

Gergely Dávid:

**Building a Case for Euro
Examinations**

A case study

The Structure of the Study

- Part 1: Linking the Grammar and Vocabulary paper to the CEFR
 - Management decision to select GramVoc only
 - North: “The most difficult task you could pick”
- Part 2: Linking Euro Examinations to the CEFR
 - Levels B1, B2, C1 of the General English suite
 - Level B2 of the Pro (ESP) suite

Productive orientation of CEF

General Linguistic Range B2

- Has a sufficient range of language to be able to **give** clear descriptions, **express** viewpoints and develop arguments **without much conspicuous searching for words**, using some complex sentence forms to do so.
- Lack of yardsticks for a test like GramVoc
 - Van Ek and Trim volumes not useful.
 - CEF provides description of 15 categories, but without level specification (pp. 108-117).

Process and Audience for the Case Study

- Four phases of action according to the Manual
 - *Familiarization*
 - *Specification*
 - *Standardization (of judgements)*
 - *Empirical validation*
- Working with the team of full-time item writers as “holders of standards”

The Familiarization Phase

- Interesting conclusions from a survey of familiarity with the CEF scales
 - Descriptors from 15 scales, 133 „items”, as in a test
 - Checking distributions, Chi-square, etc.
 - Initial facility value of responses: 0.41 Is it low? How low?
 - 16/133 descriptors nobody got the level right. Significantly more B1 descriptors.
 - In cases uncertainty, tendency to place level of descriptor higher than in CEFR. Lower Euro standards? Leniency?
 - Leniency not related to any of the scales.
- Familiarization in Manual a formative activity
 - For the revision: a survey of how well the CEFR has been internalized which could be used as baseline for subsequent steps.

The Specification Phase

a qualitative content audit

- Two lines of work
 - Elucidating item-writers' concepts
 - Expert analysis of what content (item focuses) actually goes into the test, on the basis of the scope, the gradation and stability between 2 consecutive test administrations.

The Specification Phase

Elucidating item-writers' concepts

- Item writers select “best task” for each task type and level.
 - *What is it that makes this task the best representation of the level for you?*
- Collation of descriptors that belong to the same level
 - Amounts to a “horizontal reading” of the CEFR
- Item-writers' concepts broadly match CEFR. Apparently better overall results than in familiarization phase.

The Specification Phase

Expert analysis of item focuses

Evidence of construct under-representation or construct-irrelevance?

- 2 experts identify item focuses, then jointly finalize classification of items acc. to 15 CEFR categories (pp. 108-117)
- Tests of relatedness in crosstabs
 - Overall comparisons
 - Separate analyses of administrations
 - Stability: comparison of same level across administrations
 - Tests of item complexity

Results Specification Phase

- Statistical analysis of expert classifications
 - Distribution of focuses related to task type and author (text), but not related to level and administration.
 - Results similar when two administrations at the same level were compared.
 - The number of focuses increases by level, but the distribution of the 15 focus categories is essentially not different at the three levels.
- Good coverage of the CEFR
- For the revision:
 - The qualitative content audit may be appended by quantitative analyses, deflecting the interpretation that empirical validation is something that starts at a later stage in the linking process, as is suggested by the name of phase 4.

The Standardization of Judgements: Line 1

- Investigating the gap between Local Euro standards and the CEF standards
 - Item-writers identified descriptors on the basis of collations the content of which exceeded local standards
 - Tabulation and qualitative analysis of responses. “History” of descriptors taken into account.
- Moving up the CEFR scale, the „gap” does not widen. Most conspicuous at B2, but less considerable if descriptor history is accounted for.
 - *Why do the uncalibrated descriptors represent a higher level of requirements than those that went through it?*

Standardization of Judgements

Line 2: Video rating conference

- *CEF Performance Samples*: Link to North's rating conference (1996/2000)
 - A second-best option and problems
- Encouraging results
 - Reliability of scale use: Chronbach's Alpha 0.96
 - Kendall's W: 0.85

Standardization of Judgements

Line 3: Standard Setting

- With about 20 scripts per level for both test 2003 and 2004
- An examinee-based method. Scripts carefully chosen, arranged in decreasing order of ability
 - “Overfitting” candidates
 - Info about items
- Rating done twice bearing in mind
 - Round 1: conventional Euro standards, Kendall’s W ranged 0.8 - 0.83
 - Round 2: CEF standards, Kendall’s W ranged 0.75 - 0.79
- Least accessible chapter in the Manual

Empirical Validation Phase

- Internal validation: item analyses
 - Independent and joint analyses of same level papers
 - Investigation of item and person fit (statistics)
- External validation
 - Using standard setting data from the standardization phase as ratings of the three tests
 - Calibrate (MFRM) overall test difficulties from ratings
 - Anchor item means of independent analyses to calibrated overall test difficulties
 - Use a corrected version of North's scale
 - Compare cutoffs obtained in this way with conventional Euro cutoffs.
- Almost everything has been achieved in terms of internal validation, but alignment with
 - B2 Euro standards was good
 - B1 Euro standards “sort of” acceptable
 - C1 Euro standards was poor
- Inconclusive because results from different Euro papers are added up and the GramVoc was only one of these.

Stage 2: Rethinking experiences and problems

- The legal mandate in the Hungarian context
- Problems:
 - The scarcity of enough calibrated performance samples
 - What is there is not always accompanied with performance data to make the sample transparent/accountable enough
 - The problem of calibrated samples not based on comparable measurement scales: In essence a question a single or multiple references

A2 Dialang scale

Level / Item ID	Item theta	Discrimi- nation	Facility value
A1 (theta below -.56)			
004445	-.83	5	.989
004183	-.60	4	.951
004144	-.57	4	.944
A2 (theta -.56- to -.22) Length: 0.34 logits			

CEFR scale by North

C2	Mastery	3.90 to 4.77	0.87
C1	Effectiveness	2.80 to 3.90	1.10
B2	Vantage	0.72 to 2.79	2.08
B1	Threshold	-1.21 to 0.71	1.92
A2	Waystage	- 1.22 to - 3.23	2.01
A1	Break-through	-3.24 to -4.29	1.05
Swiss	„Tourist”	-4.30 to -5.39	1.09

Scale lengths and scale properties A2-C1

Scale	Min	Max	Length (logit)	Raw score scale
Dialang	-0.56	2.01	2.57	0 -- 1
CEFR (North)	-3.23	3.89	7.12	0-1-2-3--4

How to give an adequate response to these challenges?

Choose a single point of reference

- North (2000) and Schneider's (1999) measures as basis of the scales in CEFR: a well documented piece of work

Making a “direct” linking to CEFR

Recommendation by the Manual is “indirect”:

Own items → CEFR reference items → CEFR

Own performances → CEFR descriptors

Familiarization as replication

Survey with English, German and Hungarian CEFR descriptors.

Reproduce CEFR scale locally, MFRM analysis

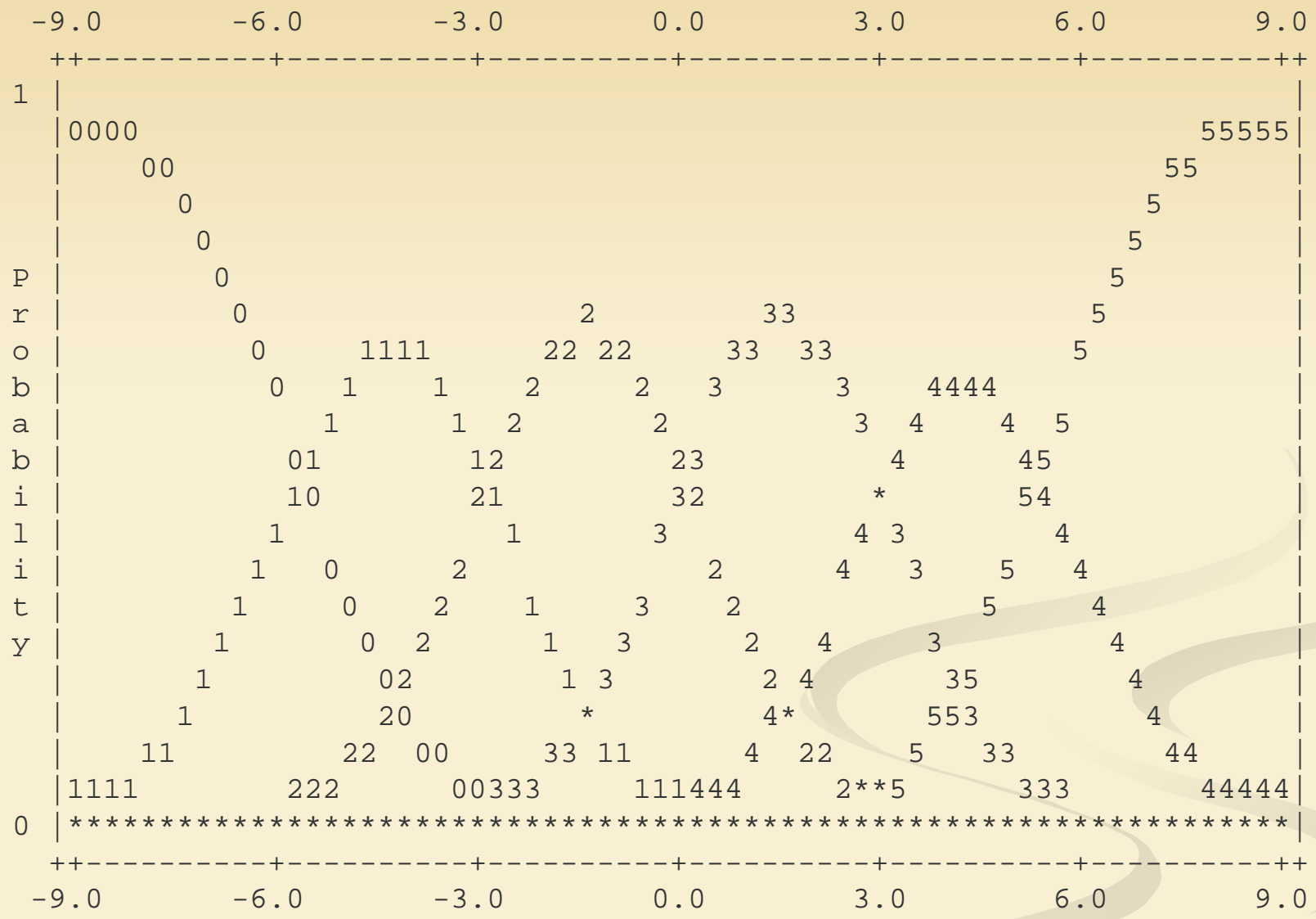
				4.61		≈ C2
			2.47		≈C1	
		0.33		≈B2		
			≈B1			
		≈A2		Stairs, steps = thresholds		
	≈A1					

Skipping standardized samples

Anchor a batch of the best descriptors from North
Obtain a CEFR sensitive overall A1-C2 scale

				3.04		C2
			0.08		C1	
		-2.85		B2		
			B1			
		A2		Item = descriptor		
	A1					

Probability Curves



Standard setting

- Identify cca. 70 appropriate candidates on the basis of analyses on the tests to be linked
- Judges rate candidate performances on all 6 papers of the test to be linked
- MFRM analyses with anchored B1, B2 and C1 steps with values from the replication of the CEFR scale
 - Information about how successive levels compare!

3.04

0.08

-2.85

C1 candidate 001

C1 candidate 002

C1 candidate 003

B2 candidate 001

C1 candidate 004

B2 candidate 001

B2 candidate 002

C1 candidate 005

B2 candidate 003

B2 candidate 004

B2 candidate 005

B2 candidate 006

B1 candidate 001

B2 candidate 007

B1 candidate 001

B1 candidate 002

B2 candidate 008

B1 candidate 003

B1 candidate 004

A2 candidate 001

Deductive logic

- If North and Schneider's work can be relied on ...
- measures of CEFR descriptors can be locally reproduced, using good descriptors and ...
- A1-C2 thresholds linked to the CEF may be determined and
- the CEFR sensitive thresholds may be used to set standards, linking not items or tasks to the CEFR, but the cutoffs, which is the basis of certification.

Thank you.

