

Interpreting scores obtained from a technology-based and face-to-face EAP speaking test.

Vivien Berry (British Council)

Fumiyo Nakatsuhara, Chihiro Inoue

(CRELLA, University of Bedfordshire)

Evelina Galaczi (Cambridge English Language Assessment)



Acknowledgement

*This research was funded and supported
by the IELTS Partners: British Council,
Cambridge English Language Assessment and
IDP (IELTS) Australia.*

RESEARCH BACKGROUND

Background

Since distance-learning through video-conferencing systems is becoming more common in tertiary education, the construct of EAP tests may need to be reconsidered. This research sheds some light on the construct of Speaking across different modes of delivery.

Different speaking test delivery modes

- **f2f mode:** Cambridge English General English tests, IELTS
- **Computer-delivered mode (administered by computer only):** Aptis, PTE, TOEFL iBT
- **Both modes:** BULATS

An alternative?

A computer-delivered f2f mode administered by an examiner on the other side of the screen, to combine advantages of both these modes.

Video-conferencing (VC) delivery mode:

- Exploratory study by the US Defense Language Institute Foreign Language Center (Clark & Hooshmand, 1992, 1992)
- Operational VC test: Computer-based Standard Speaking Test (ALC, 1999 – but no longer available)
- More recent research by Craig & Kim (2010), Kim & Craig (2012) and Zhou (2015)

Rationale

- Rapid advances in online video communication technology: e.g. Skype, Facetime, Zoom



- Ensuring accessibility and fairness in remote areas where frequent delivery of f2f tests is not possible
- Making use of recent advances in technology to offer computer-delivered tests without losing interactional features / responding to the broadening construct of speaking (increasing use of online video communication)

3 Phases of the Project

- **Phase 1 (2014):** 32 test-takers, 4 examiners in London

To explore similarities and differences in 2 modes in terms of scores, linguistic output, examiner behaviour and test-takers' and examiners' perceptions

- **Phase 2 (2015):** 99 test-takers, 10 examiners in Shanghai

To confirm the Phase 1 results, after providing examiner/test-taker training. Focus on more thorough statistical analysis of scores, effects of training, the impact of sound quality

- **Phase 3 (in progress):** c.100 test-takers, 8 examiners in Bogota, Buenos Aires, Caracas and Mexico City

To confirm the Phase 1 and 2 results, after developing a bespoke platform

Research Questions (Phase 2)

Comparing the standard f2f and VC mode:

- **RQ1:** Are there any differences in **test-takers' scores**?
- **RQ2:** Are there any differences in test-takers' **linguistic output**, specifically types of **language function**?
- **RQ3:** To what extent did **sound quality** affect performance on the test?
 - a) as perceived by test-takers, examiners and observers?
 - b) as found in test scores?
- **RQ4:** How effective was the **training** for the VC test?
 - a) for examiners as interlocutors/raters
 - b) for test-takers
- **RQ5:** What are **the examiners' and test-takers' perceptions** of the two delivery modes?

RESEARCH DESIGN

- mixed methods approach -

Data Collection

- **Test-takers:** 99 test-takers (CEFR B1 – C1) taking both f2f and VC-delivered tests
- **Examiners:** 10 trained examiners (Examiners A – J)
- **Observers:** 8 PhD students trained as observers
- **IELTS Speaking test structure:**

Part 1	Intro. and interview	4-5 mins.	6 test versions (randomly selected)
Part 2	Individual long turn	3-4 mins.	
Part 3	Two-way discussion	4-5 mins.	

- **Counter-balanced:** delivery mode
- **Rating criteria:** 1) Fluency and coherence; 2) Lexical resource; 3) Grammatical range and accuracy; 4) Pronunciation
- **Double rating:** All test sessions double-rated with video-recorded performances; Sufficient connectivity achieved through an examiner rating matrix

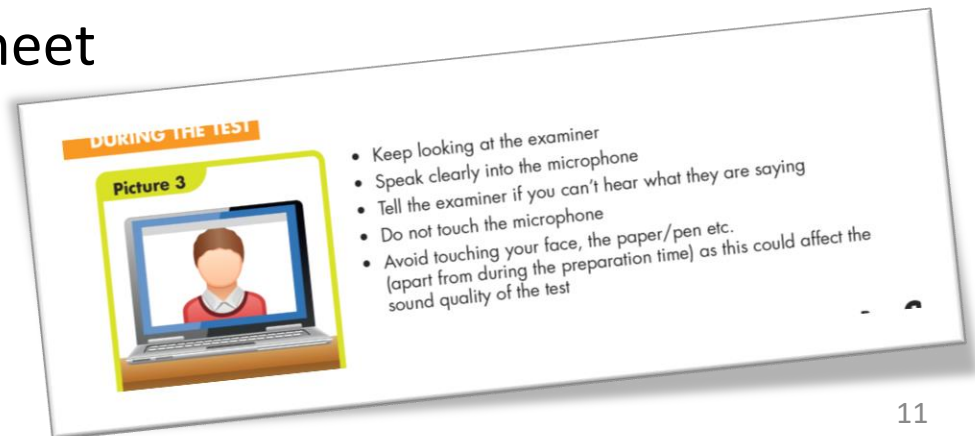
Outline of Training for VC

New VC Examiner training package:

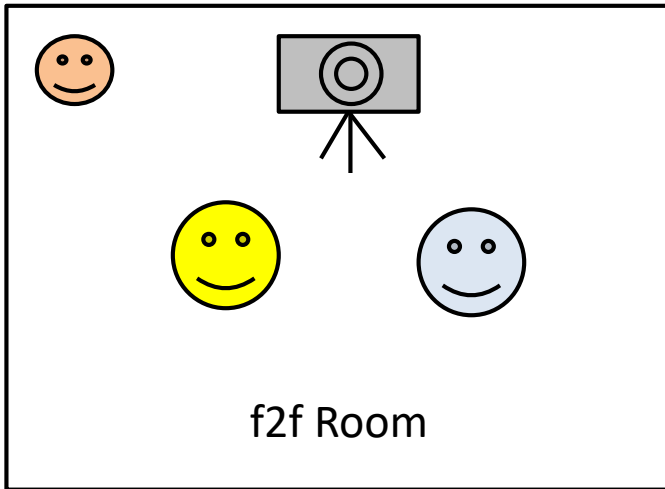
- What does not change and what does btw f2f & VC
- VC-specific interview techniques (e.g. active listening)
- Rating under VC (rating practice with 3 video-recorded tests)
- Peer practice (taking turns to do examiner, candidate and observer)
- Feedback on group practice sessions and discussion

Candidate guidelines:

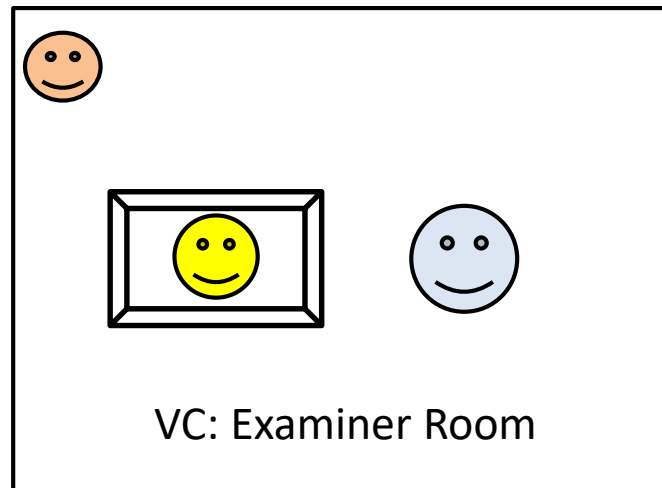
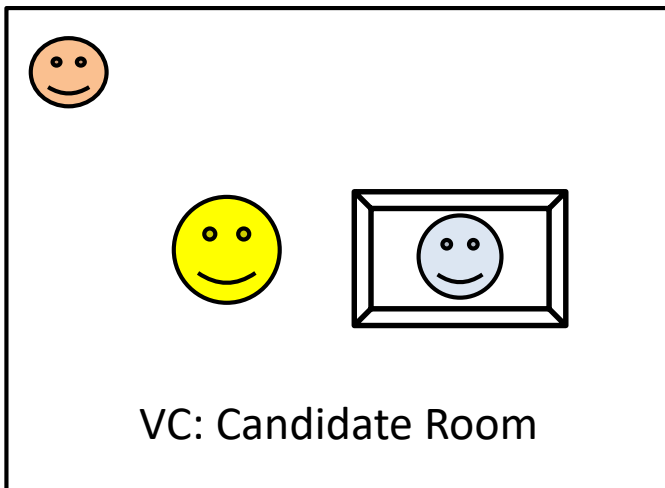
- Bilingual information sheet with clear instructions, what to expect



Gathered Data



- a) Video-recorded test performances
- b) Scores (by Examiner) on both modes
- c) Test-takers' responses to feedback questionnaire & interview
- d) Examiners' responses to feedback questionnaire
- e) Observation notes
- f) Examiners' focus group discussions



Data Analysis: Mixed-methods approach

Score comparison
(RQ1)

Language functions using
O'Sullivan et al.'s (2002)
Observation Checklist (RQ2)

Sound quality
perception (RQ3)

- a) Video-recorded test performances
- b) Scores (by Examiner) on both modes
- c) Test-takers' responses to feedback questionnaire & interview
- d) Examiners' responses to feedback questionnaire
- e) Observation notes
- f) Examiners' focus group discussions

Training effect on test-takers
and examiners (RQ4)

Examiners' and test-takers'
perceptions of the 2 modes (RQ5)

A blue rectangular button with rounded corners and a subtle gradient, featuring the word "RESULTS" in white capital letters.

RESULTS

Test Scores (RQ1) FACETS Variable map (5 facets)

Measr	Test Takers																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																		</
-------	-------------	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	----

	Measr	Real S.E.	Observed Ave.	Fair (M) Ave.	Infit MnSq
f2f	-.12	.08	5.17	5.20	.89
vc	.12	.08	5.12	5.16	.89
Fixed (all same) chi-square: 4.8, d.f.: 1, significance: .03					

0.04

4-facet analysis on each rating category

	Mode	Measr	Real S.E.	Obsrvd Ave.	Fair (M) Ave.	Infit MnSq	Fixed chi-square
Fluency	f2f	-.11	.17	5.11	5.10	.76	$\chi^2=.8$ p=.38
	vc	.11	.17	5.07	5.08	.76	
Lexis	f2f	-.20	.16	5.11	5.08	.70	$\chi^2=3.1$ p=.08
	vc	.20	.16	5.04	5.03	.83	
Grammar	f2f	-.20	.20	5.23	5.21	.86	$\chi^2=2.1$ p=.15
	vc	.20	.20	5.17	5.15	.78	
Pronunciation	f2f	-.14	.18	5.24	5.29	.76	$\chi^2=1.2$ p=.28
	vc	.14	.18	5.19	5.24	.76	

No sig. difference in scores between 2 modes

Language Functions (RQ2)

- Only one function was used by significantly more test-takers under VC mode:
 - *Ask for clarification (Part 1)* VC (63%) > f2f (27%)
(c.f. *comparing* and *suggesting* also showed sig. differences in Phase 1)
- No sig. differences in Parts 2 and 3

More similarities in elicited functions than Phase 1

Sound Quality Perceptions (RQ3)

The same 2 questions to Examiners, Test-takers and Observers in the VC rooms

Q1. The quality of the sound in the VC test [1. Not clear at all - 5. Very clear]

Examiners (M=4.36) = Observers (M=4.36) > Test-takers (M=3.71)

Clear - Very clear

OK - Clear

Q2. The effect of sound quality on candidates' performance [1. No - 5. Very much]

Test-takers (M=2.52) > Observers (M=1.66) = Examiners (M=1.54)

Not much – somewhat

No – Not much



Personally I was amazed at how few technical problems there were.
In my experience that wasn't what I was expecting. (Examiner F)

[Some problems reported in 18 of 99 tests] Occasional millisecond freezes (Examiner D), Slight skip – missed a word (Examiner J)
Computer shut down between end Part 1 and Part 2 (Examiner H)



No difference related to test-takers' proficiency levels: High (Band 6 and above), Middle (Band 5 and 5.5), Low (Below 5).

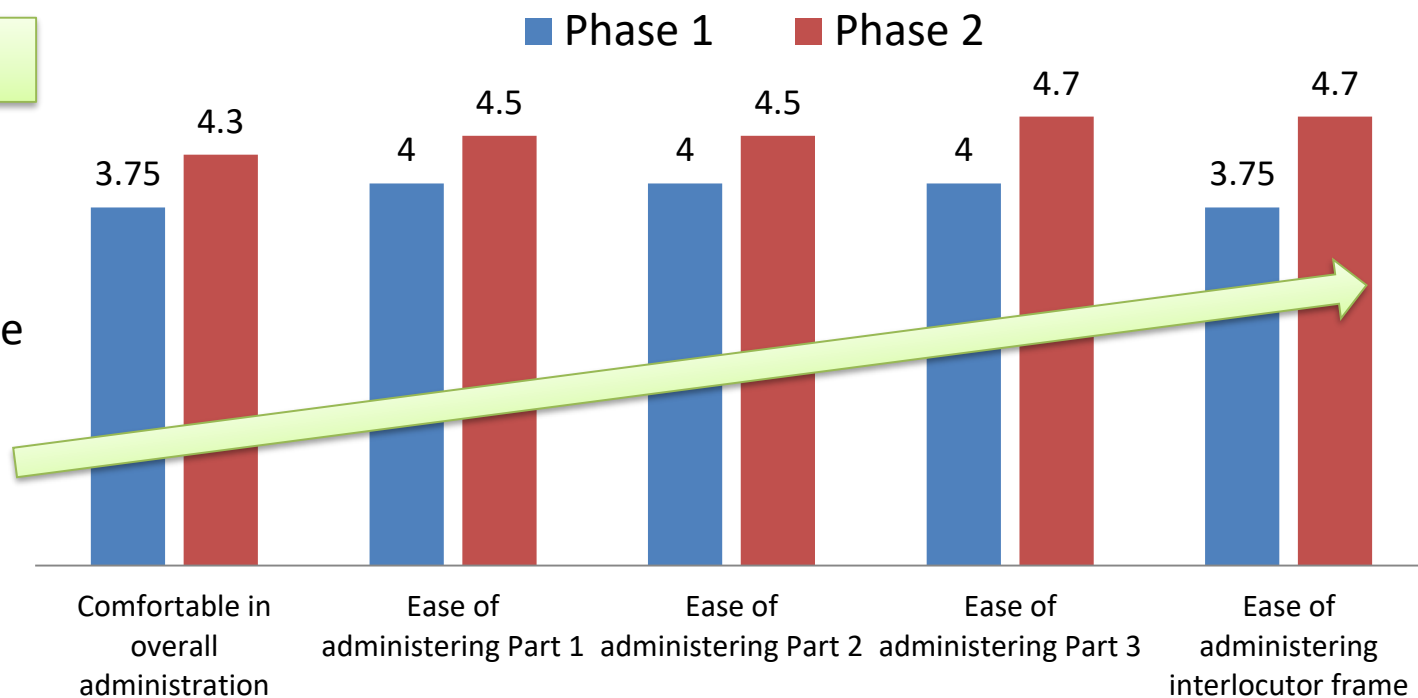
Effectiveness of VC Training (RQ4)

The VC examiner training... [1. strongly disagree – 5. strongly agree]	Mean (SD) N=10
adequately prepared me for administering the test	4.70 (0.48)
gave me confidence in handling the interlocutor frame	4.90 (0.32)
adequately prepared me for rating candidate performance	4.30 (1.06)
gave me confidence in the accuracy of my ratings	4.10 (1.10)

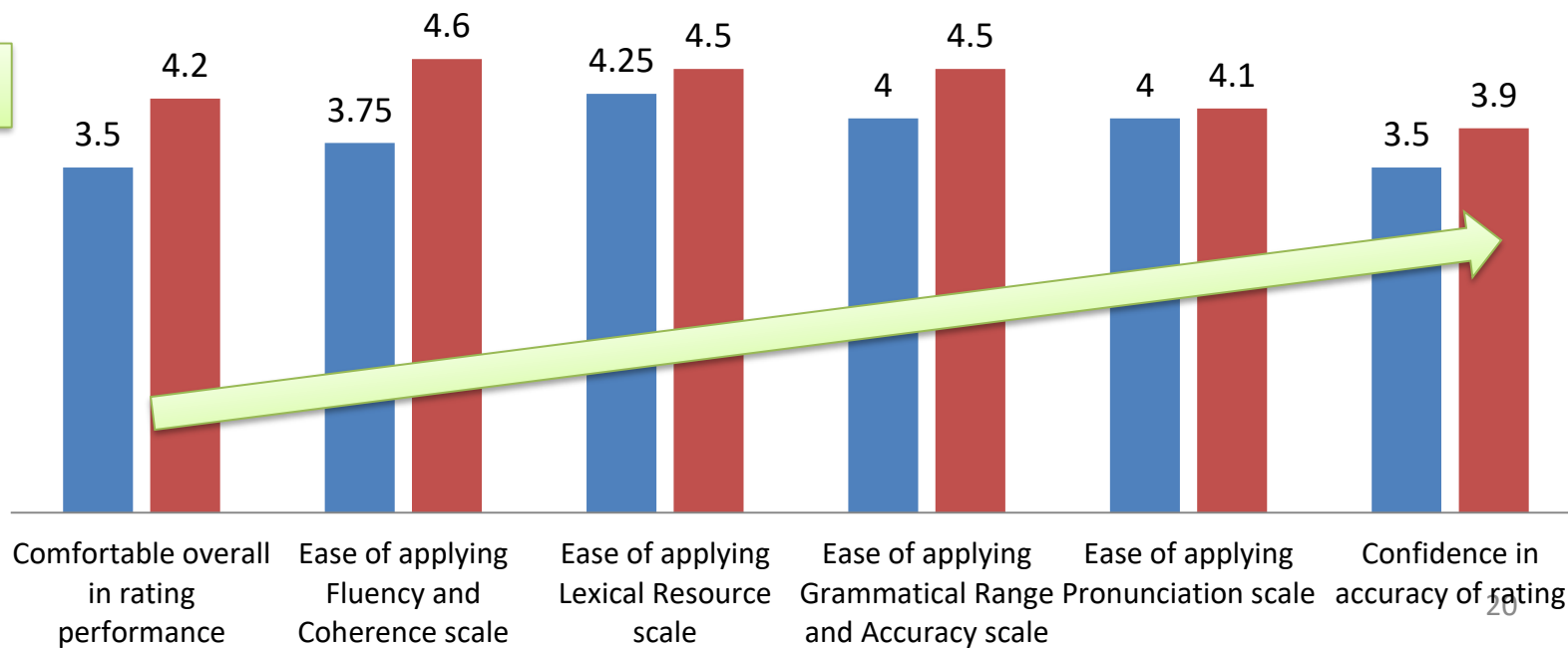
The VC Candidate guidelines	Mean (SD) N=99
Were the candidate guidelines for the VC test ... (1: Not useful; 3: OK; 5: Very useful)	3.87 (0.99) OK – Useful
Were the pictures in the guidelines... (1. Not helpful – 3. OK – 5. Very helpful)	3.65 (1.17) OK – Helpful

Administration

1. Strongly disagree
- 5. Strongly agree

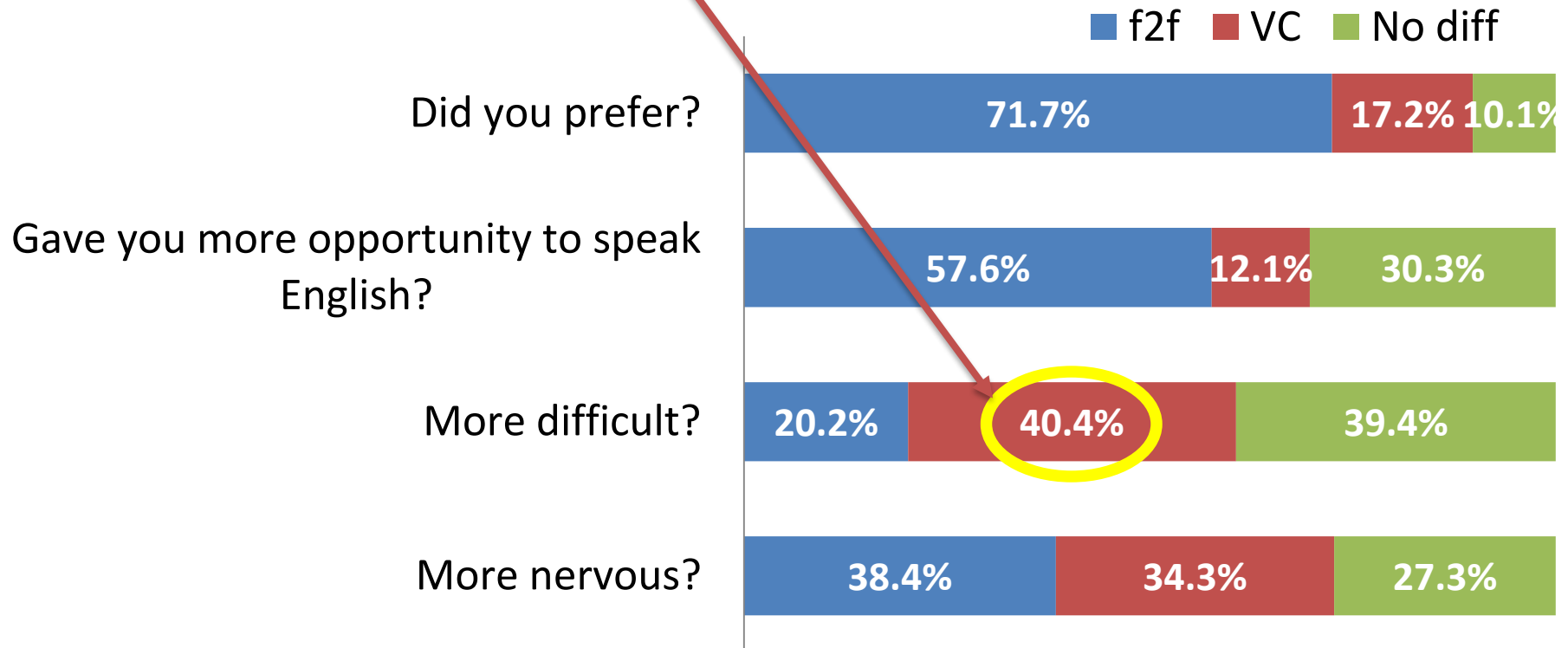


Rating



Test-takers' Perceptions of 2 Modes (RQ5)

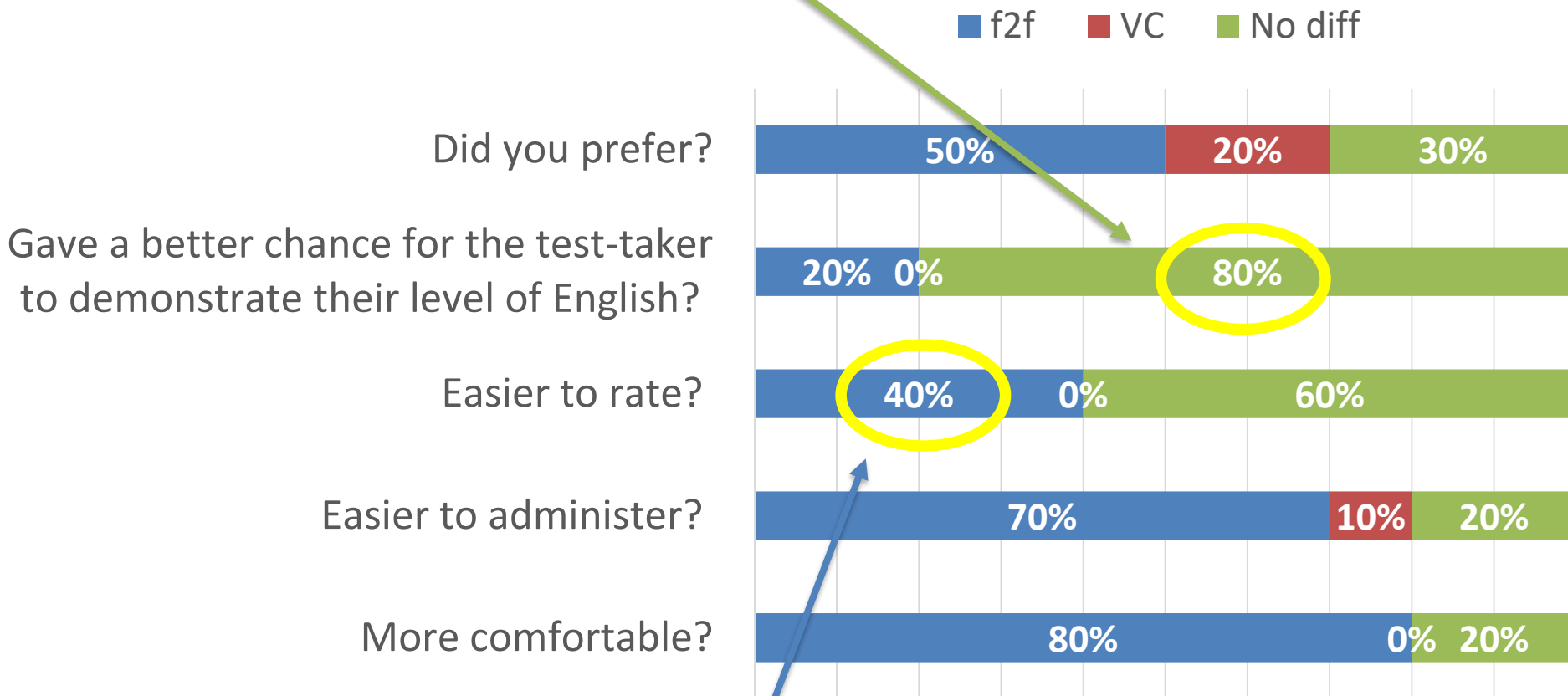
65.6% found VC was more difficult in Phase 1



f2f preferred; but generally improved compared to Phase 1

Examiners' Perceptions of 2 Modes (RQ5)

Only 50% reported no difference in Phase 1



75% found f2f rating was easier in Phase 1

Selected examiner comments & observer notes



My preference of F2F relates to my familiarity with doing it.
(Examiner D)



Video-conferencing makes it harder to be subtle. I don't feel I can use my voice as forcibly or as subtly.
(Examiner J)



When the VC candidate misheard a word, the examiner did not say the candidate was wrong, but used 'yes, but...' to guide her back to the topic.
(Observer 7 in VC-E Room)



It seemed the interactiveness depends a lot on the candidate. If the candidate is interactive, the VC conversation will appear so as well.
(Observer 5 in VC-E Room)



They just didn't seem to have a problem with the VC test at all. They were quite happy with both, the digital generation. (Examiner G)

CONCLUSIONS

Main findings of Phase 2

- **RQ1 (Scores):** Sig. difference in overall scores, but the difference was negligibly small (0.04 of a band); No sig. difference in analytic scores
- **RQ2 (Language functions):** more candidates asked for clarification on VC in Part 1
- **RQ3 (Perceived sound quality):** Examiners = Observers > Test-takers (but no difference between stronger and weaker test-takers)
- **RQ4 (Effectiveness of training):** Positively received and proved useful
- **RQ5 (Examiner/test-taker perceptions):** More positive responses on VC than Phase 1

Implications

- **In general, Phase 1 results confirmed**
 - Comparable scores, though some differences in one of the language functions and test-takers' and examiners' perceptions
- **Need for developing a bespoke platform**
 - To improve technical/sound quality & to make test administration easier
- **Revise training materials accordingly**

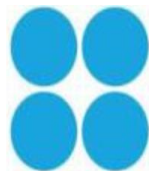
Phase 3 (in progress) with c.100 test-takers, 8 examiners in Bogota, Buenos Aires, Caracas and Mexico City

- **Aims:** To confirm the Phase 1 and 2 results, after developing a bespoke platform and after revising training materials

Thank you!



University of
Bedfordshire



**BRITISH
COUNCIL** Language
Assessment
Research



CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge