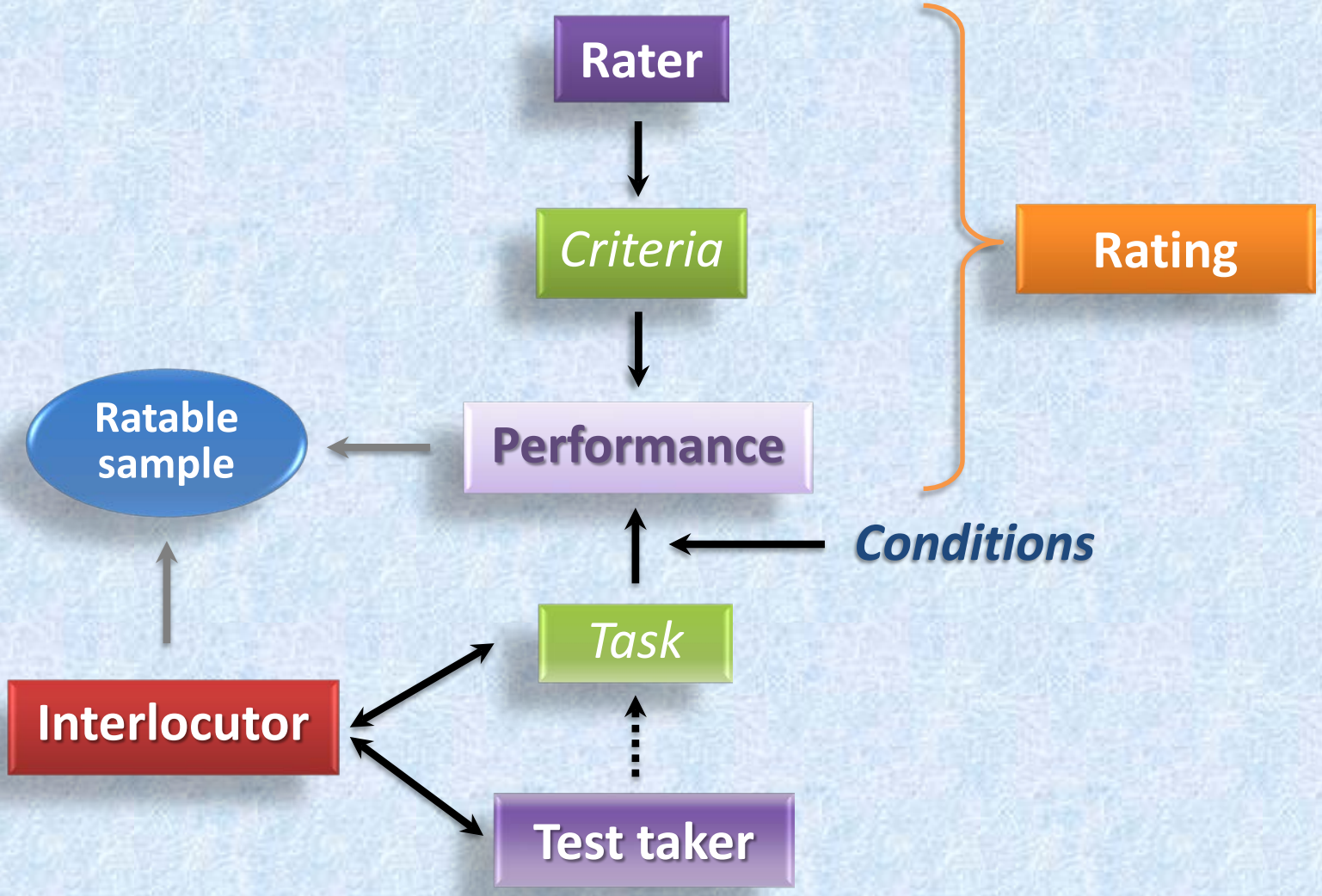
'If we do not know what raters are doing, then we do not know what their ratings mean'

Connor-Linton (1995)

Causes of rater variability



Interaction in assessment



Adapted from McNamara (1997)

The role of the interlocutor

- Shift of focus in language testing – from maximizing reliability to optimizing validity
- Potential of the interlocutor to affect the quality of the test taker's performance
  - differences in the way that interlocutors interact with test takers
  - tester personality - bias
  - tester stance
  - different elicitation techniques
- Impact on the quality of test taker's performance
  - affecting the validity and fairness of tests and the rating given

The importance of training

- Perform the required task to a **common standard**
  - gain knowledge of assessment methodology and testing principles
  - reach a common understanding/interpretation of rating scale
  - achieve consistency in the application of the rating criteria
  - minimize tester idiosyncrasies and construct-irrelevant variance
  - follow standardized testing procedures
  - enhance alignment of ratings
- Increase/maintain **professionalism** and **quality**
  - make informed decisions
  - selection and qualification/certification of testers

Ensure that the **inferences** made on the basis of the test results are **valid**, **accurate**, and **fair**

Research findings

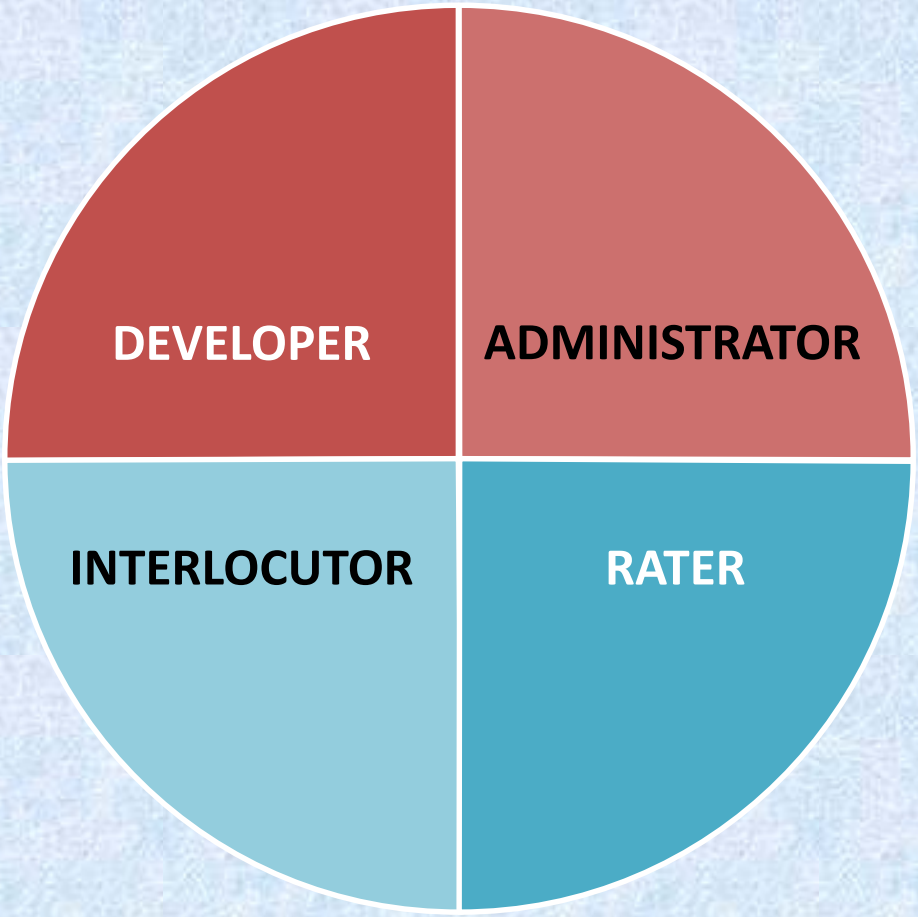
- Tester training leads to higher *inter-rater reliability*, but not necessarily to higher *intra-rater reliability*
- Tester training effects do not persist; raters tend to become more lenient over time
- NNS raters tend to be more severe re. grammar/vocab errors, but more lenient re. interference of L1 accent
- Experience is no guarantee for accurate ratings
- Rating criteria are applied more reliably when they are accompanied by benchmark performances
- Item writing guidelines are more effective when they are supported by examples of ‘strong’ and ‘weak’ items and statistical data

Research findings

Implications

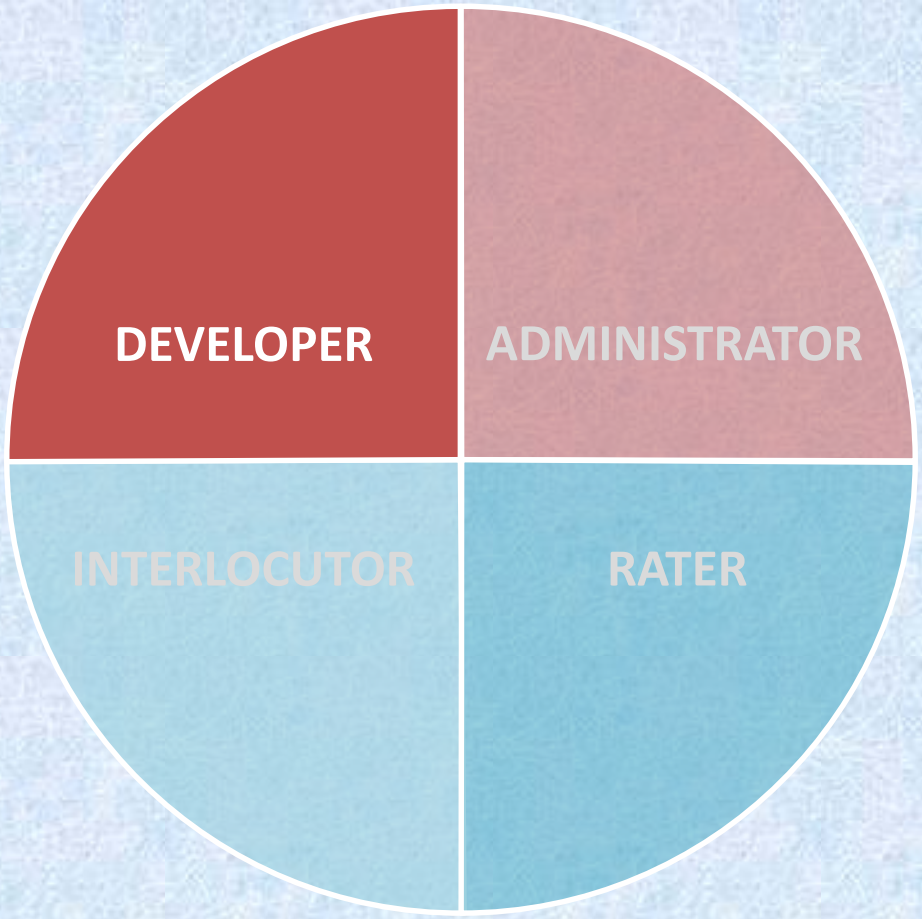
- Tester background (mother tongue, gender, teaching experience) does not play a decisive role in becoming a good examiner
- Training needs to be followed up at regular intervals to ensure that standards are maintained. Ideally, each testing session should be preceded by a renorming/recalibration session to reduce interlocutor idiosyncrasies and rater variability
- Training cannot be expected to remove all variability. The **number of test occasions** has a far greater impact on reliability than tester training or employing multiple raters
- **Practice** is the key to becoming proficient in item writing, conducting speaking tests and rating performances

Language Tester Roles

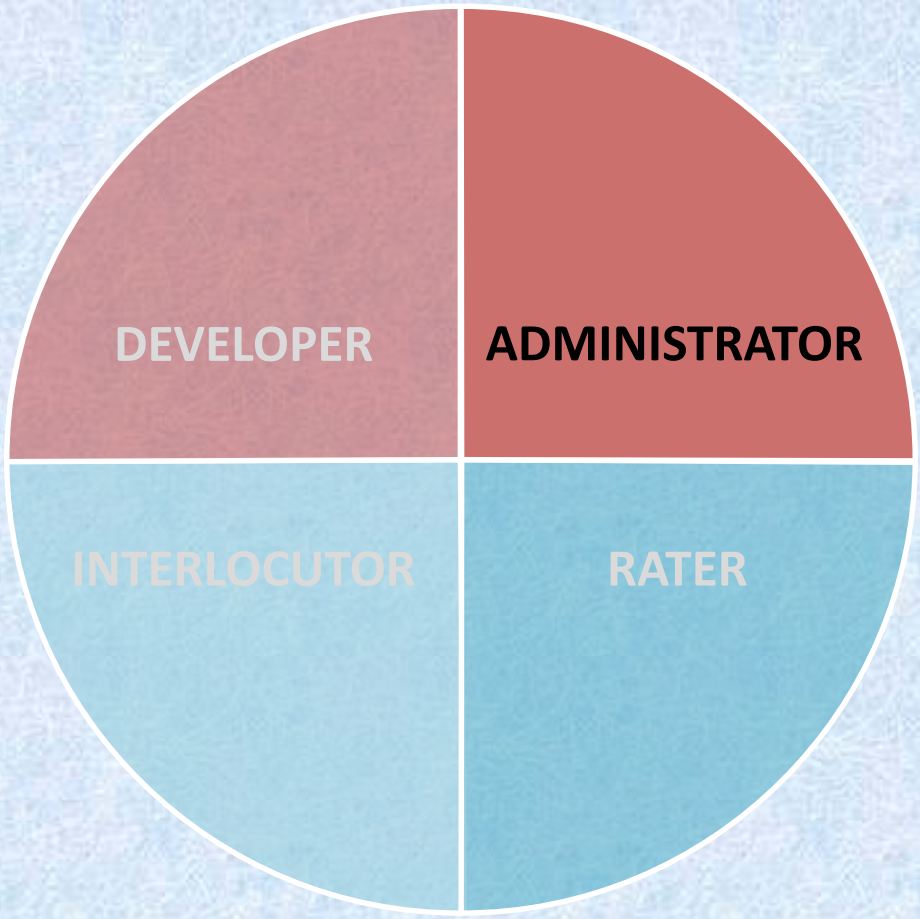


Language Tester Roles

- test specs
- CTA alignment
- item/prompt writing
- moderation
- pretesting
- test/item analysis
- validation

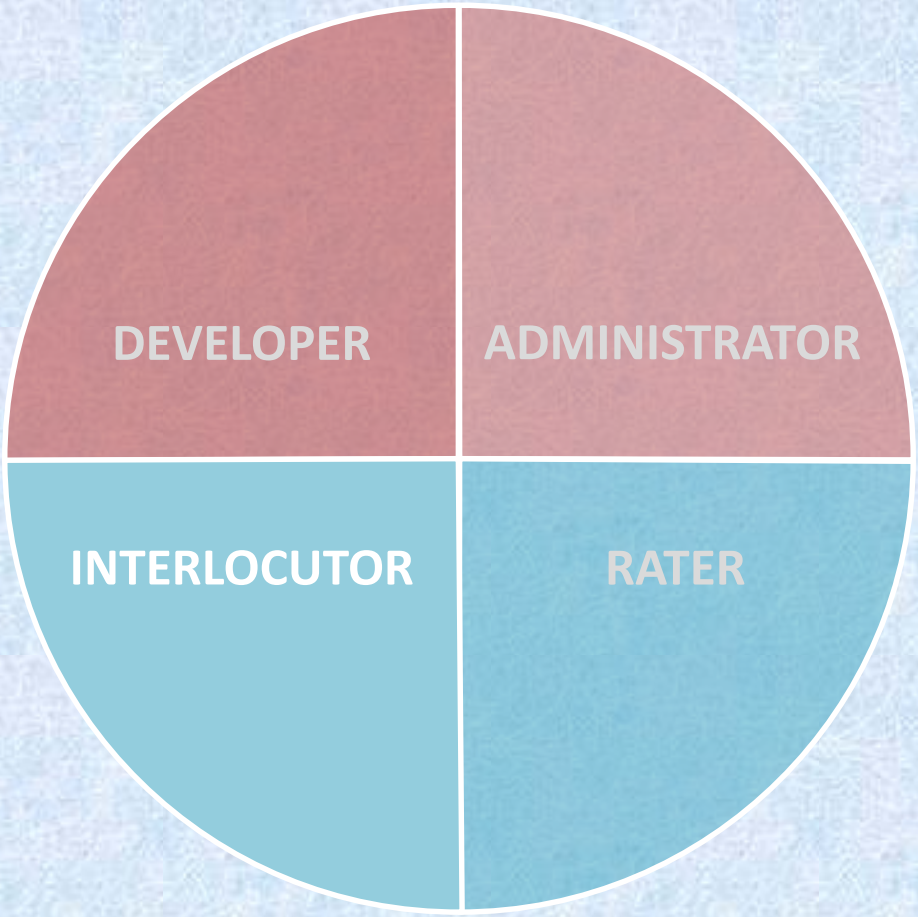


Language Tester Roles



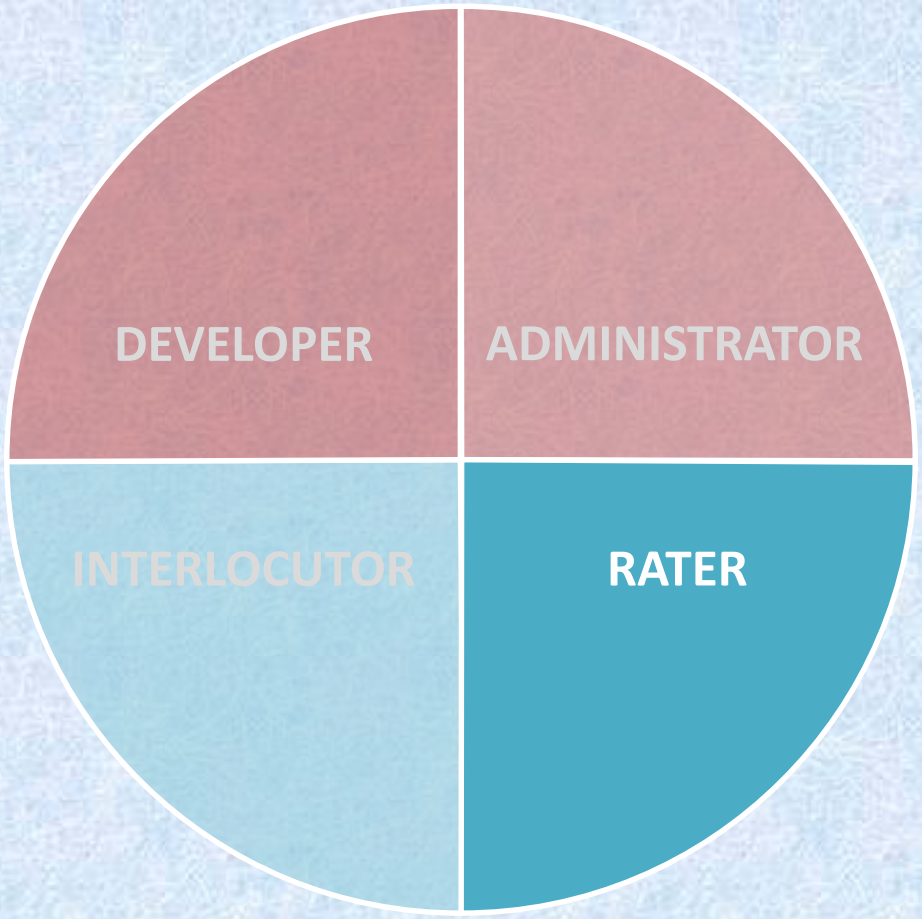
- scheduling
- scoring procedures
- test admin protocols
- test security
- retesting policy
- reporting test results
- test certificates
- reproduction/storage

Language Tester Roles



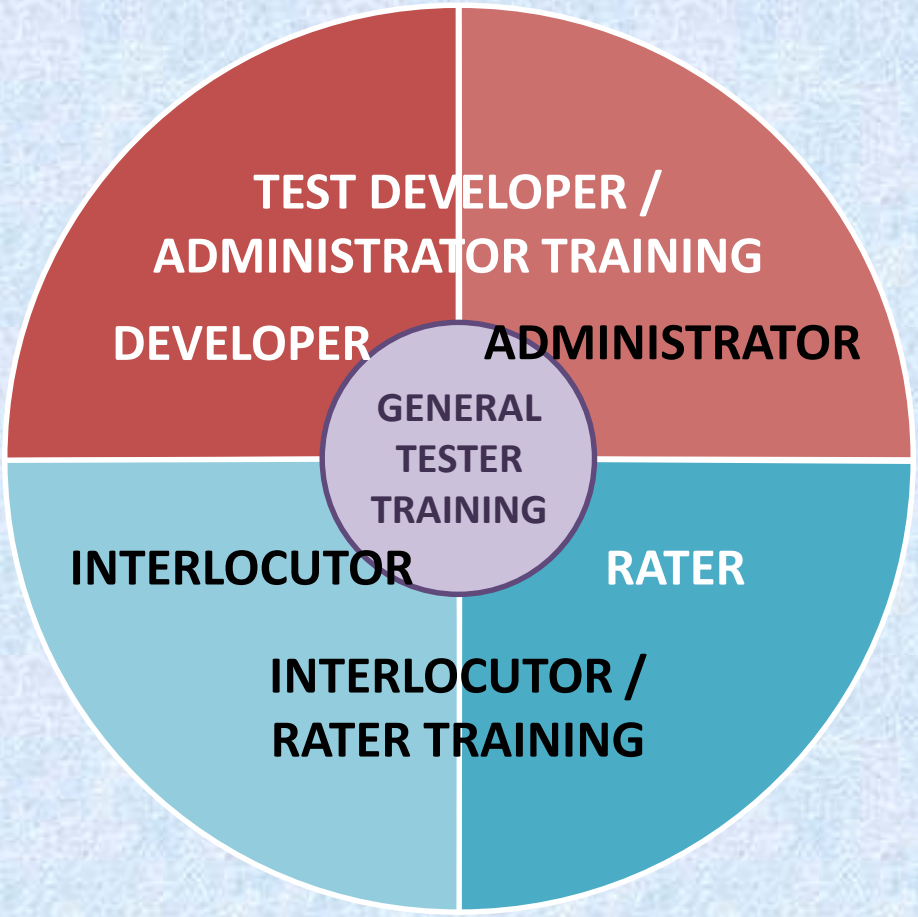
- speaking test protocol
- elicitation techniques
- ratable speech sample
- tester behaviour
- dealing with non-standard performances

Language Tester Roles



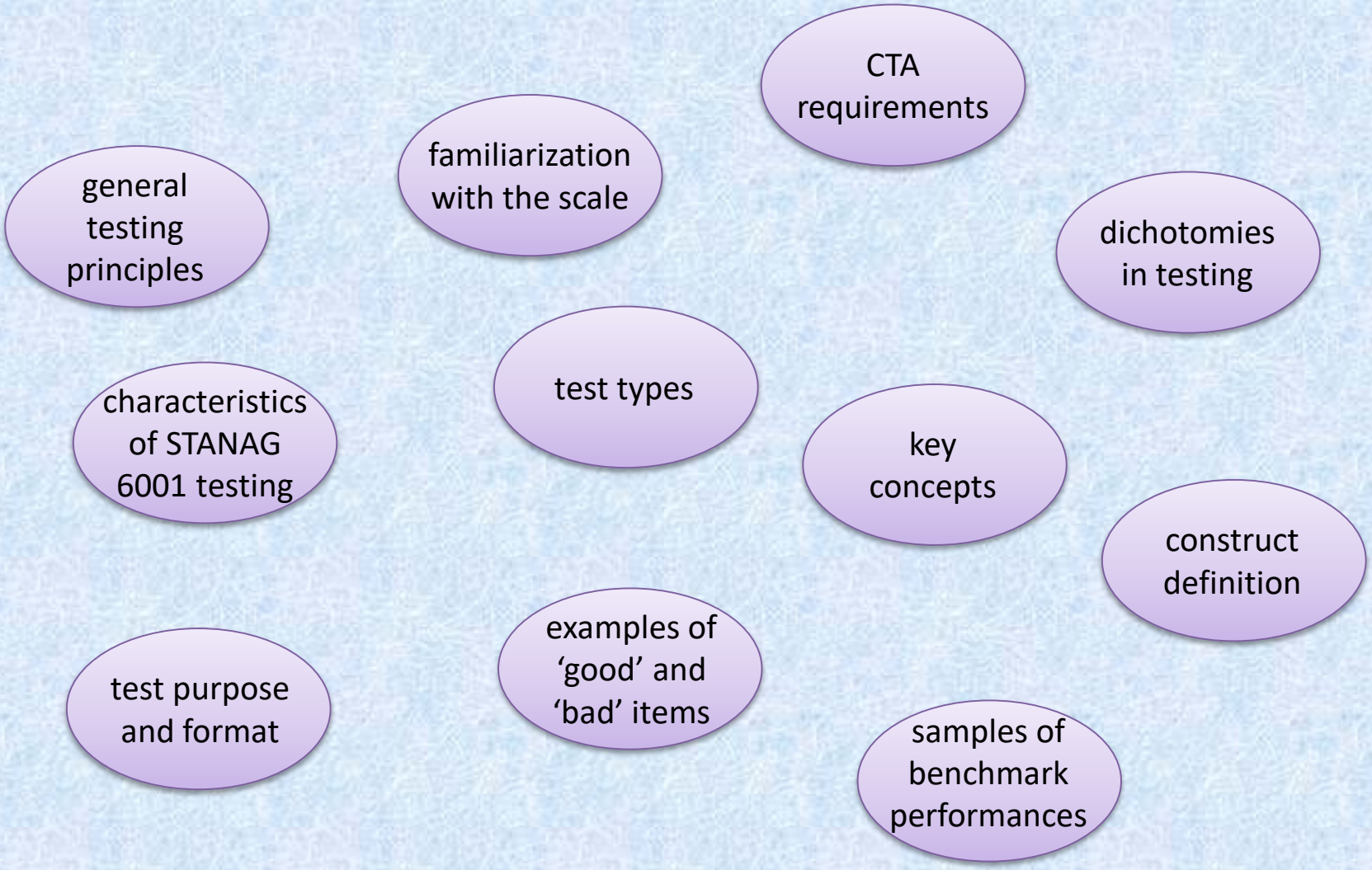
- norming/benchmarks
- rating criteria
- consistency/reliability
- analytic and holistic rating
- adjudication

Types of Tester Training



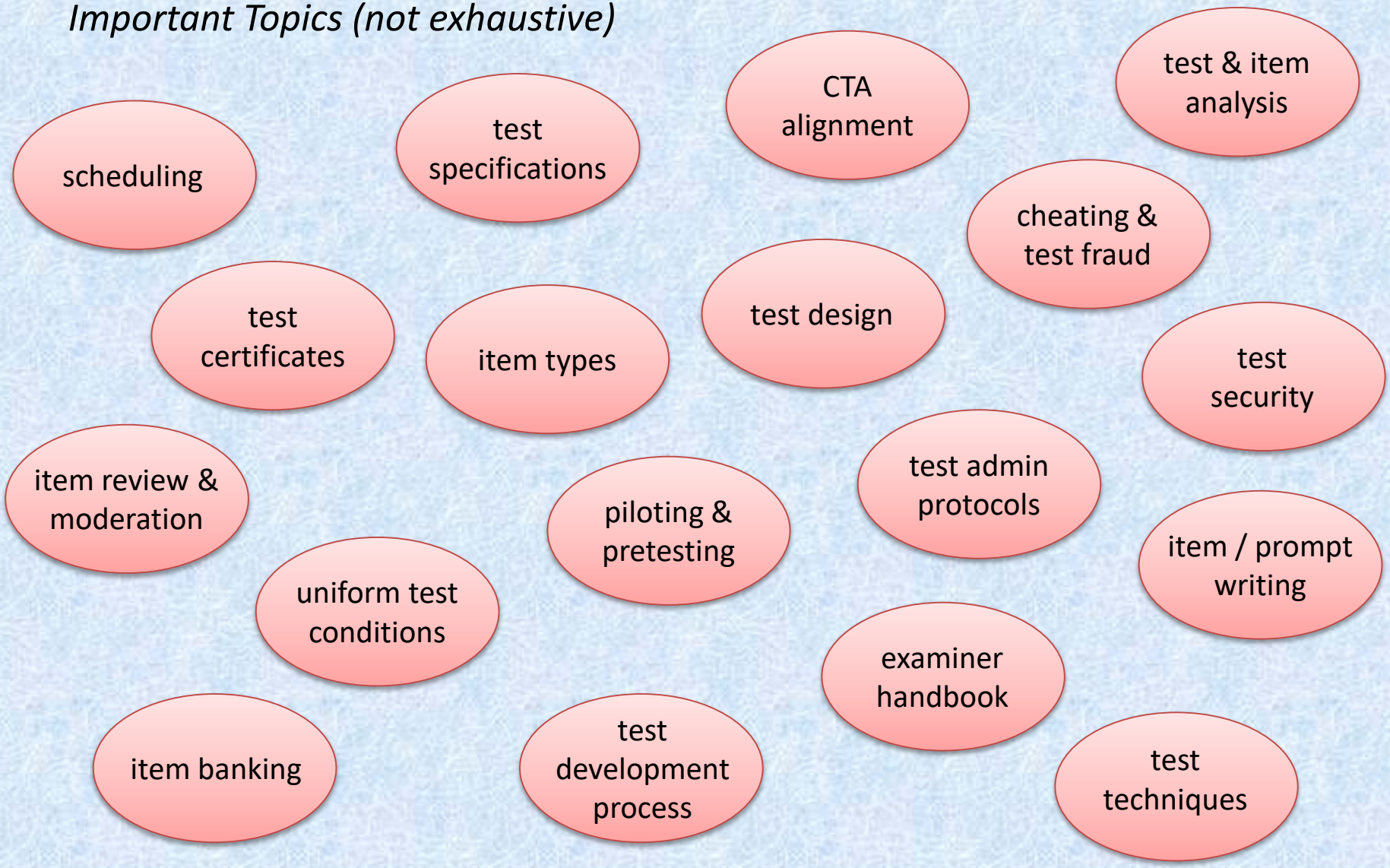
General Tester Training

Important Topics (not exhaustive)



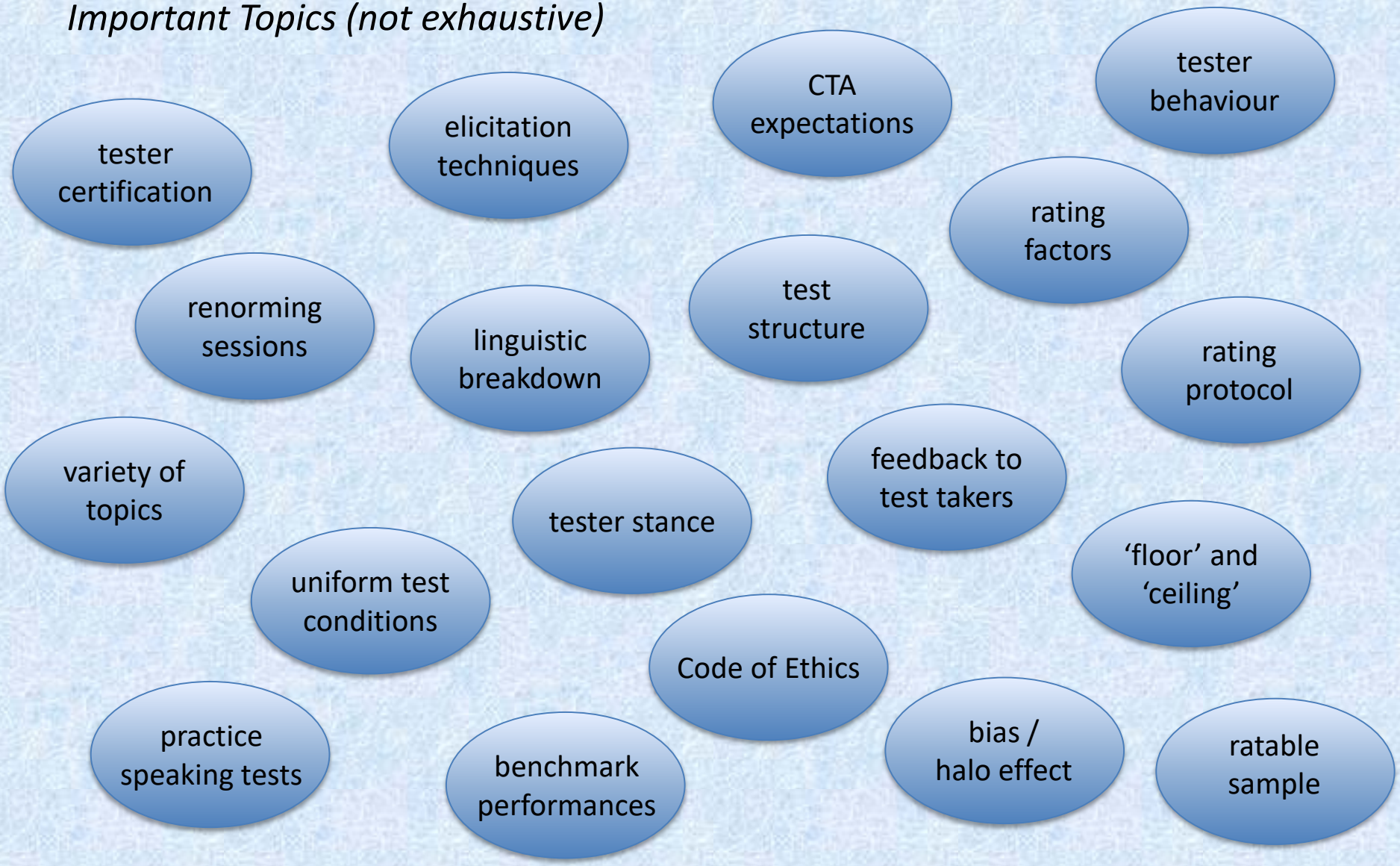
Test Developer/Administrator Training

Important Topics (not exhaustive)



Interlocutor/Rater Training

Important Topics (not exhaustive)



References / Further reading

Association of Language Testers in Europe (ALTE). (2005). ALTE Materials for the guidance of test items writers. 1995, updated 2005. http://www.alte.org/attachments/files/item_writer_guidelines.pdf

Brown, A. Interlocutor and rater training. In: Fulcher, G., and Davidson, F. (Eds.) *The Routledge Handbook of Language Testing.* Routledge: 413-425

Connor-Linton, J. (1995). Looking behind the curtain: what do L2 composition ratings really mean? *TESOL Quarterly* 29: 762-65.

Fulcher, G., and Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge.

Hogan, T.P., and Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items; What the experts say. *Applied Measurement in Education*, 20(4): 427-41.

McNamara, T.F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446-65.

Schedl, M. and Malloy, J. (2014). Writing Items and Tasks. In: Kunnan, J. (Ed.) *The Companion to Language Assessmen*t. John Wiley & Sons, 788-804.

Van Moere, A. (2014). Raters and Ratings. In: Kunnan, J. (Ed.) *The Companion to Language Assessmen*t. John Wiley & Sons, 1340-57.