TEMPUS

Modernising higher education

Testing speaking skills: why and how?

Dr Fumiyo Nakatsuhara

Dr Chihiro Inoue

CRELLA, University of Bedfordshire

European Commissi

TEMPUS

CRELLA (Centre for Research in English Language Learning and Assessment)

United Kingdom
Bangor
Belfast
Isle of Man
Lake District National Park
Dundalk
York
Blackpool
Leeds
Hull
Manchester
Dublin
Sheffield
Liverpool
Stoke-on-Trent
Nottingham
Leicester
Birmingham
Coventry
A
Pembrokeshire Coast National Park
Headington
London
Cardiff
Tilehurst
Leigh-on-Sea
Bristol
Barnstaple
Southampton
Exeter
Brighton
Portsmouth
Torquay
Truro
Plymouth
English Channel

www.beds.ac.uk/crella

CRELLA Research Staff

Prof Cyril Weir: Director of CRELLA

Prof Stephen Bax: Professor in Applied Linguistics

Prof Tony Green: Professor in Language Assessment

Dr Vladimir Zegarac: Reader in Language and Communication

Dr John Field: Senior Lecturer in Cognition in Language Learning and Assessment

Dr Lynda Taylor: Senior Lecturer in Language Assessment

Dr Fumiyo Nakatsuhara: Senior Lecturer in Language Assessment

Dr Chihiro Inoue: Post-doctoral Research Fellow

Dr Sathena Chan: Post-doctoral Research Fellow

Prof Liz Hamp-Lyons: Visiting Professor

Prof Roger Hawkey: Visiting Professor

Rebecca van der Westhuizen: Research Administrator

Aims of this workshop

- To understand the importance of **speaking tests**
- To be familiar with **different formats** of speaking tests
- To understand the **advantages and disadvantages** of different speaking formats
- To obtain a basic familiarity with **rating scales** and **rating standardisation**

Plan of the workshop

1. **Different test types**
2. **Washback effect**
3. **Speaking tests**

3.1 Interview speaking tests

- Rating interview test performance
- Interviewer variability

3.2 Paired speaking tests

- Performing a paired speaking test
- Advantages & disadvantages

1. Different test types

Test Purposes

- Proficiency tests
- Achievement tests
- Placement tests
- Diagnostic tests

Direct vs Indirect testing

- Direct Testing
- Indirect Testing

NR vs CR testing

- Norm-referenced testing
- Criterion-referenced testing

- **Proficiency tests:** to measure "people's ability in a language, regardless of any training they may have had in that language. The content is … based on a specification of what candidates have to be able to do in the language in order to be considered proficient." → We need to decide what we mean by "proficient"!
- **Achievement tests:** to measure "how successful individual students, groups of students, or the courses themselves have been in achieving objectives." → Directly related to language courses (goal and content)
- **Placement tests:** "to place students at the stage of the teaching programme most appropriate to their abilities"
- **Diagnostic tests:** "to identify learners' strengths and weaknesses"

(Hughes, 2003: 11-17)

Direct vs Indirect testing

- **Direct testing:** "requires the candidate to perform precisely the skill that we wish to measure"
- **Indirect testing:** "attempts to measure the abilities that underlie the skill in which we are interested"

(Hughes, 2003: 17-19)

NR vs CR testing

- **Norm-referenced testing:** "An individual performance is evaluated against the range of performances typical of a population of similar individuals"
- **Criterion-referenced testing:** "Individual performances are evaluated against a verbal description of a satisfactory performance at a given level."

(McNamara, 2000: 62-64, 135)

2. Washback effect

Washback effect: effect of testing on teaching and learning

- **Positive/Negative washback effect**
- **2 major types of threats to construct validity:** "tests are imperfect measures of constructs because **they either leave out something that should be included according to the construct theory** *(construct under representation)* or **else include something that should be left out** *(construct-irrelevant variance)*, or both" (Messick, 1989: 36).
- **Minimising** these 2 threats is significant for generating the ground for fostering positive washback (Messick, 1996).

Our responsibility as testers and teachers

- Testing is very important part of teaching and learning. Testing should give a positive washback effect on teaching and learning.

- If speaking is the ability which we wish to enhance, the assessment should directly test oral skills.

3. Speaking tests

What do we need to decide before giving a speaking test?

- What **aspects of language** we want to assess
- How to **elicit ratable language samples** from test-takers suitable for the aspects of language

We need to decide;

- **Rating criteria** [marking categories, levels, descriptors] [Holistic scales vs. Analytic scales]
- **Elicitation techniques / Test format** (types of questions, task types)

3.1 Interview speaking tests - Let's practise rating interview test performance!

- **Rating criteria:**
  - **Phonological control**
  - **Grammatical accuracy**
  - **Vocabulary range**
  - **Fluency**

  (Taken from ‘Common European Framework of Reference for Languages’, Council of Europe 2001)
- **Test format: interview format** with the following structure

1	**Openings** (1 minute)
2	**Conversation on familiar topics** (3 minutes) The interviewer asks the candidate to talk about him/herself.
3	**Picture Description** (2 minutes) The interviewer asks the candidate to describe a photo.
4	**Conversation on topics from the given picture** (5 minutes) The interviewer asks the candidate questions linked to the picture (from general to extended questions).
5	**Closings** (1 minute)

- Issues of Interviewer variability

Nakatsuhara, F. (2008) ELTJ 62 (3): 266-275

1. Interviewer's interactional variation

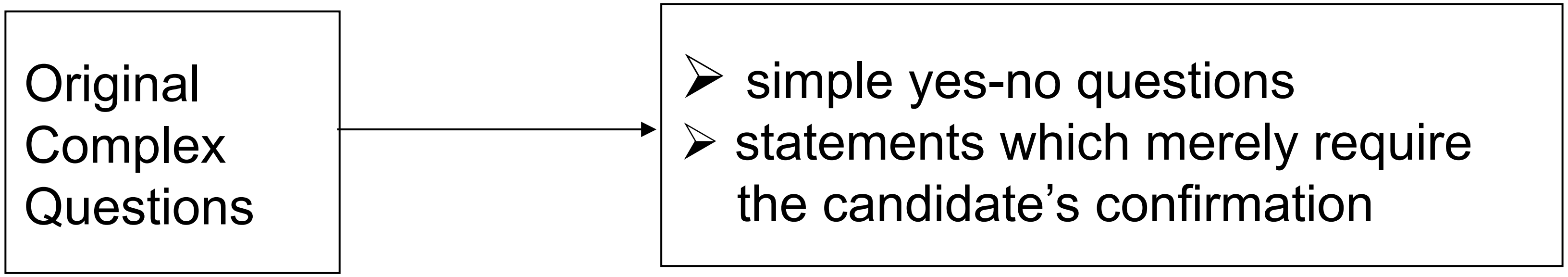
- **Interviewer Accommodation**

(e.g. Slowdown, lexical simplification, rephrasing questions, etc.)

→ **Validate the test**

But, *inconsistent accommodation*
→influence on candidate performance

□ **Deprive interviewees of opportunities to speak**



Original Complex Questions

➢ simple yes-no questions
➢ statements which merely require the candidate's confirmation

(Lazaraton, 1996)

□ **Over-accommodation for candidates at a certain level** (Ross and Berwick,1992)

2. scores affected by the interactional difference

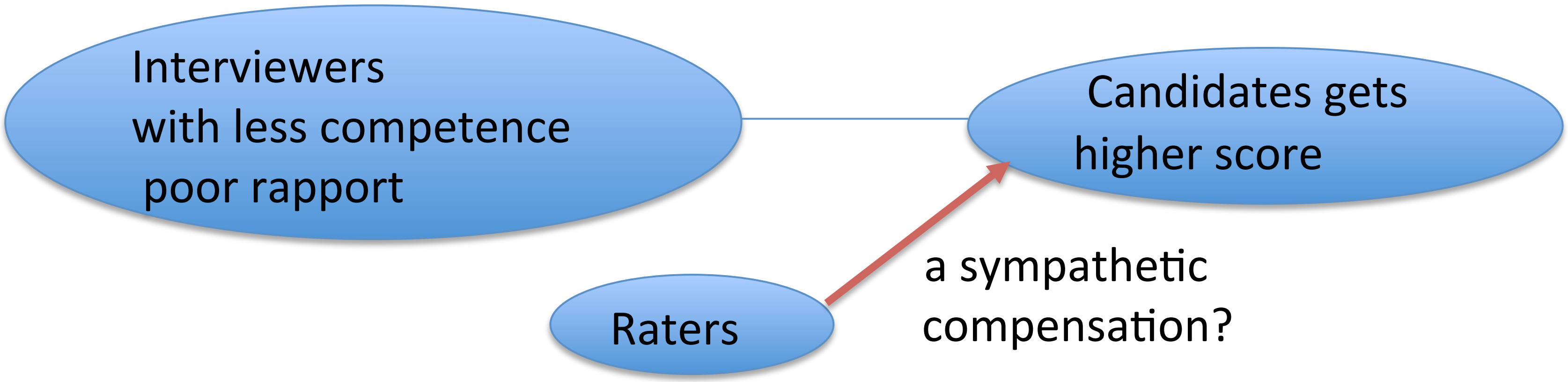
- **"Interviewer difficulty"** (Brown and Hill, 1998; Brown, 2003)

□ ***the easiest interviewer***

0.6 difference (IELTS speaking scale)

□ ***the most difficult interviewer***

- **Ratings and interviewer competence / the amount of rapport** (McNamara and Lumley, 1997)



Interviewers
with less competence
poor rapport

Candidates gets
higher score

Raters

a sympathetic
compensation?

Research Questions

1. Are there **any analytic marking categories** especially affected by the interviewer difference?

2. If so, **what interlocutor behaviour** influenced the analytic components?

Data Collection

Subjects: 1 candidate/ 2 interviewers/ 22 raters (with experience)

Interviewer A / Candidate C

Interviewer B / Candidate C

rated by 22 raters (D-Z)

Analytic rating scale [criterion-referenced]:

Pronunciation
Grammar
Vocabulary resource
Fluency
Interactional Communication

with 4 levels (0-3)

Tasks:

1. **Openings** (1 min)
2. **Conversation on familiar topics** (3 mins)
3. **Picture description** (2 mins)
4. **Conversation on topics from the given picture** (5 mins)
   [General questions → Extended questions]
5. **Closings** (1 min)

Data Analyses

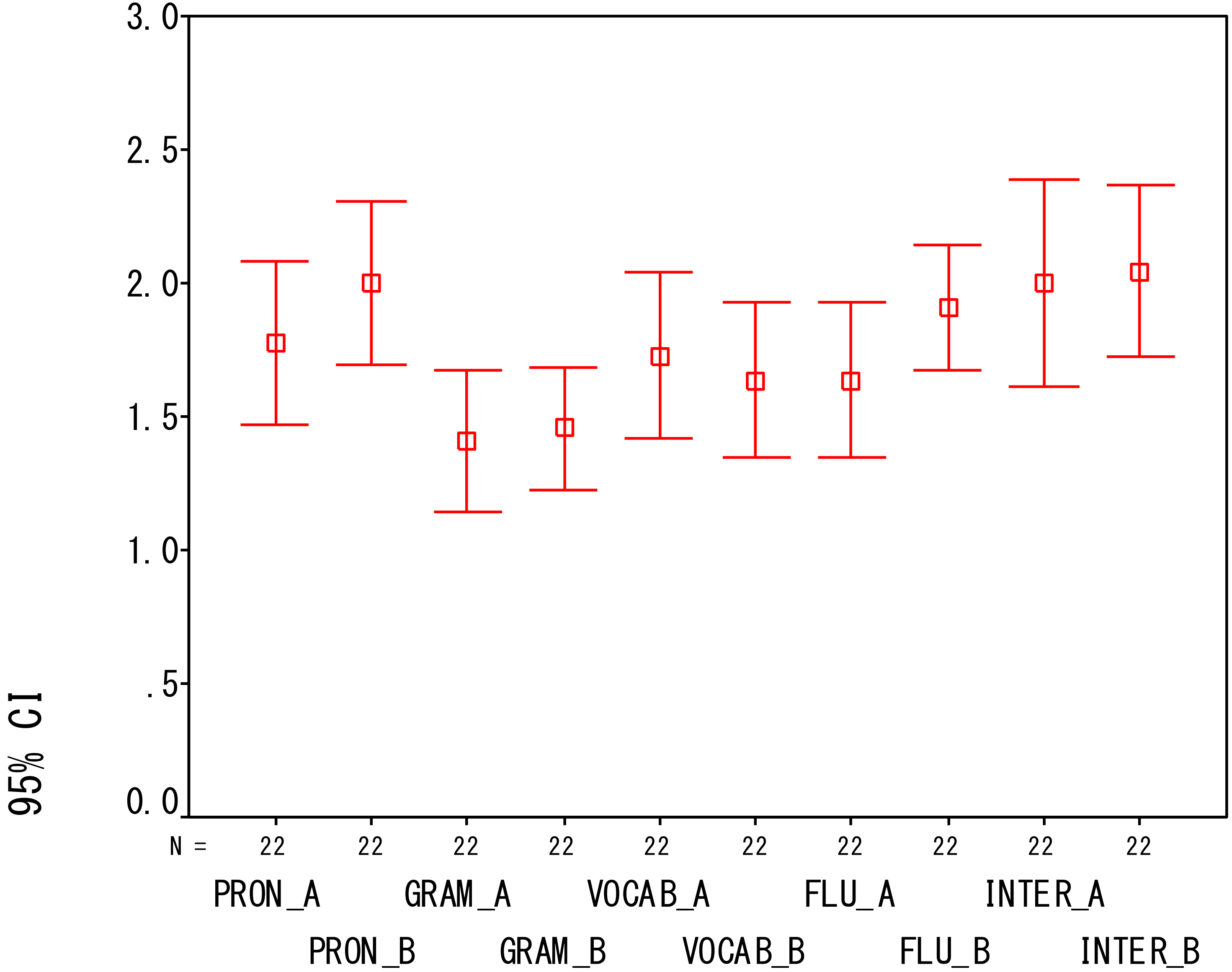
- **Paired Sample t-tests to compare the rating results of two sessions** (RQ1: Are there **any analytic marking categories** especially affected by the interviewer difference?)

- **Conversation analysis (CA)**

→ **The CA findings + commentaries of 22 raters on their rating**

(RQ2: If so, **what interlocutor behaviour** influenced the analytical components?)

Rating results



Inter-rater reliability Tests

Absolute sense: 50%; Relative sense: 77%

Paired samples t-tests

	Mean difference	SD	t	df	Sig (2-tailed)	
Pron_A – Pron_B	-.2273	.4289	-2.485	21	**.021**	**Sig**
Gram_A – Gram_B	-.0455	.3751	-.568	21	.576	
Vocab_A – Vocab_B	.0909	.5263	.810	21	.427	
Flu_A – Flu_B	-.2727	.5505	-2.324	21	**.030**	**Sig**
Inter_A – Inter_B	.0455	.7854	-.271	21	.789	
All_A – All_B	-.5000	1.7113	-1.370	21	.185	

Interviewer techniques

<table>
<tr><th>Interviewer A</th><th>Interviewer B</th></tr>
<tr><td colspan="2"><em>1) questioning and topic nomination techniques</em></td></tr>
<tr><td>Rephrasing</td><td>Stating question prompts as statements</td></tr>
<tr><td colspan="2"><em>2) topic expansion and management techniques</em></td></tr>
<tr><td>Topic recycling by various questions</td><td>Frequent topic shift</td></tr>
<tr><td colspan="2"><em>3) receipt tokens and feedback techniques</em></td></tr>
<tr><td>No feedback comment</td><td>Echoing</td></tr>
<tr><td>Non-verbal receipt tokens</td><td>Evaluative comment</td></tr>
</table>

A's topic expansion and management techniques

(1) Demanding more opinions

1 C: Yeh, freedom freedom for child n: if the child is crying, OK if crying OK
2 finish ah will be goo(h)d
3→I: Right. Do you agree with that? Or do you [( )

(2) Requesting reasons for the previous answer

1 C: =but I I think she uh:: doesn't care the ki(h)ds cry(hah)ing
2→I: All right. Hah hah ha What makes you say that?=

(3) Asking for examples

1 I: =So so in <u>your</u> idea or your point of view, what makes a good mother?
2 C: Em:: (.) manage em something they <u>should</u> manage [and [ha huh
3→I: [Right [For example?
Ahah hah ha=

B’s receipt tokens and feedback techniques

(4) Echoing the candidate’s utterance

1 I: Ah OK right so how long have you been studying here?
2 C: Uh::, about **two months**
3→I: two months?=

(5)

1 I: Uh so you watch videos.
2 C: Um: **Just watching TV** hah hah ha
3→I: Just TV

(6) Evaluative comment

1 I: So their [parents should be responsible.
2 C: [Yah yah yah
3 C: Yes.
4→I: Oh OK Yeah very good. yeah, very very good. OK.

Discussion

- **Why were different scores awarded to "Pronunciation" and "Fluency" components?**

1) *the degree of interviewer control*

A's interview:
Highly controlled

➢topics were clearly defined at every stage

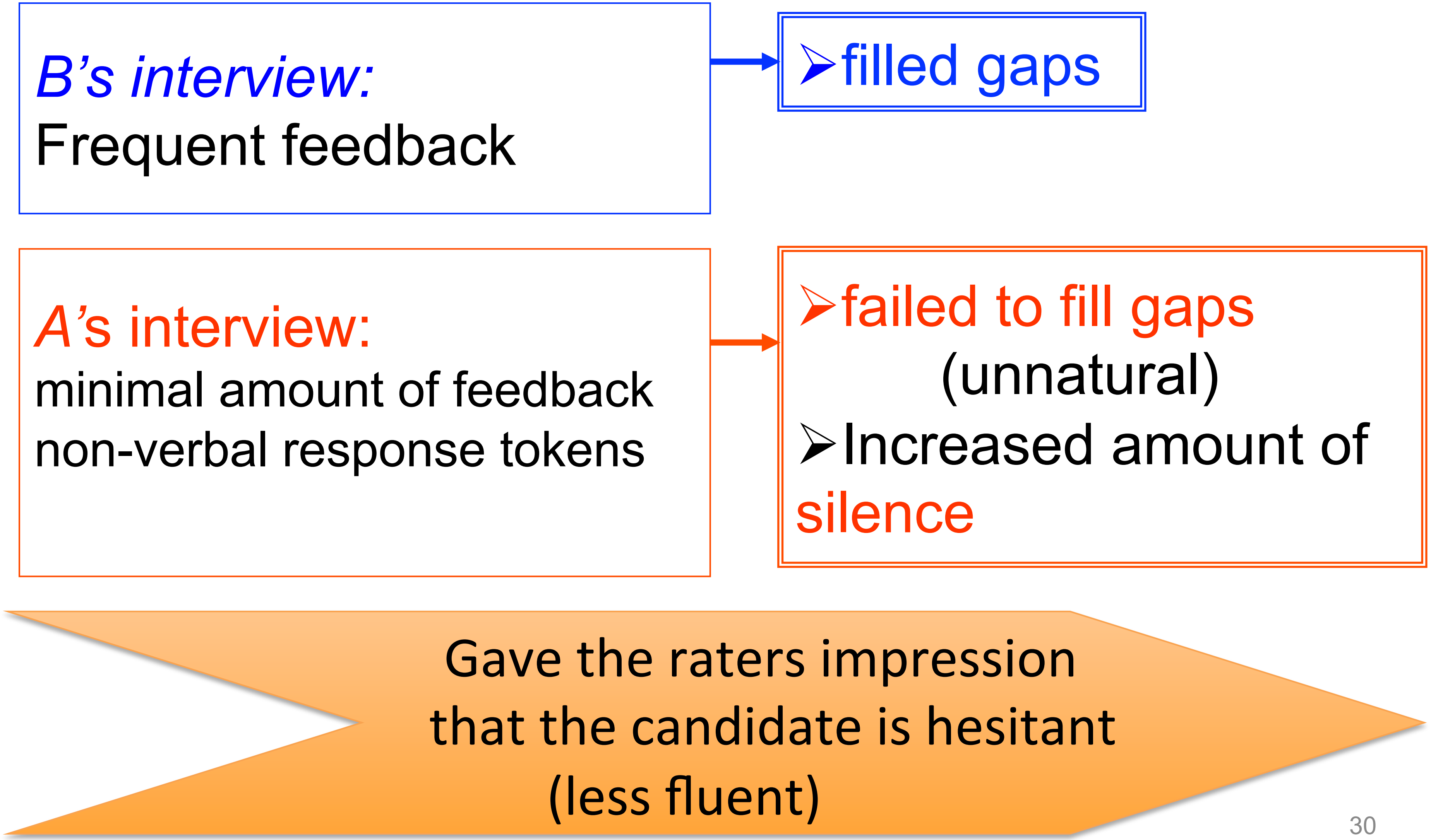
B's interview:
Less controlled
Topic shift

➢ avoided lexis whose pronunciation that she might get wrong

➢spoke more fluently whatever she wanted to talk

Avoidance Strategies
(Faerch and Kasper, 1984)

2) *the types and the amount of interviewer's feedback*



B's interview:
Frequent feedback

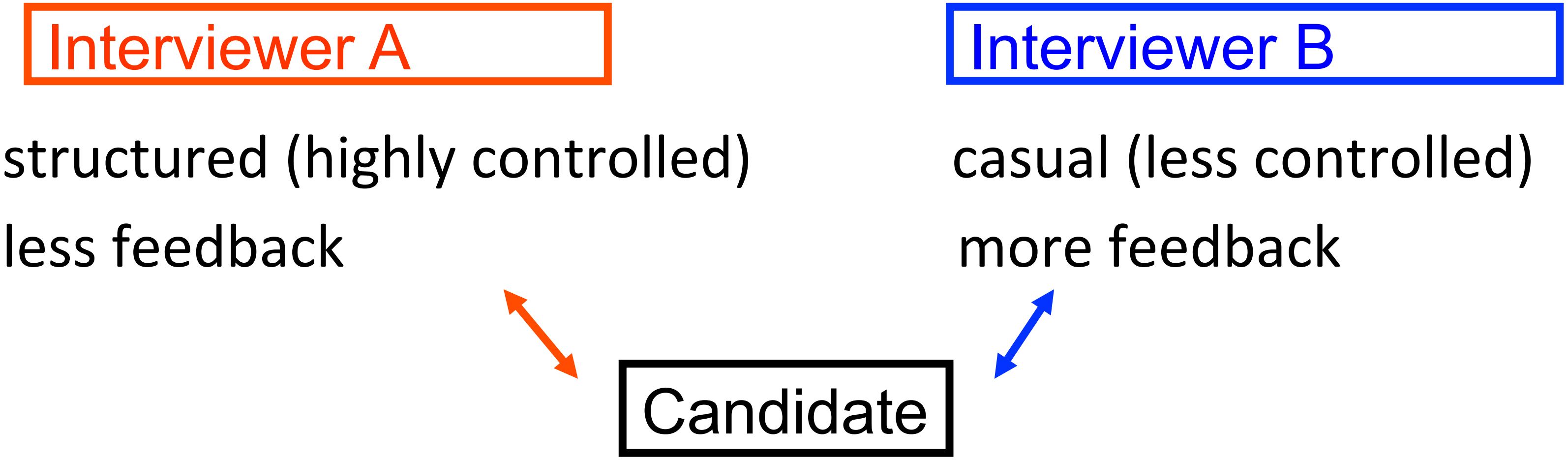
- filled gaps

A's interview:
minimal amount of feedback
non-verbal response tokens

- failed to fill gaps
  (unnatural)
- Increased amount of silence

Gave the raters impression
that the candidate is hesitant
(less fluent)

Conclusion



Interviewer A

structured (highly controlled)

less feedback

Interviewer B

casual (less controlled)

more feedback

Candidate

“Pronunciation” & “Fluency” score differences

- **More precise picture** of the possible relationship between **interviewer variation** and **rating scores** affected

3.2 Paired speaking tests - Let’s perform a paired speaking test!

Cambridge EFL examinations

KET (1993-); PET (1995-); FCE (1996-); CAE (1991-); CPE (2004-)

Ex. The structure of the CAE Speaking Test

1	Three-way conversation between the candidates and the interlocutor	3 minutes
2	Individual long turns with brief responses from second candidate	4 minutes
3	**"Two-way collaborative (problem-solving) task" Two-way interaction between the candidates**	**4 minutes**
4	Three-way conversation between the candidates and the interlocutor	4 minutes

Two-way Collaborative Task (4 minutes) **[Hotel Staff]**

◆ Imagine that both of you are co-managers of a new 4-star international hotel in London. You are looking for 7 different staff positions.

(1)Talk to each other about how demanding/ important these jobs are.

(2)Rank the 7 jobs according to the salary you are prepared to pay them, and give your reasons.

◆ *Can you both try to agree on the rank-order?*

Waiter

Chef

Receptionist

Porter

Barman

Pianist

Cleaner

Advantages of Paired formats

- Why do you think this particular test format is preferred?

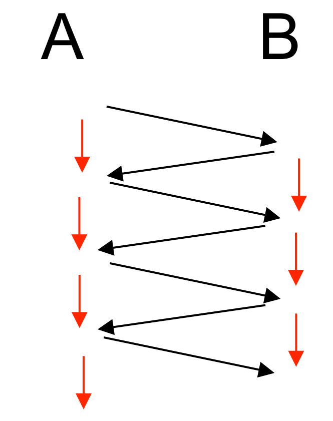
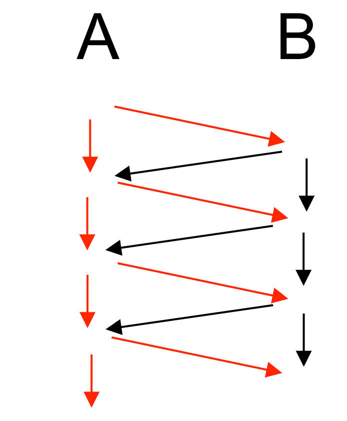
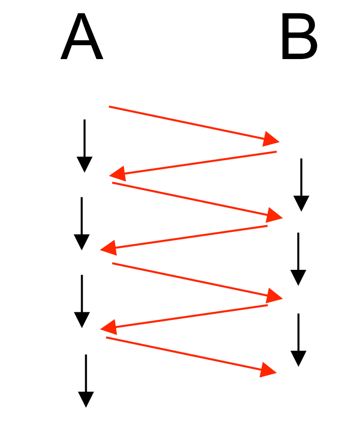
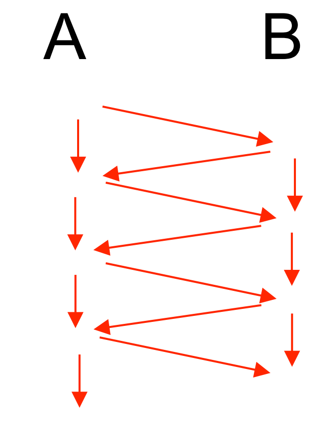
- What language functions / language aspects do you think are tested in paired formats which are **not** tested in interview formats?

Advantages of Paired formats

- Capable of **eliciting more symmetrical contribution** to the interaction from test-takers
- Capable of **eliciting a much richer and more varied language functions**
- **Positive reaction** from test-takers (less anxious), a sign of **positive washback** effect
- **Practical**: time-efficient, cost-effective, less burden and less training for the examiners

Comparisons of Interview Discourse with Conversational Discourse

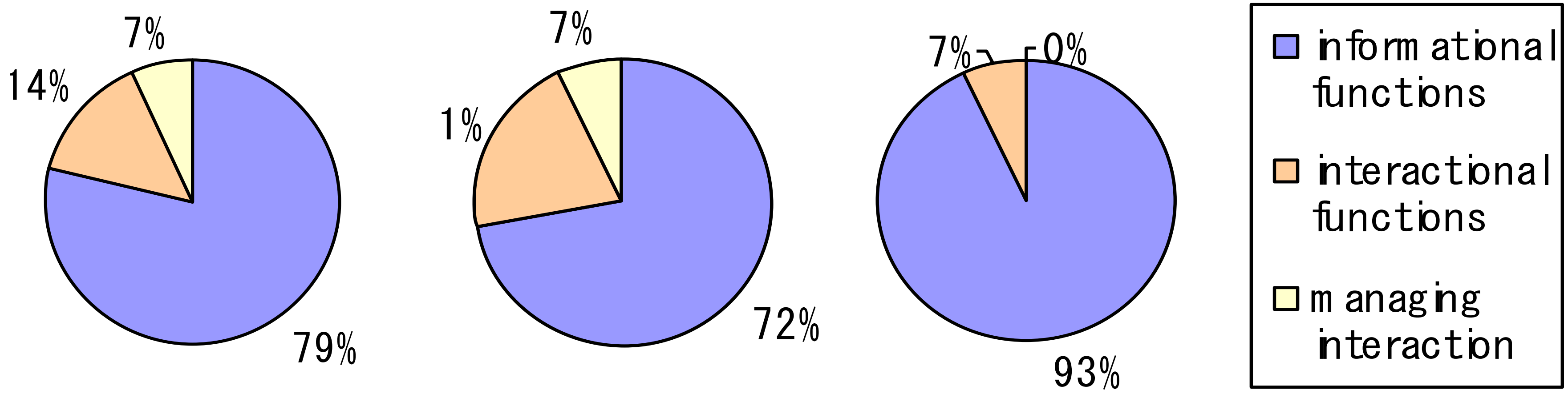
(van Leir,1989;Young and Milanovic,1992; Young,1995; Kormos, 1999)

Time				
	Pseudo-Contingency	**Asymmetrical Contingency**	**Reactive Contingency**	**Mutual Contingency**
Ex. Inter-action	• acting in a play • rituals (e.g. greeting)	• interviewing • (traditional) teaching	• rambling conversation	• negotiation • serious discussion

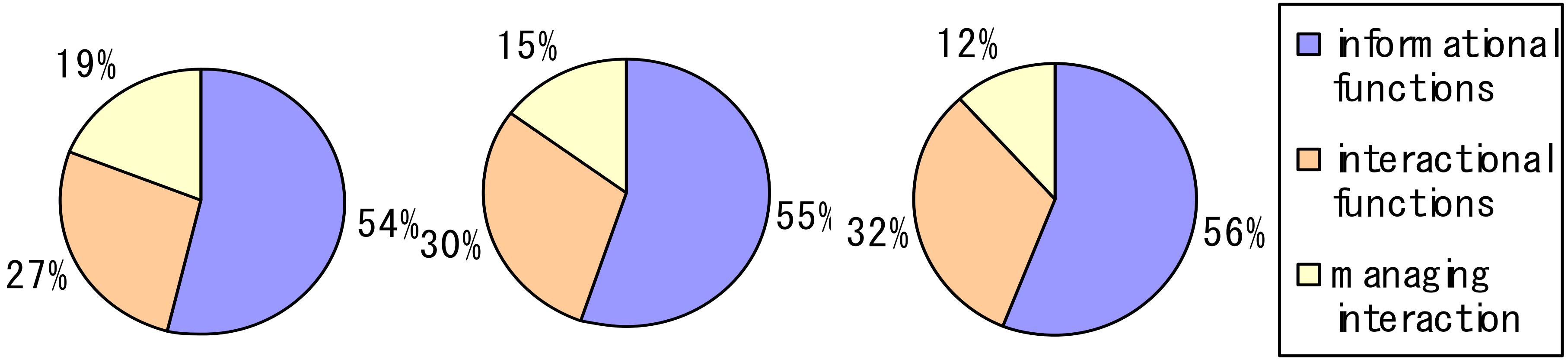
Richer language elicitation

(ffrench, 1999 cited in ffrench, 2003: 413)

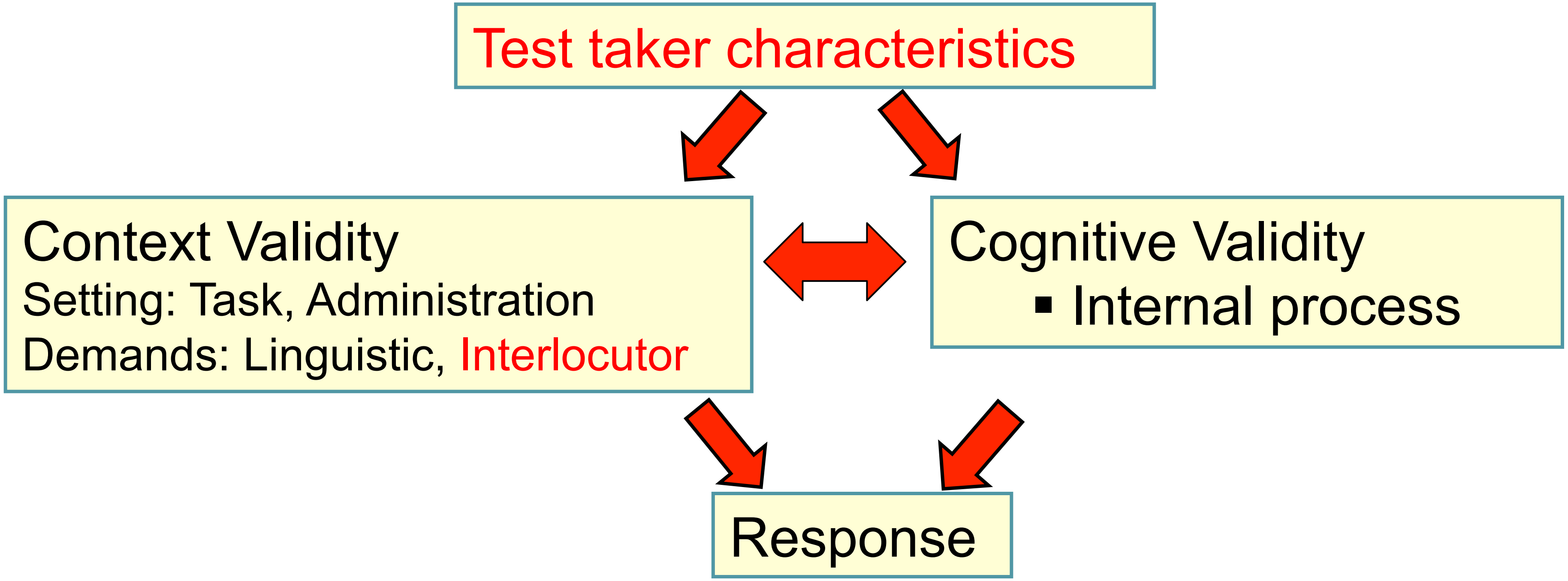
Use of language functions by percentage [individual speaking tests]



Use of language functions by percentage [paired speaking tests]



Issues related to Paired formats: Impact of test-taker characteristics



Test taker characteristics

Context Validity
Setting: Task, Administration
Demands: Linguistic, Interlocutor

Cognitive Validity
- Internal process

Response

Socio-cognitive framework for validating speaking tests
(Weir, 2005; Taylor, ed. 2011)

Test taker characteristics

Gender, Acquaintanceship, Cultural background, L1, Personality, Proficiency level (e.g. Berry, 2007; Nakatsuhara, 2013; Norton, 2005; Ockey, 2006; O'Sullivan, 2008; Van Moere & Bonk, 2004)

References

General Introduction

- **Alderson, J. C., Clapham, C. and Wall, D.** (1995) *Language Test Construction and Evaluation*, Cambridge: CUP.
- **Bachman, L. F.** (1990) *Fundamental Consideration in Language Testing*, Oxford: OUP.
- **Bachman, L. F. and Palmer, A. S.** (1996) *Language Testing in Practice,* Oxford: OUP.
- **Hughes, A.** (2003) *Testing for Language Teachers (second edition),* Cambridge: CUP.
- **McNamara, T.** (1996) *Measuring Second Language Performance*, Harlow: Longman.
- **McNamara, T.** (2000) *Language Testing*, Oxford: OUP.
- **Weir, C. J.** (2005) *Language Testing and Validation: An evidence-based approach*, London: Palgrave Macmillan.

Testing speaking

- **Fulcher, G.** (2003) *Testing Second Language Speaking*, London: Longman.
- **Luoma, S.** (2004) *Assessing Speaking*, NY: CUP.
- **Taylor, L. ed.** (2011) *Examining Speaking,* Studies in Language Testing vol. 34, Cambridge: CUP.

Washback

- **Cheng, L. and Watanabe, Y.** (2004) *Washback in Language Testing –Research Contexts and Methods-*, London, Lawrence Erlbaum Associates.
- **Green, A.** (2007) IELTS Washback in Context, Cambridge, CUP.
- **Messick, S.** (1996) 'Validity and Washback in Language Testing', *Language Testing* 13/3: 241-256.

Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe 2001)

- http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

TEMPUS

Modernising higher education

Thank you! ☺

Dr Fumiyo Nakatsuhara

Dr Chihiro Inoue

CRELLA, University of Bedfordshire

European Commissi

TEMPUS