Relating a reading test to STANAG 6001 proficiency levels: lessons learnt

Ülle Türk
Language Testing Unit
Estonian Defence Forces

The 8th Annual EALTA Conference
Siena, 5–8 May 2011

Overview

- Why?
- How?
- What did we learn?

Situation

- NATO member states must use STANAG 6001 proficiency levels to describe foreign language proficiency of military personnel
  - 4 skills: SLP (LSRW)
  - 6 levels: Level 0 (No proficiency) – Level 5 (Highly articulate native)
- Each country responsible for designing own tests
  - Small teams
  - Small test populations
- Little cooperation and standardisation

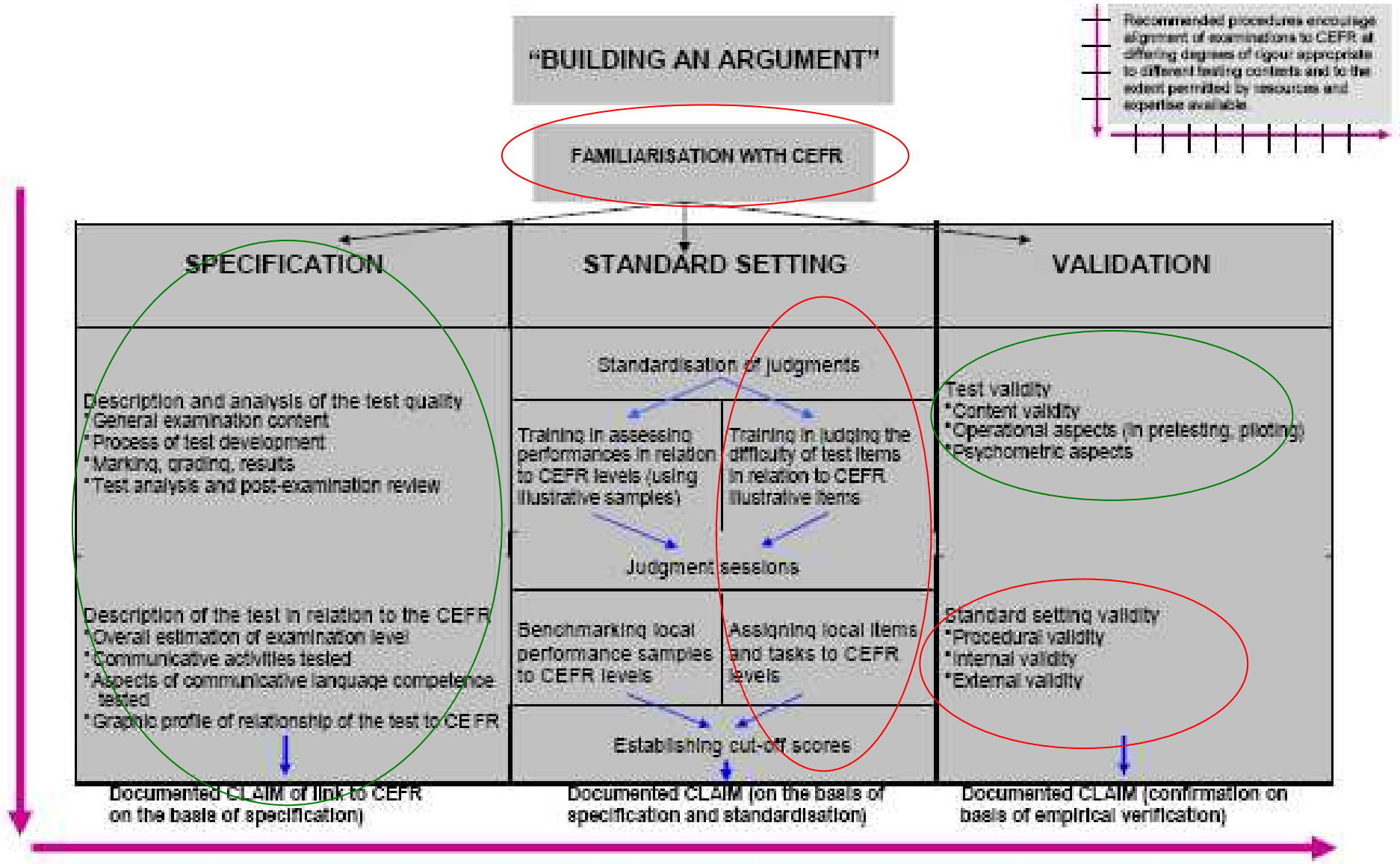
The test

- Bi-level listening and reading tests in English for levels 2 and 3
- Developed jointly by Estonian and Latvian testing teams
- Pre-tested in several central and eastern European countries
- First administered in spring 2007
- A new version developed every year
- **Problems**:
  - **Cut scores for levels 2 and 3**
  - **Equivalency of the cut scores across test versions**

Standard setting

- Establishing one or more cut scores on a test
- "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek 1993: 100)

Figure 2.2: Visual Representation of Procedures to Relate Examinations to the CEFR



Test statistics (version1)

	Est	Lat	All
No of test-takers	190	133	323
No of items	39	39	39
No of tasks	7	7	7
Mean score	25.9	21.9	24.3
St dev	8.02	8.06	8.27
Reliability (α)	0.9	0.89	0.9
SEM	2.51	2.66	2.59

Item statistics (version 1)

		Est	Lat	All
Difficulty	**Mean**	**66%**	**56%**	**62%**
	Min	37%	24%	33%
	Max	94%	89%	91%
Discrimination	**Mean**	**0.46**	**0.44**	**0.46**
	Min	0.19	0.20	0.20
	Max	0.62	0.63	0.58

The study

- Conducted over a three-day period during a Pan-Baltic neeting of language testers in October 2008
- Judges: 11 (10) language testers from
  - Estonia (4)
  - Latvia (3 $\Rightarrow$ 2)
  - Lithuania (2)
  - Denmark (1)
  - SHAPE (1)

Familiarisation: rating descriptors

- "The form includes 39 descriptors taken from the STANAG 6001 reading proficiency levels and listed in random order. In the column next to each descriptor, enter a number (0, 1, 2, 3, 4 or 5) corresponding to the STANAG 6001 level to which you think the descriptor belongs."
- **Mean correlation**: 0.90; St dev: 0.036
- **Range**: 0.86-0.98
- **All 11 judges**: 5 descriptors
  - Demonstrates understanding of abstract concepts in texts on complex topics (which may include economics, culture, science, technology) as well as his/her professional field. (22)
- **Fewer than 50% of the judges**: 7 descriptors
  - Comprehension is not dependent on subject matter. (17)

Standardisation of judgements

- Training
  - No validated illustrative samples available
  - Items used in single-level tests in earlier years in Estonia
- Judges take the test; answers discussed

Judgement session 1

- '*At what STANAG 6001 level can a test taker already answer the following item correctly?*'
- Correlation (judgements/p-values): -0.44 (-0.31-0.53)
- **No of items per level**:
  - L1 = 4.7 (0-9)
  - L2 = 17.4 (12-22)
  - L3 = 16.7 (13-20)
  - L4 = 0.22 (0-2)
- **Cut scores**:
  - L2 = L1 + L2 items = 22.1
  - L3 = L1 + L2 + L3 iterms = 38.8

What does it mean to be at a level?

- **"REDS"** (Ray Clifford, Martha Herzog)
  - **Random** (no visible evidence) = 0% to 35%
  - **Emerging** (some limited evidence) = 40% to 50%
  - **Developing** (present, inconsistent) = 55% to 65%
  - **Sustained** (consistent evidence) = 70% to 100%
- L2 = 70% of 22.1 = 15.47
- L3 = 70% of 38.8 = 27.16

Judgement session 2

- *'Think of 100 examinees who are exactly on the border between two consecutive **STANAG 6001** levels. **Estimate** how many of them will answer each item correctly and **write** the number (a whole number between 0 and 100) in the corresponding **column**.'*
- **Tentative cut scores**:
  - L1/2. mean 8.5, SD 3.61, range 4.55-14.70
  - L2/3: mean 26.16, SD 4.18, range 19.05-31.58
  - L3/4: mean 36.16, SD 2.28, range 32.45-39.00

Cut scores

	L2	L3
Basket procedure	22	39
Basket with REDS	15	27
Modified Angoff	9	26
Previously established	17	30

Lessons learnt

- Familiarisation with the scale is crucial, even with the experienced testers
- Training needed
  - With validated illustrative samples
  - In using the standard setting procedures
- One judgement session with one method not enough
- Basket procedure problematic
  - Judges opinions of item difficulty do not match the reality ⇒ Is it possible to improve with practice?
  - What does it mean to be at a level?

Thank you.

Sources

- Cizek, Gregory J. & Michael B. Bunch (2007) *Standard Setting. A Guide to Establishing and Evaluating Performance Standards on Tests.* Thsousand Oaks: SAGE Publications.
- Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*: http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp