Re-examining Speaking for teaching and testing: theory in practice

Professor Lynda Taylor
CRELLA, University of Bedfordshire, UK
Introduction

• modern languages graduate
• English language (EFL) teacher and teacher trainer
• EFL materials writer (CUP)
• test developer (for Cambridge English tests)
• researcher and consultant (Cambridge and CRELLA)
• retired? ... not really...
Background

- longstanding interest in assessment of speaking skills

- integration of applied linguistics, language pedagogy and language assessment

- quality assurance issues related to direct assessment of speaking and writing
Background

• longstanding interest in and commitment to developing assessment literacy
• papers in professional journals
• training courses in UK, Europe, Asia
• professional organisations, e.g. ILTA, EALTA, IATEFL
Aims of our session today

• to explore a **theoretical model** of spoken language ability for assessment design/evaluation purposes

• to consider how this model can **inform and shape our testing practices** when assessing speaking ability

• input to be followed up in a **workshop** later in the conference
Some serious questions we could ask about this particular test task...

- What skill is this test designer seeking to assess?
- How well does his assessment task match his target population?
- What will be his criteria for success in the task?
- How fair is the task?
- Has he designed a task that is ‘fit for purpose’ (i.e. will it meet the test designer’s aims and needs)?
Our understanding of language proficiency constantly evolves in light of...

• changes in social context and priorities
• changes in applied linguistic theory
• changes in language pedagogy
• changes in technology
Our understanding of language proficiency also evolves in light of...

- changes in assessment theory and practice
  - testing aims and purposes
  - testing volumes and frequency
  - the educational and social impact of tests
Modelling and assessing the language proficiency construct

• teachers and testers need an accessible and sound theory to serve as a springboard for their own language test development practices

• teachers and testers need access to a sound theoretical model in order to describe, analyse and critically evaluate tests and their uses
The concept of ‘validity’

• a fundamental concept for all types of measurement and measuring tools

• a key concept when designing a language assessment task or developing a test instrument
Defining validity...

The extent to which scores on a test enable inferences to be made which are appropriate, meaningful and useful, given the purpose of the test. Different aspects of validity are identified which provide evidence for judging the overall validity of a test for a given purpose.

How do we gather and present evidence of a test’s validity?

We need a model or framework for language proficiency assessment.

- **Evidence-Centred Design (ECD)**
  - Mislevy, Almond and Steinberg, 2003
- **Assessment Use Argument (AUA)**
  - Bachman and Palmer, 2010

- **Socio-Cognitive Framework (Weir, 2005)**
  - addresses the test-in-practice as well as the test-in-theory
  - serves as sound practical basis test design and development
  - serves as accessible tool for test description, analysis and evaluation
  - systematically addresses variation across proficiency levels (cf. CEFR)
A ‘socio-cognitive’ view of speaking ability

• speaking is a ‘social’ as well as a linguistic phenomenon, i.e. the user is a ‘social agent’ who needs to be able to perform certain actions in the language according to context (North 2009)

• speaking involves mental processing, i.e. the exercise of internal cognitive and metacognitive abilities such as planning and monitoring, as well as linguistic knowledge and resources
Core components of a socio-cognitive model of speaking performance

- Test taker (with personal characteristics)
- Context validity
- Cognitive validity
- Spoken performance
A socio-cognitive model for assessing speaking ability

Test taker characteristics

Context validity

Cognitive validity

STIMULUS TEST TASK

Spoken performance
Considerations when designing a speaking test task

Test taker characteristics

FEATURES OF STIMULUS TEST TASK

Context validity

Cognitive validity
Considering test taker characteristics when designing a speaking test task

TEST TAKER CHARACTERISTICS

PHYSICAL/PHYSIOLOGICAL:
- age and gender
- disability

EXPERIENTIAL:
- first language
- cultural background
- educational level
- life experience

PSYCHOLOGICAL:
- personality
- affective schemata
- motivation

Disability
What is the linguistic and socio-cultural context appropriate to the speaking task?

**CONTEXT VALIDITY**

- **Task setting**
  - Purpose
  - Response format
  - Known criteria
  - Weighting
  - Order of test items
  - Time constraints
  - Physical conditions

- **Task demands**
  - **Linguistic demands:**
    - Domain
    - Discourse mode
    - Content knowledge
    - Cultural specificity
    - Nature of information
    - Presentation
    - Lexical range
    - Structural range
    - Functional range
    - Topic familiarity
  - **Interlocutor factors:**
    - Speech rate
    - Variety of accent
    - Number
    - Acquaintanceship
    - Gender
What cognitive processing is activated by the speaking task?

Levels of cognitive processing:
- Conceptualisation (pre-verbal)
- Grammatical encoding
- Morpho-phonological encoding
- Phonetic encoding/Articulation
- Self-monitoring (→ self-repair)

Information sources:
- Speaker goals, world/local knowledge, discourse recall, rhetorical patterns
- Topic recall, knowledge of syntax, pragmatics, formulaic chunks
- Lexical and phonological knowledge
- Knowledge of articulatory settings
- Speaker goals, target utterance stored in buffer, discourse recall
A socio-cognitive model for assessing speaking ability
A socio-cognitive model for assessing speaking ability
A socio-cognitive model for assessing speaking ability
How is the speaking task scored, and is the approach appropriate?

Scoring validity

Rating:
Assessment criteria/rating scale
Rating process
Rating conditions
Rater characteristics
Rater training
Post test adjustment
Score reporting
Can we turn this theoretical framework into a practical tool for test task design and evaluation?

- Features of the stimulus test task(s)
- Features of the expected response on the task (spoken output)
- Scoring features
Using the socio-cognitive framework to analyse speaking test tasks

• by systematically examining the **features of the stimulus task(s)** in terms of their context validity and cognitive processing demands

• by systematically examining the **features of the expected response(s)** on the task, i.e. the spoken output, in terms of its contextual features

• by systematically examining the **reliability of rating** the spoken output

**Features of the stimulus test task(s)**

**Features of the expected response on the task (spoken output)**

**Scoring features**
Let’s take a specific Speaking Test and analyse it...

**Part 1** : the examiner asks general questions about test taker and a range of familiar topics, such as home, family, work, studies and interests (4-5 mins).

**Part 2** : the test taker receives a card asking them to talk about a particular topic. There is 1 minute to prepare before speaking for up to 2 minutes. The examiner then asks 1 or 2 questions on the same topic (3-4 mins).

**Part 3** : the examiner asks further questions about the topic in Part 2. These provide the opportunity to discuss more abstract ideas and issues (4-5 mins).

(11–14 minutes)
Using the socio-cognitive framework to analyse the IELTS Speaking Test tasks

- by systematically examining the features of the stimulus task(s) in terms of their context validity and cognitive processing demands
- by systematically examining the features of the expected response(s) on the task, i.e. the spoken output, in terms of its contextual features
- by systematically examining the reliability of rating the spoken output
# Features of the Stimulus Task(s)

<table>
<thead>
<tr>
<th>Context Validity Features</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>social and academic (not work)</td>
</tr>
<tr>
<td>Discourse mode</td>
<td>descriptive, biographical, expository, argumentative, instructive</td>
</tr>
<tr>
<td>Content knowledge</td>
<td>midway between general and specific</td>
</tr>
<tr>
<td>Cultural specificity</td>
<td>neutral (not specific)</td>
</tr>
<tr>
<td>Nature of information</td>
<td>mostly concrete, fairly abstract</td>
</tr>
<tr>
<td>Presentation</td>
<td>verbal (not non-verbal, i.e. graphs)</td>
</tr>
<tr>
<td>Lexical range</td>
<td>unspecified</td>
</tr>
<tr>
<td>Structural range</td>
<td>unspecified</td>
</tr>
<tr>
<td>Functional range</td>
<td>unspecified</td>
</tr>
<tr>
<td>Topic familiarity</td>
<td>familiar to unfamiliar</td>
</tr>
<tr>
<td>Intended speaker/listener relationship</td>
<td>Speech rate: normal; Variety of accent: not specified; Number of speakers: 2; Acquaintanceship: not specified; Gender: not specified</td>
</tr>
</tbody>
</table>
# Features of the Stimulus Task(s)

<table>
<thead>
<tr>
<th>Cognitive Validity Features - Levels of speaking</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptualisation</td>
<td>√</td>
</tr>
<tr>
<td>Grammatical encoding</td>
<td>√</td>
</tr>
<tr>
<td>Phono-morphological encoding</td>
<td>√</td>
</tr>
<tr>
<td>Phonetic encoding / Articulation</td>
<td>√</td>
</tr>
<tr>
<td>Self-monitoring</td>
<td>√</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cognitive Validity Features - interaction pattern and planning time</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-reciprocal, i.e. semi-direct (computer/tape-based)</td>
<td></td>
</tr>
<tr>
<td>Reciprocal, i.e. direct (face-to-face)</td>
<td>√</td>
</tr>
<tr>
<td>Planning time allowed</td>
<td>√</td>
</tr>
<tr>
<td>No planning time</td>
<td></td>
</tr>
</tbody>
</table>
Features of the Expected Response(s) – Spoken Output

<table>
<thead>
<tr>
<th>Context Validity Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>Discourse mode</td>
</tr>
<tr>
<td>Content knowledge</td>
</tr>
<tr>
<td>Cultural specificity</td>
</tr>
<tr>
<td>Nature of information</td>
</tr>
<tr>
<td>Presentation</td>
</tr>
<tr>
<td>Lexical range</td>
</tr>
<tr>
<td>Structural range</td>
</tr>
<tr>
<td>Functional range</td>
</tr>
<tr>
<td>Topic familiarity</td>
</tr>
</tbody>
</table>

* Nakatsuhara (2012)
# Scoring Validity

<table>
<thead>
<tr>
<th>Reliability: Estimated reliability of rating quality of spoken output (rater consistency)</th>
<th>0.83-0.86 (based on recent G-studies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-rater reliability (correlation between Rater 1 and Rater 2)</td>
<td></td>
</tr>
</tbody>
</table>
How/why is this approach useful?

• In speaking test design and development
  • it helps to systematically describe the critical context and cognitive validity features of their test in relation to the stimulus tasks and the expected response (i.e. the spoken output)

• In speaking test validation and evaluation
  • it helps to describe, analyse and critically evaluate tests and the uses to which they are put (i.e. their fitness for purpose)
Let’s take an example of a comparative study....

- **General Medical Council (GMC)** is a regulatory body for the medical profession in the UK
- the organisation verifies and validates international medical qualifications and registers overseas-trained doctors to practise in the UK
- for professional registration the GMC also requires evidence of international doctors’ English language proficiency
- for some years the GMC has used IELTS Academic as a test to provide such evidence of reading, writing, listening and speaking ability
In 2014...

• the GMC commissioned research to review the suitability of other language tests that might provide similar evidence (e.g. TOEFL iBT)
• the aim was to ensure the GMC was using the most secure, efficient and flexible approaches to check whether doctors have the required standard of English to practise in the UK
• CRELLA won the bid to conduct the research, and the study adopted a mixed methods approach, involving a systematic documentary review of selected tests combined with desk-based internet research and targeted follow-up with specific test providers and test user organisation
• the socio-cognitive framework was developed into a practical tool to describe, analyse, evaluate and compare a range of 9 major English language tests against IELTS as the baseline for comparison
• full report and appendices (October 2015) available on GMC website
Comparing various Speaking tests against IELTS...

<table>
<thead>
<tr>
<th>Context Validity – Features of the Stimulus Task(s)</th>
<th>Cambridge Advanced</th>
<th>Pearson Test of English - Academic</th>
<th>TOEFL iBT</th>
<th>Occupational English Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>=</td>
<td>&lt;</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Discourse mode</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>≤</td>
</tr>
<tr>
<td>Content knowledge</td>
<td>=</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>Cultural specificity</td>
<td>≥</td>
<td>=</td>
<td>≥</td>
<td>=</td>
</tr>
<tr>
<td>Nature of information</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Presentation</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>=</td>
</tr>
<tr>
<td>Topic familiarity</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>&lt;</td>
</tr>
<tr>
<td>Intended speaker/listener relationship</td>
<td>&gt;</td>
<td>&gt;</td>
<td>=</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

Key: = comparable with IELTS   > more demanding than IELTS   < less demanding than IELTS
≥ marginally more demanding than IELTS   ≤ marginally less demanding than IELTS
Comparing various Speaking tests against IELTS... (points to note)

<table>
<thead>
<tr>
<th>Context Validity – Features of the Stimulus Task(s)</th>
<th>Cambridge Advanced</th>
<th>Pearson Test of English - Academic</th>
<th>TOEFL iBT</th>
<th>Occupational English Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>=</td>
<td>&lt;</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Discourse mode</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>≤</td>
</tr>
<tr>
<td>Content knowledge</td>
<td>=</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>Cultural specificity</td>
<td>≥</td>
<td>=</td>
<td>≥</td>
<td>=</td>
</tr>
<tr>
<td>Nature of information</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Presentation</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>=</td>
</tr>
<tr>
<td>Topic familiarity</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>&lt;</td>
</tr>
<tr>
<td>Intended speaker/listener relationship</td>
<td>&gt;</td>
<td>&gt;</td>
<td>=</td>
<td>&gt;</td>
</tr>
</tbody>
</table>
Comparing various Speaking tests against IELTS...

<table>
<thead>
<tr>
<th>Context Validity – Features of the Expected Response</th>
<th>Cambridge Advanced</th>
<th>Pearson Test of English - Academic</th>
<th>TOEFL iBT</th>
<th>Occupational English Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse mode</td>
<td>=</td>
<td>&lt;</td>
<td>=</td>
<td>&gt;</td>
</tr>
<tr>
<td>Presentation</td>
<td>=</td>
<td>≤</td>
<td>≤</td>
<td>≥</td>
</tr>
<tr>
<td>Lexical level</td>
<td>? (&gt;)</td>
<td>?(=)</td>
<td>?(=)</td>
<td>? (&gt;)</td>
</tr>
<tr>
<td>Grammatical/structural level</td>
<td>? (&gt;)</td>
<td>?(=)</td>
<td>?(=)</td>
<td>?(=)</td>
</tr>
<tr>
<td>Functional range</td>
<td>&gt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&gt;</td>
</tr>
</tbody>
</table>
Comparing various Speaking tests against IELTS...

<table>
<thead>
<tr>
<th>Skills focus</th>
<th>Cambridge Advanced</th>
<th>Pearson Test of English - Academic</th>
<th>TOEFL iBT</th>
<th>Occupational English Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive processing</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

Reliability: Estimated reliability of rating quality of spoken output (rater consistency)

<table>
<thead>
<tr>
<th>Inter-rater reliability (correlation between raters, except PTE-A)</th>
<th>Cambridge Advanced</th>
<th>Pearson Test of English - Academic</th>
<th>TOEFL iBT</th>
<th>Occupational English Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Equally reliable (0.82)</td>
<td>More reliable (0.91)</td>
<td>Equally reliable (0.88)</td>
<td>Equally reliable (0.89)</td>
</tr>
</tbody>
</table>
Aims of our session today

• to explore a theoretical model of spoken language ability for assessment design/evaluation purposes

• to consider how this model can inform and shape our testing practices when assessing speaking ability
Thank you for your attention!