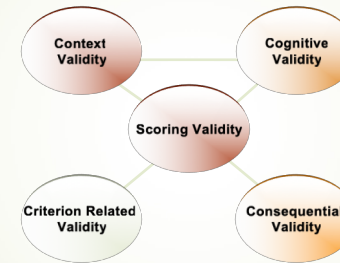


Testing Speaking Skills

Based on the lecture by Tony Green given for the CEFR-J development team

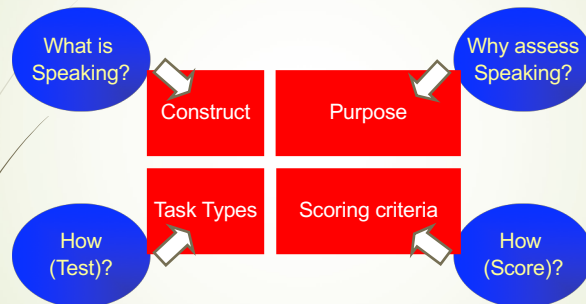
2

A framework for test validation



3

Key questions



4

What is Speaking?

Socially situated, often reciprocal oral interaction with another or others
It requires...

- phonological control
- lexico-grammatical knowledge
- knowledge of discourse
- pragmatic awareness
- strategic competence

These are different from written language

5

Speaking Processes

Cognitive dimension

- The knowledge factor
 - multiple competencies (Bachman and Palmer 1996)
- Processing factor
 - established procedures enable fluent performance with online planning reduced to acceptable amounts and timing

6

What is proficient speech?

- | | |
|--|--|
| - Shorter and less complex units of speech | + Longer and more complex (e.g. more embedded) units of speech |
| - More errors per unit of speech | + Fewer errors per unit of speech |
| - Less and more limited use of cohesive markers | + More, and more varied, use of cohesive markers |
| - Limited to use of common words | + Use of more sophisticated and idiomatic vocabulary |
| - Pauses linked to language search | + Pauses linked to content search |
| - Pauses within grammatical constituents | + Pauses between grammatical constituents |
| - More silent pause time | + Less silent pause time |
| - Shorter 'runs' of speech between noticeable pauses | + Longer runs of speech between noticeable pauses |
| - Speed of delivery noticeably below native speaker rate | + Speed of delivery not noticeably below native speaker rate |

(Tonkyn and Wilson, 2004)

7

What is proficient speech?

Social dimension

Conversational rules and moves

- Opening and closing
- Turn taking
- Holding the floor
- Interrupting

Register

- Status and roles
- Degrees of formality
- Politeness

Communication strategies

- Paraphrase
- Avoidance
- Seeking help
- Asking for repetition

Language functions

- Macro-functions
 - transmitting information
 - social interaction
 - persuading others
 - learning and thinking
- Micro-functions
 - inviting
 - apologizing
 - warning

Speech acts

- Locutionary
- Illocutionary
- Perlocutionary

8

Speaking Tasks

What speakers do with language in social contexts

Brown and Yule (1983) two language functions

- ▲ Interactional
 - ▲ To develop and maintain relationships

- ▲ Transactional
 - ▲ To convey content

Bygate (1987) two types of speaking tasks

- ▲ Factually oriented
 - e.g. description/narration/instruction/comparison
- ▲ Evaluative
 - e.g. explanation/justification/prediction/decision

9

Speaking Tasks

The difficulty of a Speaking task in another language will vary according to...

- familiarity and immediacy of task type/ topic
- support available from linguistic/ physical contexts
- number of participants/ elements involved
- relationship between speakers
- performance conditions such as time available for planning/ rehearsal

10

Why assess Speaking?

No single answer

- Different groups of language learners have different needs
e.g. some groups of adult language learners
 - International travellers: language for travel, leisure
 - Migrants: survival skills, access to employment
 - Students: presentations, seminars
 - Professionals: meetings, professional-client encounters
- Different users have different purposes when they seek information from tests
- But most users of language do need to **speak**

11

Why assess Speaking?

Validity

- + Speaking is an important element of language ability
- + Spoken language is not equivalent to written language

Reliability

- Speaking tests can be difficult to score reliably

Impact

- + Testing Speaking in important tests will (if supported by training) encourage teaching of Speaking

Practicality

- Speaking tests are a challenge logistically
- Speaking tests are often expensive

12

How should we assess Speaking?

A common-sense answer ...

The best way to test people's speaking ability is to get them to speak

(Young & He, 1998)

13

Performance Testing

Performance testing in second language proficiency assessment is traditionally used to describe the approach in which a candidate produces a sample of spoken or written language that is observed and evaluated by an agreed judging process (McNamara 1996)

14

What is Performance Testing?

1. Sample of written or **spoken** language
2. **Simulates behaviour in the real world**
 - ▲ Not like paper-and-pencil 'objective' tests
3. Observed and evaluated by **agreed judging process**

15

Performance Testing

Performance tests are characterised by relationship of test task to world beyond the test:

Ideally tasks are...

- ✓ **direct** and
- ✓ **authentic**
 - **situationally** authentic
 - **interactionally** authentic

But test can never fully represent reality

16

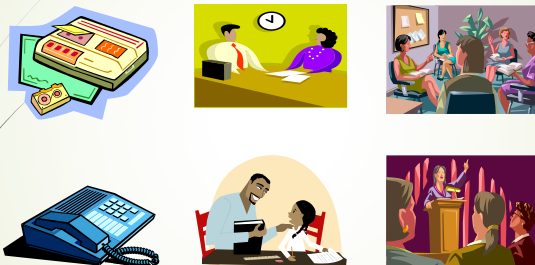
Performance Testing

McNamara (1996) – continuum of performance testing

- ➔ Strong – *focus on successful task completion*
- ← Weak – *focus on language used*

17

How do we get people to speak?



18

Speaking Test Tasks

Nature of input

- Spoken/ visual/ written
- Controlled/ impromptu
- Length/ complexity/ number of speakers
- Topic
- Planning time

Nature of (expected) response

- Length
- Interaction (mono/dialogue)
- Text type
- Functions
- Linguistic features
- Purpose

Relationship between input and response

- Reactivity: degree of reciprocity/ interaction involved
- Scope: amount and range of input to be processed in order to respond
- Directness: extent to which response can be made using information in input by or whether test taker must also rely on contextual information/ real world knowledge.

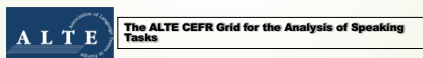
Bachman & Palmer (1996)

19

Let's take a look at some tasks...

Key English Test
(KET) A2

Certificate of Proficiency in English
(CPE) C2



20

Interlocutor Now, in this part of the test you're going to do something together. Here are some pictures of children.

Place picture sheet 1 in front of the candidates.
Select two of the pictures for the candidates to look at*.

First, I'd like you to look at pictures * and * and talk together about what you think might have happened just before the photos were taken. You have about a minute for this, so don't worry if I interrupt you.

Candidates
⌚ 1 minute

Interlocutor Thank you. Now, look at all the pictures.

I'd like you to imagine that a photographic exhibition is being assembled on the theme of Growing Up. All these photographs are to be included.

Talk together about the aspects of 'growing up' the photographers are trying to show. Then suggest two other aspects of growing up which you would like to see represented in the exhibition.

You have about three minutes to talk about this.

Candidates
⌚ 3 minutes

Interlocutor Thank you. Retrieve picture sheet.




21

Candidate B, here is some information about a bookshop. Candidate A, you don't know anything about the bookshop, so ask B some questions about it. Now A, ask B your questions about the bookshop and B, you answer them.

- BOOKSHOP**
- ◆ address ?
 - ◆ big / small ?
 - ◆ closed / Sundays ?
 - ◆ sell / travel books ?
 - ◆ telephone number ?

WORLD BOOKS
212 Main Street



Largest bookshop in the country

Get your travel books here

Monday – Saturday 10.00 am – 8.00 pm
Sunday 12.30 pm – 8.00 pm

Tel: 724 399

22

CEFR A2

CAN produce brief everyday expressions in order to satisfy needs of a concrete type: personal details, daily routines, wants and needs, requests for information.

CAN use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people.


Has a limited repertoire of short memorised phrases covering predictable situations: frequent breakdowns and misunderstandings occur in non-routine situations.

23

KET Part 2

- Supported
- Immediate
- Routine
- Predictable
- Closed
- Simple
- Concrete
- Transactional
- Single source

WORLD BOOKS
212 Main Street



Largest bookshop in the country

Get your travel books here

Monday – Saturday 10.00 am – 8.00 pm
Sunday 12.30 pm – 8.00 pm

Tel: 724 399

- BOOKSHOP**
- ◆ address ?
 - ◆ big / small ?
 - ◆ closed / Sundays ?
 - ◆ sell / travel books ?
 - ◆ telephone number ?

24

CEFR C2

CAN exploit a comprehensive and reliable mastery of a very wide range of language to formulate thoughts precisely, give emphasis and eliminate ambiguity... no signs of having to restrict what she wants to say.

Has command of a very broad lexical repertoire including idiomatic expressions and colloquialisms

CAN create coherent and cohesive text making full use of a variety of organisational patterns and a wide range of cohesive devices.

25

CPE Part 2

- Unsupported
- Unfamiliar
- Open
- Infrequent
- Complex
- Speculative
- Abstract
- Discursive
- Multiple sources



26

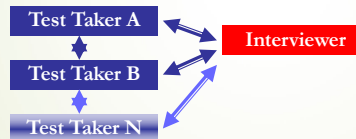
Interaction Patterns



Test mode: live or tape based (direct vs. semi-direct)



Individual, pair or group interaction



27

Authenticity

- **Tape mediated format**
 - Limited to human-machine interaction
 - Talking to a machine not like talking to a person
- **Interview format (1:1)**
 - Questioned in 1990's by van Lier (1989) & Lazaraton (1992)
 - Unbalanced power distribution: interviewer power to initiate & ask all the questions - role of test taker just to answer
 - In other interactions (conversations / discussions), participant rights & responsibilities are more balanced

28

Authenticity

• Paired or group format (n:n)

Swain (2001) 3 arguments in favour of paired or group format:

- Inclusion of more types of talk: broadens evidence of examinees' skills & supports validity of test
- Relationship between teaching & testing: paired and group tests may have positive washback on teaching & encourage more peer to peer interaction in class
- Interviewing pairs and groups is more practical and economical: reduces amount of examiner time needed

29

Paired format: challenges

However, there are a number of challenges

- Luoma (2004) notes that test takers' talk is influenced by other participants personality, communication style and possibly language level
- Problems of interpreting *individual* scores based on *jointly* constructed interaction among participants (McNamara 1997; Swain 2001)

30

Paired format: challenges

- Concern that test takers may not get equal opportunity to display speaking skills at their best
- Cambridge ESOL approach to this challenge is to use 4 tasks - only 1 is paired
- Varied task format gives opportunity for interviewer to redress any imbalance in opportunities 2 test takers have to show speaking ability: BUT not all test takers may get an equal opportunity to talk under all conditions

31

So...

Performance tests are

- ▲ highly complex events involving multiple variables, some of which may be difficult to predict or control
- ▲ Multiple variables or facets of measurement interact to produce a score

Reducing subjectivity (CEFR, p 188)

- **specification of the content** of the assessment
- using **pooled judgements** to select content
- adopting **standard procedures**
- basing judgements on **specific defined criteria**
- requiring **multiple judgements**
- undertaking appropriate **training**
- checking validity, reliability by **analysing assessment data**

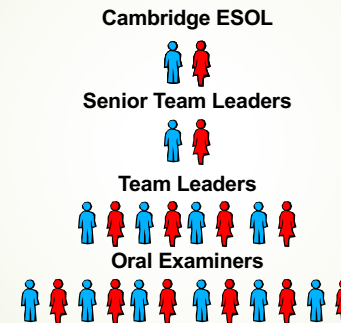
33

Reliability in Speaking tests



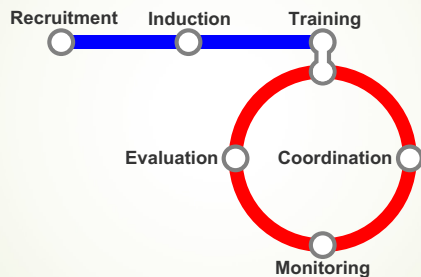
34

Examiner training



35

Training of examiners



36

Rating Scale Criteria

- ▶ Holistic:
 - ▶ Global Achievement (Cambridge ESOL, Main Suite examinations)
- ▶ Analytic:
 - ▶ Range, Accuracy, Fluency, Interaction, Coherence (CEFR)
 - ▶ Grammatical Resource, Lexical Resource, Discourse Management, Interactive Communication, Pronunciation (Cambridge ESOL, Main Suite examinations)

37

Holistic rating scales

- **Positive features**
 - Practicality: fast
 - Rating holistically may be more naturalistic
- **Disadvantages**
 - No useful diagnostic information: single score
 - Not always easy to interpret: raters not required to use same criteria to arrive at score

38

Analytic rating scales

- **Positive features**
 - Can provide diagnostic information if scores reported separately
 - Potentially clear, explicit and detailed
 - Usually more reliable (multiple scores)
 - Useful in training raters to focus on our construct
 - Potentially useful in guiding learners
- **Disadvantages**
 - Time-consuming
 - May overburden raters

39

Rating Scale Criteria (CEFR Appendix 1)

Methods

- Intuitive
- Qualitative
- Quantitative

Qualities

- ✓ Positiveness
- ✓ Definiteness
- ✓ Clarity
- ✓ Brevity
- ✓ Independence

40

Let's take a look at some scales...

IELTS

- **Purpose**
 - Post-secondary/ university level
- **Mode of delivery**
 - Interview: candidate & examiner
- **Test content**
 1. Introduction & overview
 2. Individual long turn
 3. Two-way discussion
- **Scoring**
 - 9-band analytic scale

TSE

- **Purpose**
 - Scores used primarily in decisions about admission to colleges and universities
- **Mode of delivery**
 - Semi-direct (tape-based)
- **Test content**
 1. Story based on six-picture sequence
 2. Describe graph
 3. 2 questions about ideas
 4. 3 questions in workplace situation
- **Scoring**
 - 5-point holistic scale

41

Number of observations

More test tasks



More raters



42

Speaking: Validity and Reliability

"Reliability is concerned with minimizing the effects of measurement error, while validity is concerned with maximizing the effects of the language abilities we want to measure"
(Saville 2003:69)

Potential tension between validity and reliability in performance assessment.

43

Tensions

← **Validity/ Impact**

← 'Direct testing'

← Variety of 'Task types'

← More functions tested

← Fewer inferences required

← Administration complex

← More positive impact on teaching and learning

→ **Reliability/ Practicality**

→ 'Indirect testing'

→ Fewer 'task types'

→ Fewer functions tested

→ More inferences required

→ Administration simpler

→ Less positive impact on teaching and learning