Enhancing scoring validity of direct speaking tests: Construct(s) measured by live, audio and video rating modes

Fumiyo Nakatsuhara, Chihiro Inoue & Lynda Taylor
CRELLA
University of Bedfordshire

BC New Directions in English Language Assessment Conference 2016

CRELLA
University of Bedfordshire

Acknowledgement

- *This research was funded by the 2014-15 IELTS Funded Research Programme.*

- *Any opinions, findings, conclusions or recommendations expressed in this presentation are those of the presenters and do not necessarily reflect the views of the IELTS partners.*

CRELLA
University of Bedfordshire

RESEARCH BACKGROUND

Double marking in speaking tests

Double marking of spoken performance is essential (e.g. AERA, APA and NCME, 1999)

Practical constraints in large-scale tests

Rating systems of speaking tests include:

- Double marking with 2 'live' examiners (e.g. GEPT, Cambridge main suite)
- Double marking with a live examiner and a post-hoc audio rater (e.g. BULATS, TEAP in Japan)
- Single marking of audio-recorded performance (part-scoring) (e.g. TOEFL, Aptis)
- Single marking but all tests recorded for quality assurance purposes (e.g. IELTS, Trinity)

Rapid advances in computer technology → easier gathering and transmission of test-takers' recordings in a sound and video format

Good moment to investigate different rating modes! ☺

Studies into different rating modes (Pre-2001 IELTS)

- Inter- and intra-rater correlations of audio ratings were higher than those of video ratings (Styles, 1993).
- For 10 out of 27 students, audio rating was a band lower than live rating (Conlan et al., 1994).

Listening perceptions of speech samples delivered by different modes

- Listeners rely on visual information in understanding the spoken text (e.g. Raffler-Engel, 1980; Burgoon, 1994; c.f. McGurk & MacDonald, 1976).
- In listening tests, presenting video could facilitate understanding better than audio-only materials (e.g. Wagner, 2008; 2010), though it may lead to distraction (e.g. Bejar et al., 2000).

To what extent are the constructs assessed under audio and video rating conditions comparable?

Research questions

RQ1: Are there any differences in **examiners' scores** when they assess test-takers' performance under live, audio and video rating conditions?

RQ2: Are there any differences in the volume and nature of **positive and negative features of test-takers' performance** that examiners report noticing when awarding scores under audio and video rating conditions?

RQ3: Are there any differences **in examiners' perceptions** towards test-takers' performance between audio and video rating conditions?

RESEARCH DESIGN

Data collection

IELTS Speaking test:

Part 1	Introduction and interview (4-5 mins)
Part 2	Test-taker long turn (3-4 mins)
Part 3	Examiner and test-taker discussion (4-5 mins)

Analytic rating categories (9 levels):

1) Fluency and Coherence
2) Lexical Resource
3) Grammatical Range and Accuracy
4) Pronunciation

Data collection

Existing data (Nakatsuhara, 2012)

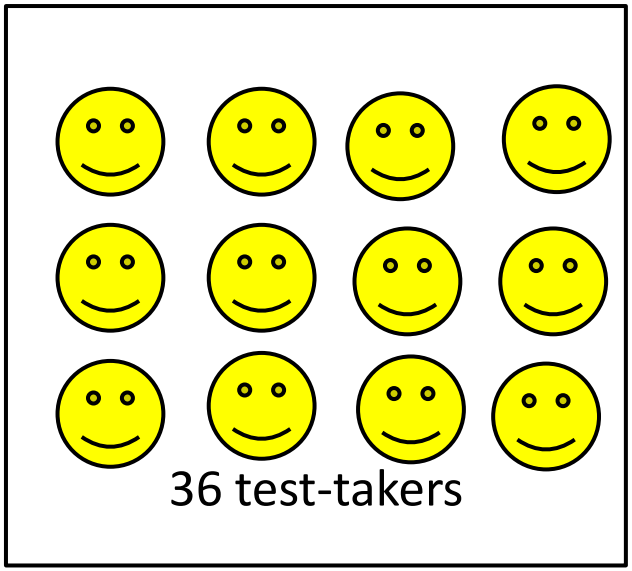
- Audio and video recordings of 36 IELTS Speaking Test sessions (Bands 3.0 to 8.0)
- Live rating scores + Examiners' written comments
- Audio rating scores + Examiners' written comments

Newly collected data

- Video rating scores + Examiners' written comments
- Examiners' (A, D, E, F) verbal reports for 4 audio and 4 video recorded performances (Bands 4.0, 5.0, 6.0 and 7.0)

Examiner ID	A	B	C	D	E	F
Live	x	x	x			
Audio	x	x	x	x		
Video	x			x	x	x

Gathered data and analysis



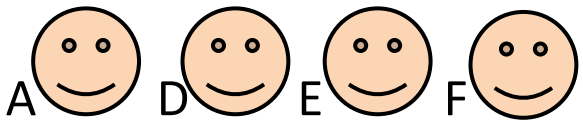
a) Live rating scores + Examiners' comments
b) Audio rating scores + Examiners' comments
c) Video rating scores + Examiners' comments

MFRM analysis on scores (RQ1)

Quantifying the degree of positiveness (RQ2)
- Positive
- Both positive and negative
- Negative

} MFRM analysis

4 Examiners' verbal reports on audio and video performances (4 test-takers each)

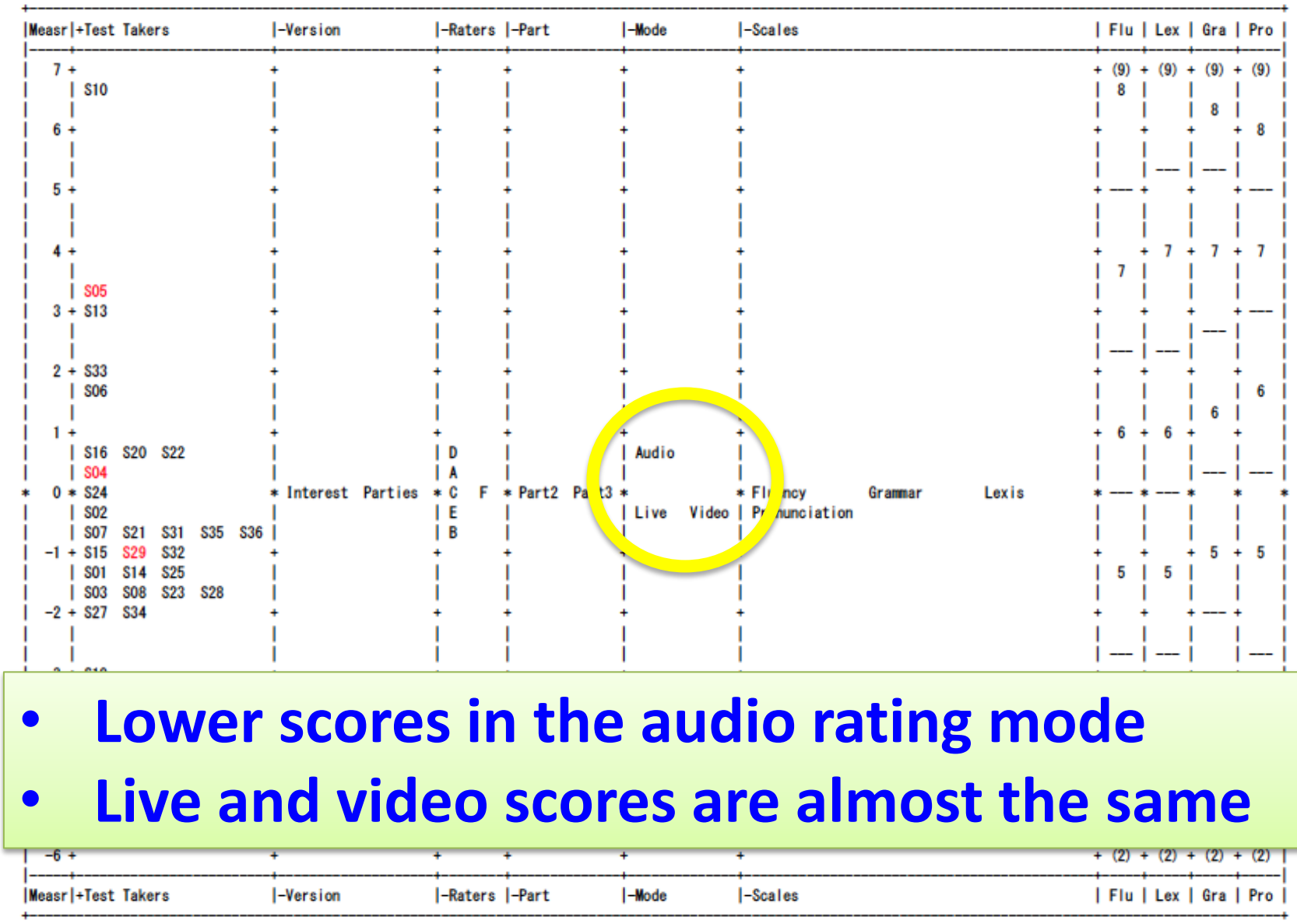


Transcribed + Thematic analysis (RQ3)

RESULTS

Score analysis (RQ1)

6-facet analysis



- **Lower scores in the audio rating mode**
- **Live and video scores are almost the same**

4 sets of 5-facet analysis (within each rating category)

Summary of paired comparisons with fair average scores

	Live	Sig.	Audio	Sig.	Video	Sig.	Live
Fluency	5.32	>	4.91	<	5.06	<	5.32
Lexis	5.05	>	4.63	<	5.08	=	5.05
Grammar	5.05	>	4.63	<	5.09	=	5.05
Pronunciation	5.29	>	4.85	<	5.27	=	5.29

>: Significantly larger than, <: Significantly smaller than, =: No sig difference

- **Lexis, Grammar, Pronunciation: Live=Video>Audio**
- **Fluency: Live>Video>Audio**

Examiners' written comment analysis (RQ2)

Classification of examiner comments

Positive	*e.g. Wide enough range to discuss topics at length. Generally paraphrases successfully (Lexis, Band 6)*
Both positive & negative	*e.g. Wide variety of structures, including subordinate clauses. Some inaccuracies persist. Occasional self-corrections. Generally accurate. (Grammar, Band 7)*
Negative	*e.g. Patches that are unclear and mispronunciations are frequent. (Pronunciation, Band 4)*
Unclassified	*e.g. Possible 5 in latter part but overall 4. (Fluency, Band 4)*

MFRM analysis on examiners' comments

	Measure	Real S.E.	Fair Ave.	Infit MnSq
Fluency (Live)	-.39	.18	2.07	.98
Grammar (Live)	-.36	.21	2.05	1.27
Fluency (Audio)	-.23	.13	1.98	.99
Lexis (Live)	-.14	.18	1.93	1.04
Lexis (Video)	-.05	.13	1.88	.96
Pronunciation (Live)	-.03	.19	1.88	.83
Lexis (Audio)	.03	.14	1.84	1.15
Fluency (Video)	.09	.13	1.81	.77
Grammar (Audio)	.15	.14	1.78	1.08
Grammar (Video)	.27	..	72	.88
Pronunciation (Audio)	.32	.14	1.70	1.02
Pronunciation (Video)	.34	.13	1.69	1.02

More negative

Examiner comments on Fluency

	Measure	Real S.E.	Fair Ave.	Infit MnSq
Fluency (Live)	-.39	.18	2.07	.98
Grammar (Live)	-.36	.21	2.05	1.27
Fluency (Audio)	-.23	.13	1.98	.99
Lexis (Live)	-.14	.18	1.93	1.04
Lexis (Video)	-.05	.13	1.88	.96
Pronunciation (Live)	-.03	.19	1.88	.83
Lexis (Audio)	.03	.14	1.84	1.15
Fluency (Video)	.09	.13	1.81	.77
Grammar (Audio)	.15	.14	1.78	1.08
Grammar (Video)	.27	.13	1.72	.88
Pronunciation (Audio)	.32	.14	1.70	1.02
Pronunciation (Video)	.34	.13	1.69	1.02

More negative

Examiner comments on Lexis

	Measure	Real S.E.	Fair Ave.	Infit MnSq
Fluency (Live)	-.39	.18	2.07	.98
Grammar (Live)	-.36	.21	2.05	1.27
Fluency (Audio)	-.23	.13	1.98	.99
Lexis (Live)	-.14	.18	1.93	1.04
Lexis (Video)	-.05	.13	1.88	.96
Pronunciation (Live)	-.03	.19	1.88	.83
Lexis (Audio)	.03	.14	1.84	1.15
Fluency (Video)	.09	.13	1.81	.77
Grammar (Audio)	.15	.14	1.78	1.08
Grammar (Video)	.27	.13	1.72	.88
Pronunciation (Audio)	.32	.14	1.70	1.02
Pronunciation (Video)	.34	.13	1.69	1.02

More negative

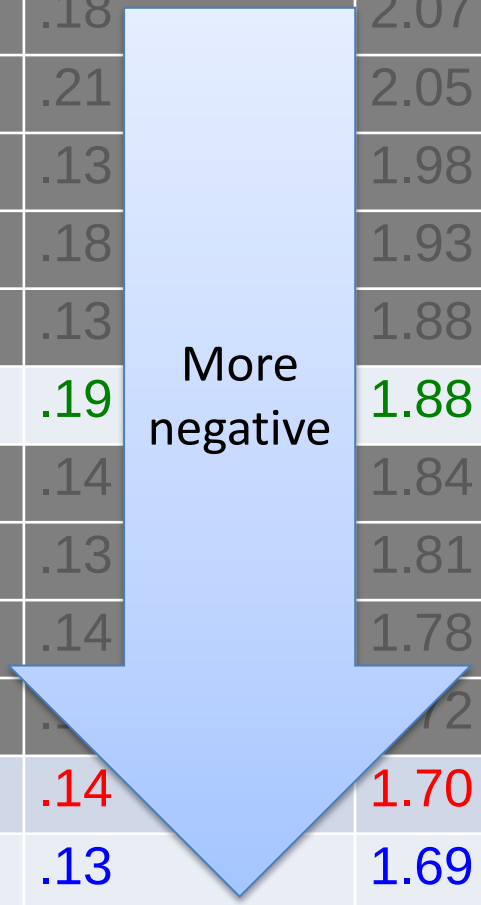
Examiner comments on Grammar

	Measure	Real S.E.	Fair Ave.	Infit MnSq
Fluency (Live)	-.39	.18	2.07	.98
Grammar (Live)	-.36	.21	2.05	1.27
Fluency (Audio)	-.23	.13	1.98	.99
Lexis (Live)	-.14	.18	1.93	1.04
Lexis (Video)	-.05	.13	1.88	.96
Pronunciation (Live)	-.03	.19	1.88	.83
Lexis (Audio)	.03	.14	1.84	1.15
Fluency (Video)	.09	.13	1.81	.77
Grammar (Audio)	.15	.14	1.78	1.08
Grammar (Video)	.27	.1	72	.88
Pronunciation (Audio)	.32	.14	1.70	1.02
Pronunciation (Video)	.34	.13	1.69	1.02

More negative

Examiner comments on Pronunciation

	Measure	Real S.E.	Fair Ave.	Infit MnSq
Fluency (Live)	-.39	.18	2.07	.98
Grammar (Live)	-.36	.21	2.05	1.27
Fluency (Audio)	-.23	.13	1.98	.99
Lexis (Live)	-.14	.18	1.93	1.04
Lexis (Video)	-.05	.13	1.88	.96
Pronunciation (Live)	-.03	.19	1.88	.83
Lexis (Audio)	.03	.14	1.84	1.15
Fluency (Video)	.09	.13	1.81	.77
Grammar (Audio)	.15	.14	1.78	1.08
Grammar (Video)	.27	.13	1.72	.88
Pronunciation (Audio)	.32	.14	1.70	1.02
Pronunciation (Video)	.34	.13	1.69	1.02



MFRM analysis on Examiners' comments

- **Lexis, Grammar, Pronunciation:** no sig difference between Audio and Video
- **Fluency:** more negative comments on Video than Audio

Possibly, influence of a visually-oriented descriptor in Fluency: 'is willing to speak at length…' (Band 6)

➢ 16 comments on (un)willingness in Video (8 in Audio)

e.g. Seems unwilling to develop turns…

Appears willing to produce long turns

Summary of the findings so far

Score analysis (RQ1)

- Audio ratings were lower than live and video ratings.

Examiners comments analysis (RQ2)

- Both audio and video ratings directed examiners' attention to the similar numbers of negative performance features of test-takers.

➢ Examiners under the video condition seemed to use such negative evidence in moderation in awarding scores.

Verbal report analysis (RQ3)

4 main themes emerged:

1. Video providing a fuller picture of communication
2. Possible difference in scores between two modes
3. Different examining behaviour / attitudes between two modes
4. Implications for future double-rating methods

1. Video providing a fuller picture of communication

1.1. Video helping examiners understand what test-takers are saying

- *Without the video, I think I would have misunderstood quite a lot of what she said without the visual help of 'belt' and 'fighting', I may not have understood what she was talking about (S29, Examiner F, Video).*

1. Video providing a fuller picture of communication

1.2. Video giving more information beyond what test-takers are saying

- *She did look quite animated for a little bit there, some good eye contact and she used her hands a bit, but then she goes back to leaning on her elbow, putting her face in her hand. She's either not interested or not motivated or just out of her depth (S09, Examiner E, Video).*

1. Video providing a fuller picture of communication

1.3. Video helping examiners understand what test-takers are doing when dysfluency or awkwardness occurs

- *Some disfluency here, and you can tell from her face it's because she doesn't really understand 'celebrations', it's showing that it's a lack of comprehension rather than thinking of the ideas (S04, Examiner E, Video).*

2. Possible difference in scores between two modes

2.1. Different features are noticed/ attended/ accentuated in the two modes

- *The mistakes he's making are much more obvious here [in audio]… I hear lots of little pauses and gaps. Though my impression [in video] was that he was fluent, here my impression is that he is hesitant. He is almost like a different person (S05, Examiner F, Audio).*

2. Possible difference in scores between two modes

2.2. Comments directly related to scoring

- *Well, maybe I'm more critical of the lower-level students with the video because I can see that they're not understanding, whereas if you're just listening, it could be just searching for the right words or content, rather than not understanding. The higher-level students, I possibly mark them a bit higher with the video because I can see how relaxed they look and how good their body language is (Examiner E, General comments).*

3. Different examining behaviour / attitudes between two modes

- *I feel I can concentrate a lot more when I don't have the visual input (Examiner D, General comments).*

- *She's signalling "I'm thinking. I'm going to give you an answer in a second, just as soon as I get it in my head." You can see where she's keeping her mouth open, so she is indicating, "I haven't finished" (Examiner F, General comments).*

4. Implications for future double-rating methods

4.1. Preferred mode of double-rating

- One examiner preferred the video mode because it offered a more rounded picture of communication and she was more confident in her scores.
- Two examiners preferred the audio mode because they were used to double-rating with the audio, and that is how they were trained to double-rate.
- The other examiner stated that she did not have any preference, and it was just a matter of getting used to either mode.

4. Implications for future double-rating methods

4.2. Implications for examiner training and standardisation

- *We do standardisation and we do all audio, but when we do the training, we do it with video (Examiner F, S29, Part 3, Audio).*
- *In some ways, we're having to do so much [during live exams] that we're not really taking in much of that, so maybe when it's live, I'm not sure how many of the other cues I'm getting. There's so much of the swan on the water that's paddling…and I'm not sure if there's much mental space left to take in non-verbal cues as well (Examiner D, General comments).*

CONCLUSIONS

Main Findings

Rating scores under live, audio and video conditions (RQ1)

- Audio ratings were lower than live and video ratings.

Examiners comments analysis (RQ2)

- Examiners in both audio and video modes noticed the similar numbers of negative performance features, but in video, rating scores were not affected as much.

Examiner perceptions towards audio and video modes (RQ3)

- Visual information helped examiners

a) to understand what the test-takers were saying,

b) to see what was happening beyond what the test-takers were saying, and

c) to understand with more confidence the source of test-takers' hesitation, pauses and awkwardness in their performance.

Implications

1. The constructs tested under the video condition are much closer to those under the live test condition, and the audio rating seems to assess narrower constructs than video rating.

To what extent should non-linguistic features be considered?

2. In video rating, examiners seemed able to provide more informed judgements (rich visual information, no time pressure).
3. A number of examiners' verbal reports related to the fluency and pronunciation features. → *importance of visual information for assessing these features?*
4. Double rating with video seems more appropriate, as long as the test aims to assess the wider constructs of face-to-face interaction with paralinguistic and reciprocal features.
5. Making the rating modes consistent by using videos would make training and standardisation of examiners more effective.

THANK YOU!

Fumiyo Nakatsuhara, Chihiro Inoue & Lynda Taylor
CRELLA
University of Bedfordshire