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

- Application of rating criteria
- Central tendency
- Halo / ordering effect
- Inconsistencies / randomness
- Ambiguous questions
- Bias
- Severity / leniency
- Background / mother tongue
- Interpretation of rating scale
Interaction in assessment

- Rater
- Criteria
- Performance
- Task
- Test taker
- Interlocutor
- Conditions
- Rating

Ratable sample

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

- Administrator
- Developer
- Interlocutor
- Rater
Language Tester Roles

- Developer
  - Test specs
  - CTA alignment
  - Item/prompt writing
  - Moderation
  - Pretesting
  - Test/item analysis
  - Validation

- Administrator

- Interlocutor

- Rater
Language Tester Roles

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

- **ADMINISTRATOR**

- **INTERLOCUTOR**

- **RATER**
Language Tester Roles

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

- DEVELOPTER
- ADMINISTRATOR
- INTERLOCUTOR
- RATER

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

Important Topics (not exhaustive)

- general testing principles
- test types
- characteristics of STANAG 6001 testing
- test purpose and format
- familiarization with the scale
- examples of ‘good’ and ‘bad’ items
- CTA requirements
- dichotomies in testing
- key concepts
- construct definition
- samples of benchmark performances
- examples of benchmark performances
Test Developer/Administrator Training

Important Topics (not exhaustive)

- scheduling
- test specifications
- CTA alignment
- cheating & test fraud
- test & item analysis
- test design
- test admin protocols
- test security
- item types
- uniform test conditions
- piloting & pretesting
- examiner handbook
- item / prompt writing
- test techniques
- item review & moderation
- item banking
- test development process
- uniform test conditions
- item review & moderation
Interlocutor/Rater Training

Important Topics (not exhaustive)

- tester certification
- elicitation techniques
- CTA expectations
- tester behaviour
- rating factors
- rating protocol
- variety of topics
- linguistic breakdown
- test structure
- feedback to test takers
- ‘floor’ and ‘ceiling’
- uniform test conditions
- tester stance
- Code of Ethics
- bias / halo effect
- ratable sample
- practice speaking tests
- benchmark performances
- renorming sessions


