Standardising the assessment of writing in English across nations in the military

Ülle Türk
Language Testing Unit
Estonian Defence Forces

The 6th Annual EALTA Conference,
Turku, 4-7 June 2009

Outline

- Background
- Aims of the project
- Procedure
- Results
- Conclusions

Background

- NATO member states – STANAG 6001 describing foreign language proficiency
- Each country responsible for designing own tests
- 4 skills – SLP (LSRW)
- Small teams
- Small test populations
- Little cooperation and standardisation
- Last year's EALTA conference – meeting to agree on future cooperation

Aims of the project

- To select a number of sample scripts that
  - have been written in response to a variety of prompts
  - demonstrate English language proficiency at STANAG levels 1-4
  - could later be used as
    - benchmark performances in assessing writing and in rater training
    - sample performances for teachers and test takers
- To study the possibility of carrying out standardisation via email.

Participants

- Denmark (1)
- Estonia (5)
- Latvia (4)
- Lithuania (3)
- SHAPE (2)
- Slovenia (5)

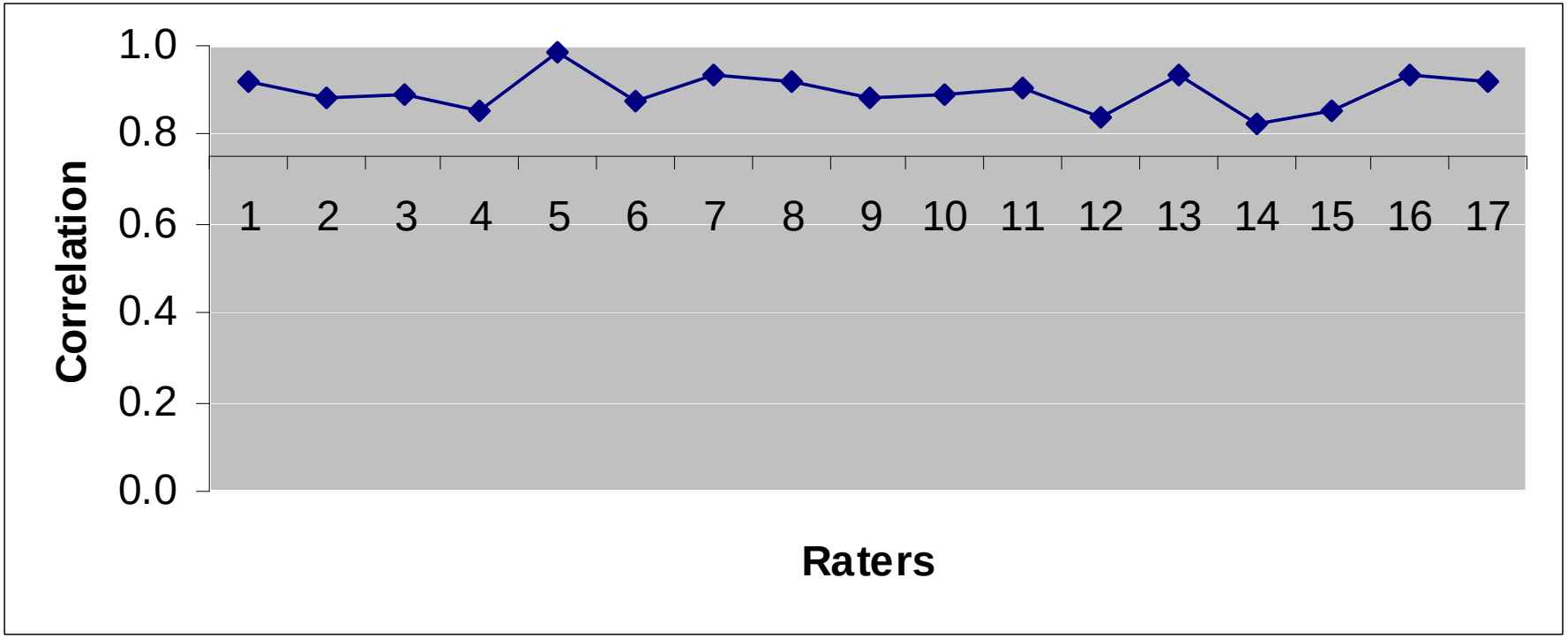
Procedure

- Each participating country/institution selects 4 scripts, including problem scripts
- Scripts are collected, coded and sent to all participants
- Scripts are marked following the procedures established in each country
  - STANAG level descriptors used
  - Weak, standard and strong performances at each level identified
- Comments provided
- Results analysed

STANAG 6001 levels: writing

Level 0	No functional writing ability
Level 1	Can write to meet immediate personal needs.
Level 2	Can write simple personal and routine workplace correspondence and related documents such as memoranda, brief reports, and private letters, on everyday topics.
Level 3	Can write effective formal and informal correspondence and documents on practical, social, and professional topics.
Level 4	Can write the language precisely and accurately for all professional purposes including the representation of an official policy or point of view.
Level 5	Writing proficiency is functionally equivalent to that of a well-educated native writer.

Raters rating descriptors



Mean correlation: 0.89 (SD =.04)

Range: 0.83 (R14) to 0.98 (R05)

Task types

- 27 scripts
  - 12 letters
  - 4 (+ 5) essays
  - 1 report
  - 1 memorandum

  A first draft of a lecture (2)
  Paper for a newsletter (1)
  Paper/ letter/essay (1)

Rating scripts

- **Task:**
- If the script were written for a STANAG 6001 test in your country/ institution, which level would it be awarded?
- Do you consider it a weak, standard or strong performance at the awarded level?
- Why?

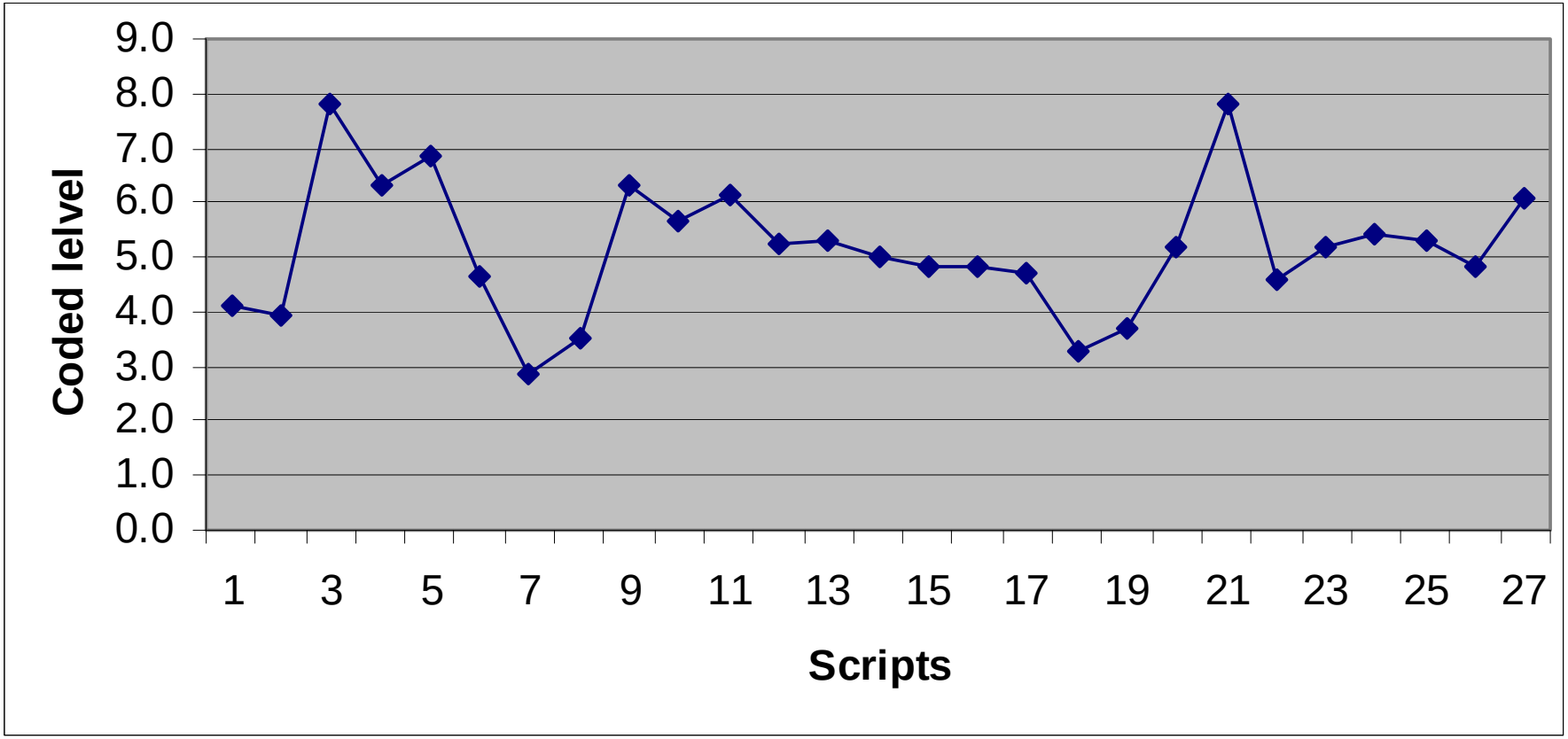
Analysis and results

- No scripts at L4
- Coding:
  - L1 weak = 1
  - L1 standard = 2
  - L1 strong = 3
  - L2 weak = 4
  - L2 standard = 5
  - L2 strong = 6
  - L3 weak = 7
  - L3 standard = 8
  - L3 strong = 9

Results

- Mean rating: 2.8–7.8 (St dev: 0.00-1.47)
- 1-3 (L1): 1 script
- 4-6 (L2): 24 scripts
- 7-9 (L3): 2 scripts
- 15 scripts (55.6%) – agreement on the level, though usually not on whether it is weak, standard or strong performance at that level

Mean ratings



Mean rating: 5.2 (SD = 1.44)

Ratings by country

Country / institution	Mean	St dev
C01	5.3	1.27
C02	5.0	1.26
C03	5.8	1.30
C04	5.1	1.64
C05	5.0	1.70
C06	5.0	1.48

Correlations between country ratings

C02	0.741				
C03	0.670	0.761			
C04	0.644	0.768	0.746		
C05	0.584	0.692	0.761	0.585	
C06	0.730	0.786	0.845	0.715	0.670
	C01	**C02**	**C03**	**C04**	**C05**

Task types and ratings

Task Type	Range	Mean	St dev
Letter (12)	2.8-6.1	4.6	1.03
Essay (9)	3.7-7.8	5.4	1.33
Other (6)	4.7-7.8	5.9	1.21

Way forward

- 1 L1 script, 12 L2 scripts, 2 L3 scripts
  - Analysis of scripts ⇒ good benchmarks?
- Collecting more scripts, particularly at L3
- Scripts based on a variety of task types
  - Did we start at the wrong end?
- Looking at scripts that caused disagreement
  - Can we reach agreement?
  - What features make them problematic?
- Expanding the circle to include more countries