UNIVERSITY of CAMBRIDGE
ESOL Examinations

Experts in Language Assessment

Quantitative and Qualitative Methodologies in the Development of Rating Scales for Speaking

Evelina D. Galaczi
University of Cambridge ESOL Examinations

EALTA, June 2009

ALTE

PART OF THE CAMBRIDGE ASSESSMENT GROUP

Cambridge ESOL rating scales for speaking

- Revised in 2006-2007
- Live roll-out in December 2008
- Why revise?
  - Last revision in Dec ’02
  - Since then:
    - Questionnaires to examiners ➔ Happy with the scales
    - BUT scale truncated ➔ Little use of bottom end of scale

Levels and bands

- Five levels:
  - A2 to C2 of CEFR
- Ten bands at each level
  - Labelled ‘0’, ‘1’, ‘1.5’, ‘2’, …. ‘4.5’, ‘5’
- Stacked up into a common scale

Cambridge ESOL Scale for Speaking				
CPE 5.0				
CPE 4.5				
CPE 4.0				
CPE 3.5				
CPE 3.0	CAE 5.0			
CPE 2.5	CAE 4.5			
CPE 2.0	CAE 4.0			
CPE 1.5	CAE 3.5			
CPE 1.0	CAE 3.0	FCE 5.0		
	CAE 2.5	FCE 4.5		
	CAE 2.0	FCE 4.0		
	CAE 1.5	FCE 3.5		
	CAE 1.0	FCE 3.0	PET 5.0	
		FCE 2.5	PET 4.5	
		FCE 2.0	PET 4.0	
		FCE 1.5	PET 3.5	
		FCE 1.0	PET 3.0	KET 5.0
			PET 2.5	KET 4.5
			PET 2.0	KET 4.0
			PET 1.5	KET 3.5
			PET 1.0	KET 3.0
				KET 2.5
				KET 2.0
				KET 1.5
				KET 1.0

Assessment Criteria

- **Grammar and Vocabulary**
  - Accurate and appropriate use of a range of **grammatical forms and vocabulary.**
- **Discourse Management**
  - Ability to link utterances together to form **coherent speech**, without undue hesitation.
- **Pronunciation**
  - Ability to produce **intelligible** utterances (stress, intonation, individual sounds).
- **Interactive Communication**
  - Ability to take an active part in the **development of discourse** (initiating and responding, maintaining the interaction).

Wording of performance descriptors

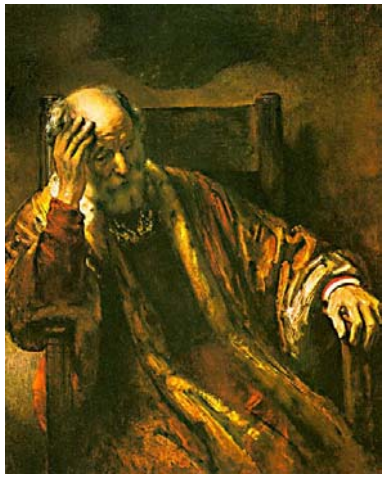
- Positive
- Definite and clear
- Brief
- Independent

- A “can do” approach
  - e.g. ‘*Poor use of stress, rhythm and intonation frequently obscures the message*’ ➔ ‘*Is mostly intelligible, despite limited control of phonological features’*
- Jargon-free
  - e.g. *Turn-taking strategies* ➔ “*initiating*”, “*responding*”, “*development of the interaction*”
- A descriptor longer than a two-clause sentence cannot realistically be referred to during the assessment process (CEFR Manual)
- Meaning should not be relative to other descriptors in the scale

Scale Development Methodologies: Mixed methods

- Commitment to validation prior to use
  (Milanovic et al, 1996; Taylor, 2000; ffrench, 2003)

- "Armchair approach"
  - Expert judgement primary



- Empirically-driven approach
  - Data primary: quantitative and qualitative analyses of candidate and rater performances; stakeholders' input

Intuitive Phase

- **Content experts**:
  - Review of current ESOL practice in light of the literature
- Set out **design principles** for revised assessment scales
  - Grammar and Vocabulary split up at C1 and C2
  - No Discourse Management at A2
  - Pronunciation descriptors identical at C1 and C2
- Sub-criteria for each assessment criterion/scale
  - Which assessment criteria at which level?

Sub-criteria

- **Grammar and Vocabulary**
  - Control
  - Range
  - Appropriacy
- **Discourse management**
  - Extent
  - Relevance
  - Coherence
  - Cohesion
- **Pronunciation**
  - Intonation
  - Stress
  - Individual sounds
  - At A2: Intonation not assessed
- **Interactive Communication**
  - Initiating
  - Responding
  - Development
  - At A2: Responding; Support required

Empirical Phase

- **Scaling "Mix'n'match" exercise**
  - Rank ordering of descriptors
- **Verbal protocol** study with raters
- **Discourse analysis** of candidate language
  - Identification of discourse features associated with differently marked performances
- **Extended trial** using Multi-facet Rasch analysis to assess performance of revised criteria

Scaling “Mix’n’Match” study

- 64 draft descriptors distributed to 31 oral examiners
  - Examiners at all levels of Speaking Examiner framework
- Descriptors matched to a level on the Cambridge ESOL scale for speaking
- Relative “difficulty” of the descriptors estimated through FACETS and based on examiner ratings
- Examiner ratings compared with the levels intended by the scale developers
- Consistency of examiner performance investigated

Findings

- Broad agreement between intended levels and examiner ratings

  ➔ encouraging evidence for the validity of the revised scales

- For some descriptors the ratings contradicted the intended levels

  ➔ Need for rewording or careful exemplification

Intended Level	*Descriptor*	*Fair Average*
A1	Has a basic vocabulary of isolated words and phrases.	1.15
A1	Shows only limited control of a few grammatical forms.	1.85
A2	On the whole, uses simple grammatical forms to convey meaning.	2.81
A2	Vocabulary is adequate to talk about everyday situations.	2.99
C1	**Vocabulary is adequate to deal with a variety of language functions.**	**3.42**
B1	Has a good degree of control of simple grammatical forms.	3.52
C1	Uses a range of grammatical forms.	3.52
B1	**Uses a range of vocabulary when talking about everyday situations.**	**3.65**
C1	Uses appropriate vocabulary, but may lack flexibility.	3.77
B2	Has a good degree of control of simple grammatical forms, and attempts some complex grammatical forms without obscuring meaning.	3.82
C2	**Uses a range of vocabulary when dealing with a variety of language functions.**	**4.13**
C2	Maintains control of a range of grammatical forms.	4.38
B2	**Vocabulary is adequate to give and exchange views on a range of topics.**	**4.47**
C1	Has a good degree of control of simple and complex grammatical forms.	4.71
C2	Uses a wide range of grammatical forms.	4.92
C2	Uses vocabulary with flexibility to express meanings.	5.12
C2+	Uses a wide range of vocabulary to deal with a variety of language functions.	5.32
C2+	Maintains control of a wide range of grammatical forms.	5.62
C2+	Uses a wide range of grammatical forms with flexibility and ease.	6.25
C2+	Uses vocabulary with flexibility and ease to express meanings.	6.31

VERBAL PROTOCOL STUDY

- What do raters pay attention to when using the revised descriptors?
- 8 oral examiners
- Awarded marks to videotaped performances at levels A2 to C2
- Completed a detailed questionnaire

Findings

- Balanced references to all 4 criteria across the tests
  - Grammatical accuracy not main driver
- Greater specificity, transparency, brevity and clarity of descriptors
  - “It is clearer what is expected at each band of each level, with less subjective interpretation.”
- Greater ease of processing and applying the descriptors
  - “Easier to apply perhaps just because they aren’t quite so wordy.”
- Use of positively worded descriptors
  - “The greater emphasis on positive achievement is welcome. One is encouraged to concentrate on what the candidate is capable of.”
- Need for greater clarity with some of the descriptors

Feedback from oral examiners

- “What constitutes ‘a good degree of control’, ‘limited control’?”
- “The ‘range’ aspect was quite difficult to judge.”
- “What is meant by ‘some complex grammatical forms’?”

➔ **GLOSSARY OF TERMS**

➔ **CAREFUL EXEMPLIFICATION IN EXAMINER TRAINING**

Feedback from oral examiners

- “I found that using ‘control’ rather than ‘accuracy’ forced assessments to look at grammatical forms over a number of utterances rather than just focusing on individual mistakes.” (GV)
- “The removal of ‘incoherent’ and ‘coherent’ and their replacement with notions of ‘repetition’ and ‘digression’ are easier to match to performance.” (DM)
- “A judgment on intelligibility is easier to apply.” (PR)
- “I like the reference in IC to ‘linking contributions to those of other speakers’. This is useful and immediately comprehensible.” (IC)

Marking trial

- Aim: to provide statistical evidence of the adequate functioning of the revised scales
- 12 raters
- 19 full tests and 51 test parts at A2 to C2; range of nationalities
- Multi-facet Rasch analysis

Findings

- Examiner harshness/leniency
  - Different levels, but within acceptable parameters
  - ➔ Raters interpreting scales in similar ways
- Internal consistency of examiners
  - Within acceptable parameters (Infit Mean Square 0.5 – 1.5)
- Examiner agreement
  - Moderately high
- Comparison between revised and current marks
  - A wider range of the scale used, especially at the lower end

An ecological system

Revised assessment scales

An ecological system

What came next?

- Communication with stakeholders
- Examiner training and standardisation:
  - On-line and Face-to-face
  - Challenge: illustrating distinctions between adjacent levels

Thank you.

For more information:

Galaczi.E@cambridgeESOL.org.uk