

Benchmarking of Videotaped Oral Performances in Terms of the CEFR

Gunter Maris, José Noijons, Evelyn Reichard

EALTA colloquium on standard setting and its relevance to the CEFR, Athens, HAU, 2008



Relevance of the CEFR

- Levels of competence in terms of *descriptors*
- Indicators of what L2 users *can do*
- Indication of *minimum* performance levels

CEFR & the Need for Clarification

- What do some descriptors mean exactly?
- Differences between descriptors at different levels
- Use of linguistic jargon
- Lack of detail, precision, context

Benchmarking Project

- Benchmarking of oral performances
- Cooperation with Institute for Curriculum Development (SLO)
- Aim: to illustrate CEFR performance levels
- Videotaped performances based on can-do statements in the ELP

Production of Tasks and Videos

- Production of tasks
 - Ranging from A1- C2
 - Emphasis on A2 - B2
- Production of videotaped performances
 - Students (12-18 years old) and adults
 - Native speakers as antagonists
 - Non-testing environment

General aim: not to judge persons but to collect performances



Selection of Performances for Benchmarking

- Pre-estimation by Cito and SLO staff
- Selection includes performances above and below (pre-estimated) standards
- Total number of selected performances: ca 60
- Sets of 20 performances on three CD-ROMs

Selection of Judges

- Participants at ALTE and EALTE conferences
- 14 countries participating
- CD-1: 26 judges, CD-2: 38 judges, CD-3: 18 judges

Instruction to Judges

- Detailed instructions (cf Manual)
- No control over judgement process

But:

- Reputable organisations
- Voluntary participation
- General agreement with aims

Judgment Procedure

- What level is the task at?
- What level is the performance at?
 - Using the criteria going with the judged level of the task
 - Criteria taken from the CEFR

Description of Task

Descriptor	Task Description
I can express in a polite way my opinion, my conviction, agreement and aversion	<p>You got talking to someone who is complaining about the high taxes he/she has to pay. You think it is only fair to pay taxes because</p> <ul style="list-style-type: none">● it is a way to support those who have less money than you have;● you think it is important to have such things as motorways and a police force.

Rating Scheme (1)

Criterion	Performance I - 1	Performance I - 2	Performance I - 3	Performance I - 4	Performance I - 5	Performance I - 6	Performance I - 7	Performance I - 8
Level of TASK	A1 A2	A1 A2	A1 A2	A1 A2	A1 A2	A1 A2	A1 A2	A1 A2
	B1 B2	B1 B2	B1 B2	B1 B2	B1 B2	B1 B2	B1 B2	B1 B2
	C1 C2	C1 C2	C1 C2	C1 C2	C1 C2	C1 C2	C1 C2	C1 C2
+ = Performance <i>above</i> task level o = Performance <i>at</i> task level - = Performance <i>below</i> task level								
Range								
Accuracy								
Fluency								
Interaction								
Coherence								

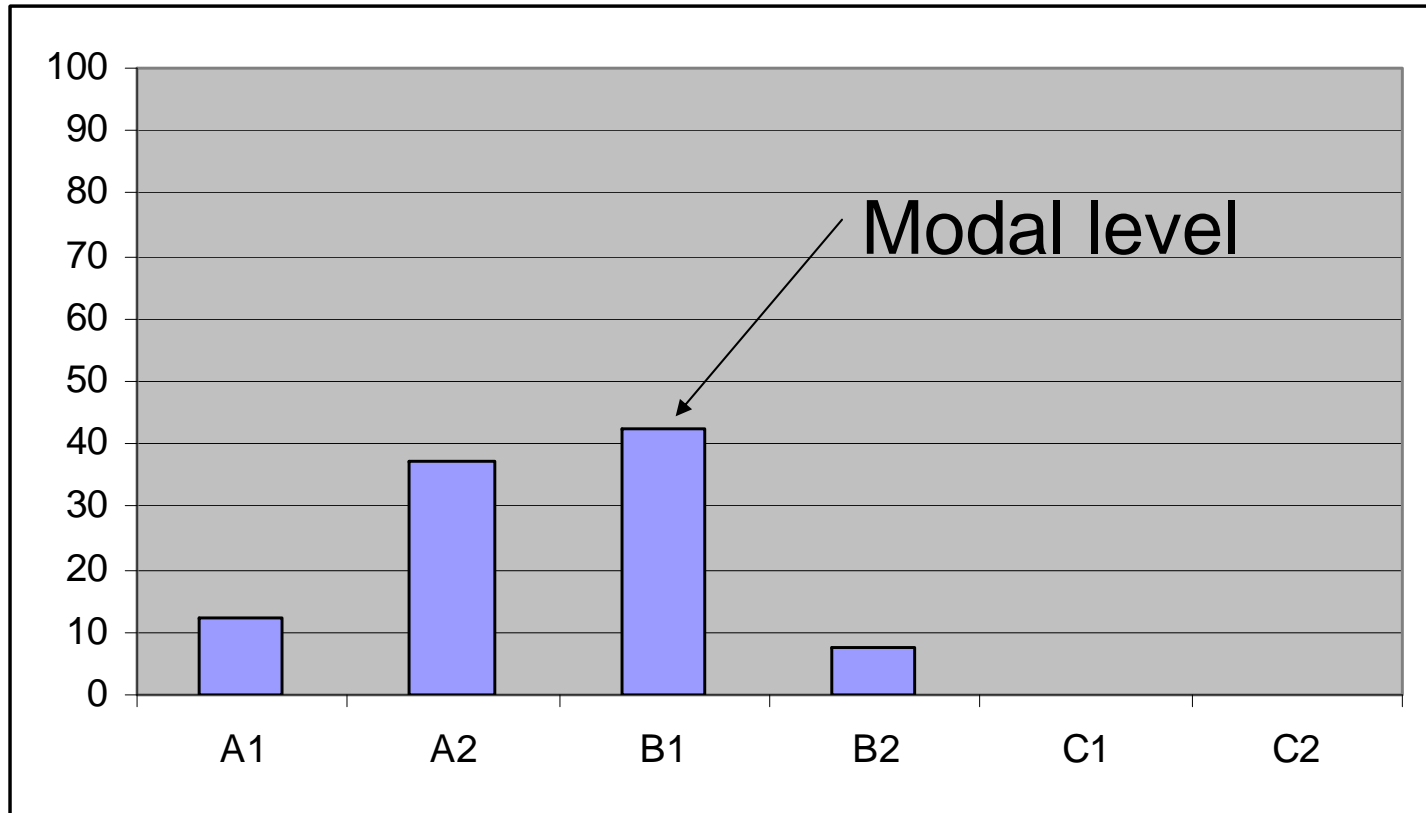


Rating Scheme (2)

Level of task	A1	A2
	B1	B2
	C1	C2
+ = Performance above task level o = Performance at task level - = Performance below task level		
Range		
Accuracy		
Fluency		
Interaction		
Coherence		



Judging task levels



- Agreement generally low
- Coarser levels improve agreement, but just the levels A, B, and C is not enough

Judging performances

Judge	Rated level of task	Rated level of performance	Modal level					
			CEFR-level					
			A	A	B	B	C	C
			1	2	1	2	1	2
1	B1	above	0	0	0	1	1	1
2	B2	at or below	1	1	1	1	0	0
3	B2	above	0	0	0	0	1	1

Judging performances

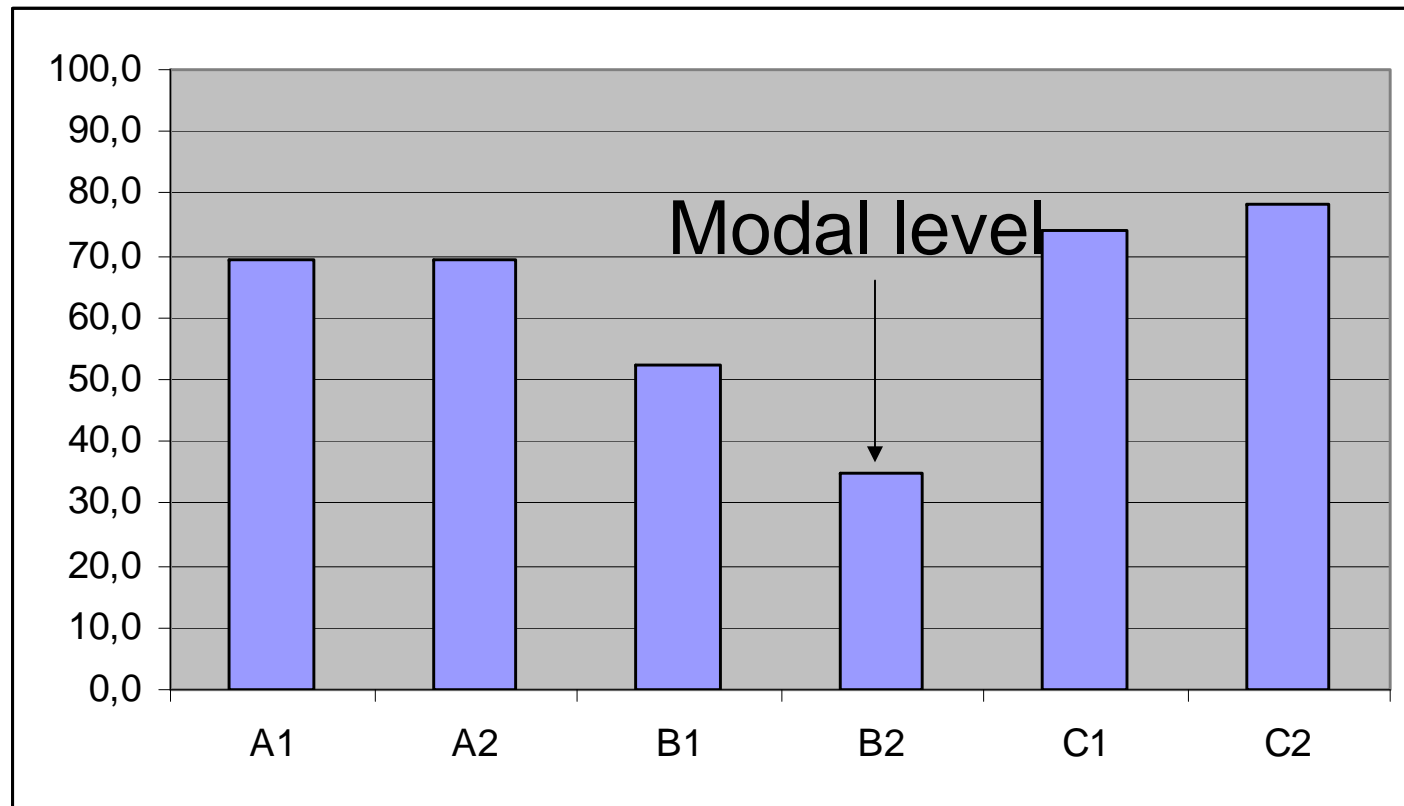
Judge	Rated level of task	Rated level of performance	CEFR-level					
Exclusion percentage stays the same if we add a <i>left</i> neighbour to B2			A	A	B	B	C	C
			1	2	1	2	1	2
1	B1	above	0	0	0	1	1	1
2	B2	at or below	1	1	1	1	0	0
3	B2	above	0	0	0	0	1	1

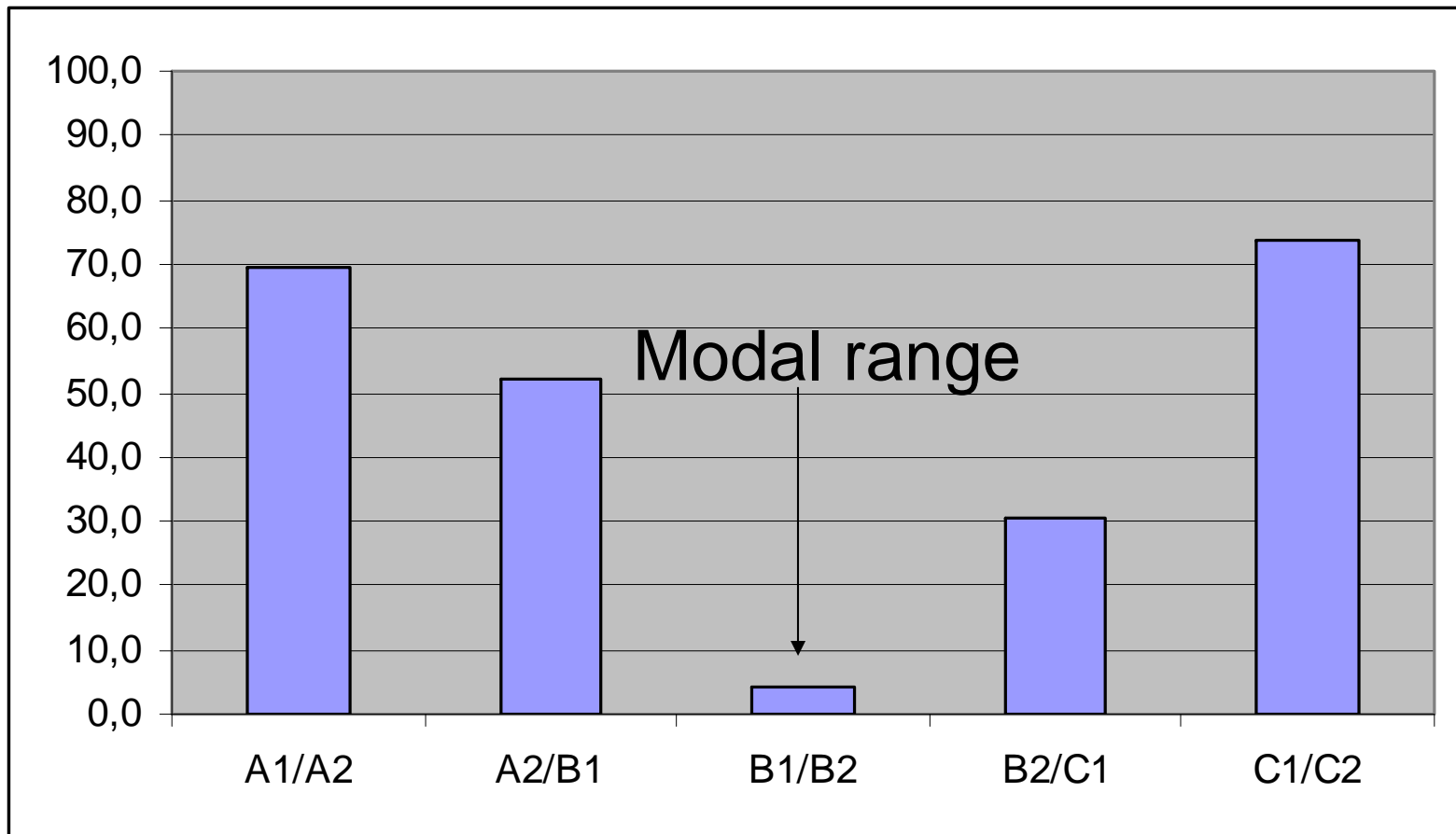
Judging performances

Judge	Rated level of task	Rated level of performance	CEFR-level					
Exclusion percentage improves If we add a <i>right</i> neighbour to B2			A	A	B	B	C	C
			1	2	1	2	1	2
1	B1	above	0	0	0	1	1	1
2	B2	at or below	1	1	1	1	0	0
3	B2	above	0	0	0	0	1	1

- Judges *exclude* levels
 - They tell us which performance levels do not apply
- 3 modal levels (B2, C1, and C2) with equal exclusion percentages
- 1 modal range (B2/C1) with exclusion percentage zero

A typical performance (CD1- III 3)





- Agreement generally low
- Coarser levels improve agreement, but just the levels A, B, and C is not enough

Conclusions

- Lower levels of agreement for fine-grained level distinctions (A1, A2, B1, B2, C1, C2)
- Acceptable levels of agreement when levels are aggregated (A, B, C)
- Relevance to stakeholders?
- Judgements on multiple performances give higher agreement on fine-grained levels
- Danger of focussing on person rather than performance
- Benchmarking of oral tasks prior to testing

