CAMBRIDGE

Where
your world
grows

Research Notes 86

Exploring video-call paired speaking tests

ISSN: 1756-509X
December 2023

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Issue 86/December 2023

Guest Editors

Evelina Galaczi, Director of Research, Cambridge University Press & Assessment

Hye-won Lee, Senior Research Manager, Cambridge University Press & Assessment

Editor

John Savage, Research, Cambridge University Press & Assessment

Typeset in the United Kingdom by Abound Creative

Contents

Insights into video-call paired speaking tests: Research background and design

Hye-won Lee, Research, Cambridge University Press & Assessment
Evelina Galaczi, Research, Cambridge University Press & Assessment

Introduction

In this special issue of *Research Notes*, we explore the feasibility of assessing paired speaking via video-call (VC) from multiple perspectives in the context of the Speaking paper in the Cambridge English Qualifications (CEQs). Traditionally, speaking assessment has been delivered and rated by human examiners under circumstances where test-takers physically co-locate with their examiners. Such tests can tap into a wider range of the construct including bi- or multilateral interaction; however, they pose complexity and challenges around logistics of test administration and global accessibility. With the advancement of technology, it has become possible to automate test delivery and rating and maximise operational practicality through computer-delivered monologic speaking tests; however, such assessment formats are constrained by what the technology available at the time can deliver and are often limited to 'narrower' tasks in construct coverage.

Against the backdrop of limitations in the existing test approach, assessing speaking via VC has received attention over recent years, being seen as a viable option to maximise logistical efficiency yet preserve the interactional nature of the speaking construct. A handful of research has been conducted to acquire an understanding of this new test delivery mode, most of which involved comparing it to the in-person counterpart from various aspects – test scores, linguistic outputs, and stakeholder perceptions. A few key findings have emerged:

- comparable holistic and analytics scores were found between the two delivery modes (e.g., Craig and Kim 2010, Nakatsuhara, Inoue, Berry and Galaczi 2017)
- in test-taker spoken responses, more frequent use of 'asking for clarification' and 'reformulation' were elicited in the VC mode (e.g., Cooke 2015. Nakatsuhara, Inoue, Berry and Galaczi 2021)
- more effort on turn management was perceived in the VC mode by the examiners; both the examiners and test-takers noticed different patterns of non-verbal behaviour in the VC mode (e.g., Kim and Craig 2012, Lee, Patel, Lynch and Galaczi 2021).

The learnings about VC speaking tests thus far are primarily from individual contexts, where only a one-on-one interaction takes place between one examiner and test-taker each. They cannot be simply extrapolated to more complex contexts such as a paired test-taker format of the CEQs Speaking test, which necessitates a need to undertake a programme of research looking into how the VC format may have an impact on the tests involving multiple speakers. Besides contributing to knowledge advancement in the subject, findings from the research would offer empirical evidence to support validity claims of the new delivery mode and gain insights for when operational roll-out is considered.

With this background, a research project was launched and aimed at addressing the following questions:

In a comparison between the in-person (F2F) and VC mode of paired speaking tests, are there differences in:

- Test-taker scores?
- Language functions elicited in the test-taker language?
- Examiner and test-taker perceptions?

A summary of the three research strands collectively was discussed in Lee, Mullooly, Devine and Galaczi (2023), focusing specifically on similarities and differences in the interactional nature of the test. In contrast, the current issue reports more details from each strand, presented in an individual article, and covers a wider range of aspects in depth.

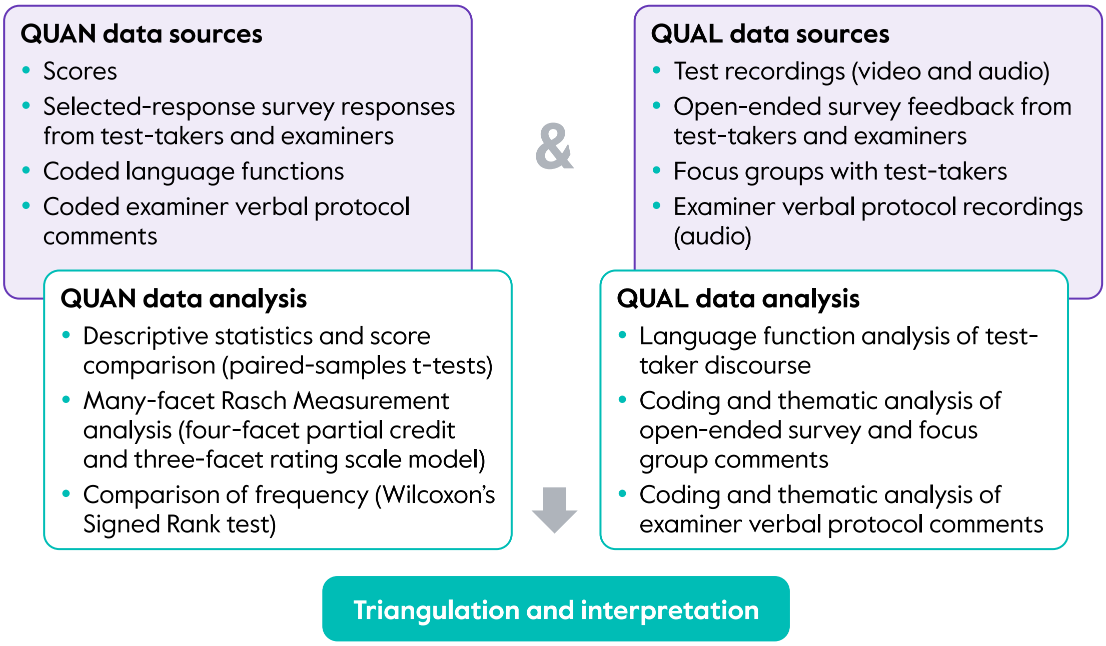
For a comprehensive investigation of the VC mode in its own right, the following aspects of the test are also explored:

- examiner rating and administration behaviour during the VC test
- operational considerations in reflection of the issues identified and observed in the VC test.

Please note: discussion does not encompass aspects relating to test security as it is outside the scope of the current research project but needs to be thoroughly examined before operationalising the VC mode at scale.

Overall research design

The five areas of investigation were addressed under the mixed methods research design visually summarised in Figure 1.



QUAN data sources

- Scores
- Selected-response survey responses from test-takers and examiners
- Coded language functions
- Coded examiner verbal protocol comments

&

QUAL data sources

- Test recordings (video and audio)
- Open-ended survey feedback from test-takers and examiners
- Focus groups with test-takers
- Examiner verbal protocol recordings (audio)

QUAN data analysis

- Descriptive statistics and score comparison (paired-samples t-tests)
- Many-facet Rasch Measurement analysis (four-facet partial credit and three-facet rating scale model)
- Comparison of frequency (Wilcoxon's Signed Rank test)

QUAL data analysis

- Language function analysis of test-taker discourse
- Coding and thematic analysis of open-ended survey and focus group comments
- Coding and thematic analysis of examiner verbal protocol comments

Triangulation and interpretation

Figure I: Mixed methods research design

Various types of both quantitative and qualitive data were collected and triangulated to develop a fuller understanding. The study was conducted first in the context of B2 First (henceforth B2), which is the most widely administered test in the CEQs, followed by both A2 Key (henceforth A2) and B1 Preliminary (henceforth B1) simultaneously to examine if the findings from the B2 trial could be generalised to the lower proficiency levels. Details of the collected data and how they were analysed will be discussed in each corresponding article of the issue, but before moving onto individual pieces, we would like to describe the features of the study that are shared across all research strands discussed in the remainder of this collection of articles.

Test materials

As mentioned above, the study focused on the first three levels of the CEQs: A2, B1 and B2. The 'for Schools' version of A2 and B1 were used, as they were an appropriate match for the average age of the test-taker participants for both tests. The Speaking paper of A2 for Schools lasts 8 to 10 minutes and is in two parts, each containing two phases (www.cambridgeenglish.org/exams-and-tests/key-for-schools/exam-format). In the first part, the interlocutor asks the test-takers interview-type questions, while Part Two focuses on discussion. That of B1 for Schools, lasting 12 minutes, includes these elements as well as an individual 'long turn' (monologue) for each candidate of up to one minute (www.cambridgeenglish.org/exams-and-tests/preliminary-for-schools/exam-format). B2 Speaking, taking the longest (14 minutes) among the three, follows a similar test format as in the B1 level (www.cambridgeenglish.org/exams-and-tests/first/exam-format). Table 1 below summarises the format of the Speaking paper for the three levels.

Table 1: Speaking exam formats for the three levels

A2 for Schools		B1 for Schools		B2	
Part	**Time required**	**Part**	**Time required**	**Part**	**Time required**
Interview	3–4 mins	Interview	2 mins	Interview	4 mins
		Extended turn	3 mins	Long turn	3 mins
Discussion	5–6 mins	Discussion	4 mins	Collaborative task	3 mins
		General conversation	3 mins	Discussion	4 mins

Examiners used the current Cambridge Assessment English Speaking Test Assessment Scales: Grammar and Vocabulary (GV), Pronunciation (P), Interactive Communication (IC), and in the B1 and B2 test Discourse Management (DM) . These can be found at www.cambridgeenglish.org/Images/563269-a2-key-for-schools-speaking-assessing-speaking-performance.pdf, www.cambridgeenglish.org/Images/563276-b1-preliminary-assessing-speaking.pdf, and www.cambridgeenglish.org/images/168619-assessing-speaking-performance-at-level-b2.pdf. The following table provides brief descriptions for each assessment criterion.

Table 2: Speaking test assessment scales

Assessment criteria	Description
Grammar and Vocabulary	Accurate and appropriate use of a range of grammatical forms and vocabulary
Discourse Management (B1 and B2 only)	Coherence, extent and relevance of each test-taker's individual contribution, whether in monologue or dialogue
Pronunciation	Ability to produce intelligible utterance to fulfil the task requirements
Interactive Communication	Ability to take part in the interaction appropriately using language to achieve meaningful communication

To ensure that all test-takers completed different tasks for their F2F and VC test, two test versions were used for each level. All test materials were reviewed by a group of experienced assessment staff prior to use. This was done to ensure all materials were suitable, accessible and representative of the typical test content. At test-level, consideration was also given to making sure both tests were of comparable difficulty in order not to unduly impact on the scores received. Minor modifications were made to interlocutor frames for online delivery. These were primarily concerned with clarity of instruction in the VC environment.

Participants

Test-takers

In total, 242 volunteer test-takers took part in this study: 92 A2 for Schools, 92 B1 for Schools, and 58 B2 test-takers from six different global locations in total – Italy, Mexico, Portugal, Romania, Spain, and Vietnam (see Table 3). They were learners studying towards taking their live exam, with most due to do so within a month of participating in the study.

Table 3: Participant demographics by country

Country	A2 for Schools	B1 for Schools	B2
Italy	26	30	23
Mexico	14	12	0
Portugal	0	0	4
Romania	16	16	0
Spain	26	28	31
Vietnam	10	6	0
Totals	**92**	**92**	**58**

Additional background information was collected through a survey, which was completed by 57 A2 for Schools, 70 B1 for Schools, and 44 B2 test-takers. Within that sample of 171, the majority of test-takers were between 13 to 15 years old for the 'for Schools' version and 16 to 25 years old for B2, with the youngest aged 10 to 12 years old for the 'for Schools' version and around 14/15 years old for B2; 76 were male and 89 female; their first languages included Spanish (37%), Italian (35%), Romanian (15%), Vietnamese (9%), Catalan (3%), and Portuguese (1%). Overall, the demographic information on the study participants reflects that of the exam population in terms of age and gender, making the findings and conclusions drawn broadly generalisable to A2 for Schools, B1 for Schools, or B2 Speaking.

Examiners

33 certificated Speaking Examiners (SEs) in total participated in the research. All the examiners reported more than six years of English Language Teaching experience and a majority had five or more years' experience as A2 for Schools, B1 for Schools, and/or B2 Speaking Examiners. More than a third of them were Team Leaders as well as SEs, i.e. they had higher levels of experience and training.

Participants' experience with the internet and VC technology

The use of VC technology was an important background variable and was focused on in all surveys administered to the participants.

In the B2 trial, most test-takers were using VC regularly or occasionally for socialising or for their studies. However, some test-takers had never used VC for these purposes (20% and 14% reported never using VC for socialising or studying respectively). Only two examiners reported never using VC for social purposes and only three reported never using VC in teaching.

In the follow-up A2/B1 for Schools trial, by far the most frequent use of VC software was for study purposes, with only nine test-takers (7%) reporting they never used VC for this purpose. Only one examiner reported never using VC for social purposes versus 17 examiners (58%) who reported social use of VC one to two times a week or more. Similarly, as one might anticipate given the frequency of Covid-19 lockdowns and the closure of school sites, most examiners (82%) reported using VC for teaching one to two times a week or more.

Based on these findings, it can be assumed that the participants in these studies had enough familiarity with VC technology, and that lack of familiarity would not play a major role in the findings reported.

Test arrangements

Prior to delivering the tests, all the examiners attended an initial familiarisation session for the VC mode. This online training covered some minor adaptations to the test materials from the standard F2F mode, as well as guidance for delivering the different parts of the test in the VC mode and completing the online marks form.

All tests were completed between July and August 2020 for the B2 trial and between April and June 2021 for the A2/B1 for Schools trial. They were carried out across 13 test centres (four in Italy, four in Spain, two in Mexico, one in Portugal, one in Romania, and one in Vietnam). F2F tests were held at each centre, under strict adherence to all relevant local social distancing regulations in place at the time. VC tests were generally conducted with the four participants (two test-takers and two examiners) all in different locations. Although in some instances the assessor, who is responsible for awarding analytic scores, was present in the same room as the other examiner, the interlocutor, who manages the actual delivery of the test, test-takers were always in separate rooms to each other and the examiners for VC tests. Most test-takers sat the VC test from their home; in a few instances they were at the test centre.

In some instances, CEQ Speaking tests may be taken in a group of three. Although not a main focus of the research, two groups of three B1 for Schools test-takers were recorded taking a VC test during a small-scale initial phase of that trial. Examiners acting as the assessor in such instances reported increased difficulty distinguishing between speakers compared to the F2F mode. Groups of three would also present different challenges for participants online in terms of directing specific questions to individuals or opening up the conversational floor, which are more readily resolved in person. Because of this combination of factors, it was decided to concentrate on the paired format for the remainder of the trial as by far the most common form of delivery and exclude data from the trios, with further investigation into triadic online discussion offering rich potential for future research.

At most centres, one examiner acted as interlocutor for all tests (both F2F and VC), whilst another carried out the role of assessor. Even when examiners switched roles during a session, all test-takers had the same interlocutor and assessor for both versions of the test. This was done to minimise the effect of examiner role (assessor or interlocutor) on the scores.

The length of time between test-takers' first and second tests was kept as short as logistically possible. At most centres, this meant test-takers completed both tests either on the same day or within at most a week of each other. At two centres, the gap between tests was longer, at between two to four weeks.

The order of taking the VC or F2F test was controlled in order to minimise an order effect, i.e., to ensure that findings were not affected by everyone taking one of the test modes before the other one. The aim was for approximately 50% of test-takers to be F2F first and VC second with the test order of the remaining 50% reversed, which was generally met for all three levels.

Examiners submitted their marks electronically rather than on a paper mark sheet. Examiners were instructed to submit their marks individually and not to discuss awarded scores with their co-examiner.

How to read this issue

Overall, the five papers in this special issue present validity evidence from different lenses which supports the comparability of the VC and F2F test modes of interest here. In this opening piece of the issue, the overall research background and study specifics have been presented as the contextual and methodological information which underpins each of the five individual articles. We have chosen a certain order of presenting the individual articles in order to incrementally build the validity argument for the VC paired speaking test. The reader can make a discretionary decision on how to approach the order and choice of individual articles, depending on their own interest or circumstances. Each article should stand on its own, and the current overarching introduction can be revisited when the reader is required to remind themselves of the rationale or design of the larger research project.

References

Cooke, S (2015) *Configuring the game of speaking: Interactional competence in the IELTS Oral Proficiency Interview across two modes of response*, unpublished Master's dissertation, Lancaster University.

Craig, D A and Kim, J (2010) Anxiety and performance in video-conferenced and face-to-face oral interviews, *Multimedia-Assisted Language Learning* 13 (3), 9–32.

IBM Corp. (2017) *IBM SPSS Statistics for Windows, Version 25.0*, Armonk: IMB Corp.

Kim, J and Craig, D A (2012) Validation of a video-conferenced speaking test, *Computer-Assisted Language Learning* 25 (3), 257–275.

Lee, H, Mullooly, A, Devine, A and Galaczi, E (2023) Exploring interaction in video-call paired speaking tests: A look at scores, language, and perceptions, *Applied Linguistics*. Advance online publication.

Lee, H, Patel, M, Lynch, J and Galaczi, E (2021) *Development of the IELTS video call speaking test: Phase 4 operational research trial and overall summary of a four-phase test development cycle*, IELTS Partnership Research Papers 2021/1, IELTS Partners: British Council/Cambridge Assessment English/IDP: IELTS Australia.

Linacre, J M (2020) *Facets computer program for many-facet Rasch measurement, Version 3.83.3*, Oregon: Winsteps.com.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2021) Video-conferencing speaking tests: Do they measure the same construct as face-to-face tests?, *Assessment in Education: Principles, Policy & Practice* 28 (4), 369–388.

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press.

Wright, B and Linacre, M (1994) *Reasonable mean-square fit values*, available online: www.rasch.org/rmt/rmt83b.htm

Looking into an innovative test mode in paired speaking from the perspective of scores

Hye-won Lee, Research, Cambridge University Press & Assessment

Introduction

Scores in a test are not random numeric figures but point indicators of one's ability to manipulate what has been assessed and therefore carry larger meaning than what they appear as just simple numbers. Due to this major role they play in testing practices, qualities of assigned scores have been put at the centre of validity considerations and examined carefully along with other facets of validity. For instance, in argument-based approaches to validation, a chain of inferences is laid out around test scores and their use (Chapelle and Lee 2022). In another approach to validity evidence-gathering, scoring validity and a focus on scores is one of the dimensions in the socio-cognitive framework, which has guided many test development and validation activities (Taylor 2011).

When comparing an existing delivery mode of a test to a newer mode of what supposedly tests the same or a similar construct, scores from the test modes to be compared tend also to be analysed first above all. Various statistical methods are used for comparison to scrutinise how much the scores from the modes of your interest are in agreement, correlated, and/or at a similar level of difficulty (e.g., Bernstein, van Moere and Cheng 2010, Kiddle and Kormos 2011, Stansfield and Kenyon 1992). In recent years, video-call (VC) tests have started to be deemed as a practical alternative to in-person (or face-to-face (F2F)) assessment of spoken language (Nakatsuhara, Inoue, Berry and Galaczi 2017), and the comparability of both test modes has been investigated from multiple perspectives, including assigned scores. Reasonable comparability of scores has been found in both holistic and analytic scales (Clark and Hooshmand 1992, Craig and Kim 2010), in the classroom setting (Kim and Craig 2012), and for high-stakes tests such as IELTS (Nakatsuhara et al 2017, 2021). However, most research on VC speaking tests was conducted from individual contexts where only one examiner and one test-taker interact. There is limited research from paired/group contexts involving two or more test-takers, which the larger study described in the introductory article of this issue targets.

Under the mixed methods research design of the larger study, this article explores the feasibility of VC paired speaking tests from the perspective of test scores. The scores collected between the VC and F2F mode are statistically compared to examine the following research question:

Are there differences in test-takers' scores in the F2F and VC mode:

- As a whole?
- Per rating category?

To ensure the robustness of the scores being examined, the qualities of the examiners participating in the current study such as rater severity, consistency, and agreement were also analysed and factored into interpretation of the findings.

Methods

Data collection

As described in the introductory paper of the issue, the 19 B2 First (henceforth B2), and 29 B1 Preliminary for Schools (henceforth B1 for Schools) and A2 Key for Schools (henceforth A2 for Schools) examiners, five of which participated in both trials, rated test-taker performances against the scales used for the standard A2 for Schools, B1 for Schools, and/or B2 Speaking tests – the Global Achievement Scale (holistic) and the Assessment Scales (analytic) (the Assessment Scales for A2 Key for Schools, B1 Preliminary for Schools, and B2 First). As in the standard test, a pair of examiners, the interlocutor and the assessor, were present in each test and marked different aspects of performance. The interlocutor awarded a mark for Global Achievement, and the assessor awarded marks for three/four[1] individual criteria in the Assessment Scales: 1) Grammar and Vocabulary, 2) Discourse Management, 3) Pronunciation, and 4) Interactive Communication. Examiners were instructed to take only an assessor or interlocutor role for both test modes with the same test-takers to ensure consistency in the setting. It was however allowable for examiners to switch roles in the middle of a session. The marks were entered by the examiners independently via the online platform Totara.

Six A2 for Schools, six B1 for Schools, and seven B2 video-recorded tests from the trials were double-marked by 27 out of the 29 A2 for Schools and B1 for Schools

[1] Discourse Management is not assessed at A2 level, so there are three analytic criteria: Grammar and Vocabulary, Pronunciation, and Interactive Communication.

examiners[2] or all 19 B2 examiners to ensure the connection (linkage) across all test-takers and examiners required for Many-Facet Rasch Measurement (MFRM) analysis. One pair, either in the F2F or VC speaking test, was randomly chosen from randomly selected participating test centres and for coverage of all participating test-taker language groups. The video recordings of the selected pairs were provided in Kiteworks, a secure file sharing platform. The examiners were asked to award analytic marks to all recordings except those for which they already acted as the assessor; in all cases they were asked to award analytic marks only. To reflect the standard marking process as far as possible, they were advised to watch each recording through only once without pausing or rewinding. All the marks were collected digitally again using Totara.

Data analysis

The marks awarded in the F2F and VC speaking test were compared using both classical test theory (CTT) analysis from the lens of paired sample t-tests (SPSS 25, IBM Corp. 2017) and MFRM analysis via the FACETS 3.83.3 analysis software (Linacre 2020). Paired samples t-tests are suitable as a group-level measure of comparison of the scores, and MFRM is suitable for shedding light on features of individual rater marking behaviours such as harshness and consistency. Analyses from the two perspectives were triangulated and provided complementary insights into the findings.

Results and discussion

Classical test theory (CTT) analysis

Figures 1 to 6 present the frequency of the Global Achievement scores test-takers received in the two test delivery modes across Levels A2, B1 and B2. They are a useful overview of overall distribution of test-taker proficiency. Most of the scores cluster around Scores 4.0, 4.5, and 5.0 for A2 for Schools and B1 for Schools and around Scores 3.5, 4.0, and 4.5 for B2, reflecting the fact that the participating test-takers are those who think they are ready to take and likely to score at the higher end of the scale.

[2] Two examiners did not participate in multiple marking.

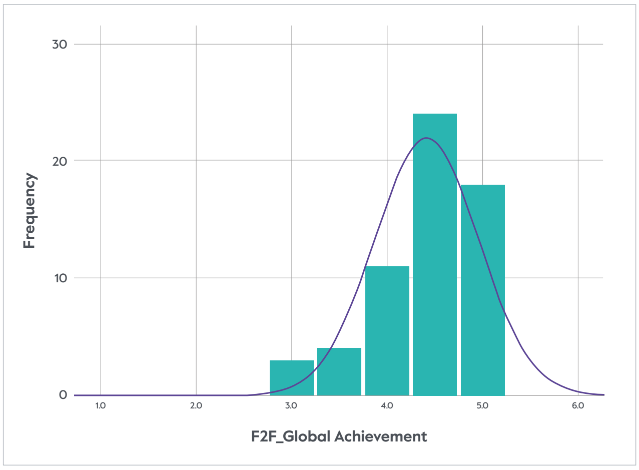


Figure 1: F2F Global Achievement scores of A2 for Schools

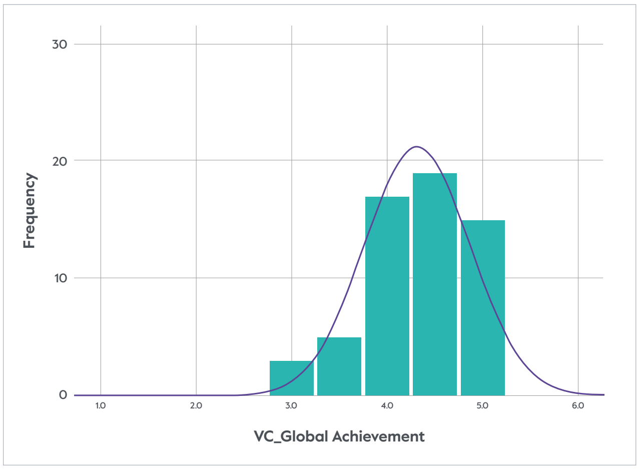


Figure 2: VC Global Achievement scores of A2 for Schools

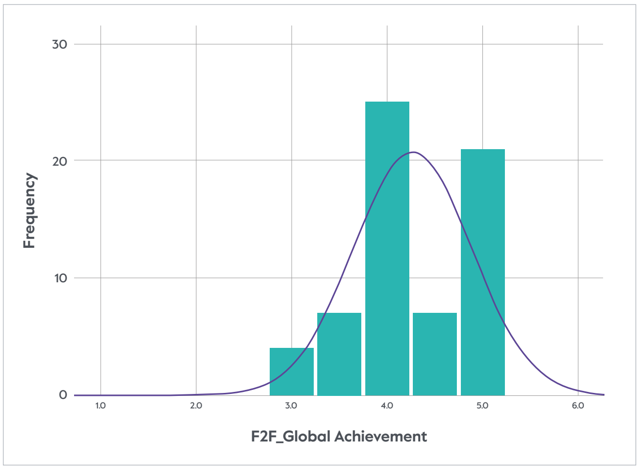


Figure 3: F2F Global Achievement scores of B1 for Schools

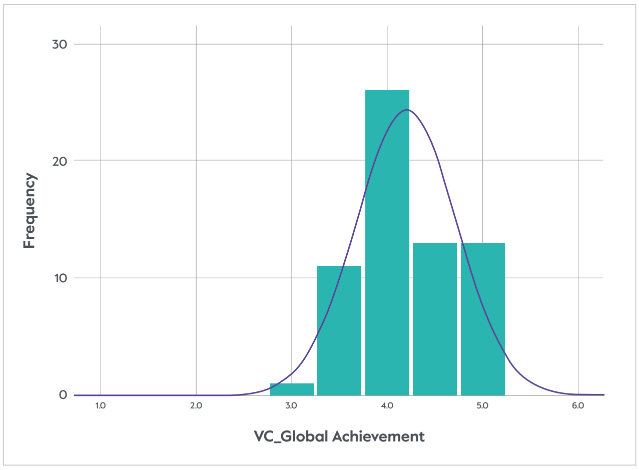


Figure 4: VC Global Achievement scores of B1 for Schools

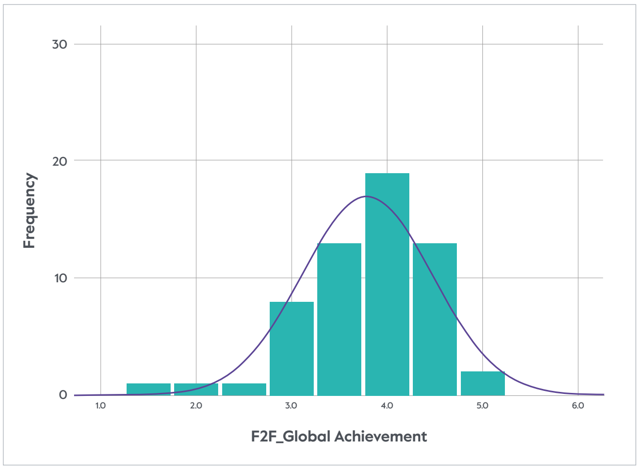


Figure 5: F2F Global Achievement scores of B2

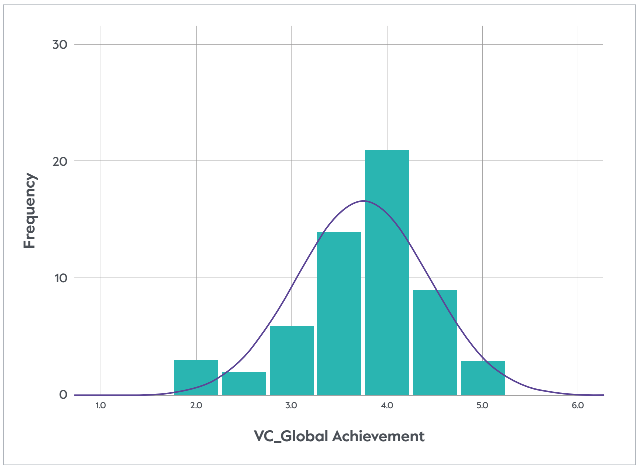


Figure 6: VC Global Achievement scores of B2

Tables 1 to 3 show descriptive statistics of test scores awarded in the two test delivery modes for A2, B1 and B2, and inferential statistics to compare the means using paired-samples t-tests.

Table 1: Descriptive statistics and paired-samples t-tests of test scores of A2 for Schools (*N*=60)

Rating category	Test mode	Mean	SD	Max	Min	Mean diff.	t	Sig. (2-tailed)
Grammar and Vocabulary	F2F	4.283	.585	5.0	2.5	-.017	-.261	.795
	VC	4.300	.652	5.0	2.5			
Pronunciation	F2F	4.467	.503	5.0	3.0	-.016	-.375	.709
	VC	4.483	.504	5.0	3.5			
Interactive Communication	F2F	4.322	.540	5.0	3.0	-.025	-.327	.745
	VC	4.347	.625	5.0	2.5			
Global Achievement	F2F	4.407	.545	5.0	3.0	.085	1.427	.159
	VC	4.322	.555	5.0	3.0			

Table 2: Descriptive statistics and paired-samples t-tests of test scores of B1 for Schools (*N*=64)

Rating category	Test mode	Mean	SD	Max	Min	Mean diff.	t	Sig. (2-tailed)
Grammar and Vocabulary	F2F	4.086	.574	5.0	3.0	.008	.163	.871
	VC	4.078	.579	5.0	3.0			
Discourse Management	F2F	4.117	.627	5.0	3.0	-.033	-.704	.484
	VC	4.150	.646	5.0	3.0			
Pronunciation	F2F	4.180	.530	5.0	3.0	-.023	-.554	.581
	VC	4.203	.582	5.0	3.0			
Interactive Communication	F2F	4.289	.502	5.0	3.0	.039	.897	.373
	VC	4.250	.577	5.0	3.0			
Global Achievement	F2F	4.266	.617	5.0	3.0	.063	1.305	.197
	VC	4.203	.525	5.0	3.0			

Table 3: Descriptive statistics and paired-samples t-tests of test scores of B2 (*N*=58)

Rating category	Test mode	Mean	SD	Max	Min	Mean diff.	t	Sig. (2-tailed)	Effect size (d)
Grammar and Vocabulary	F2F	3.414	.636	5.0	1.5	.061	1.412	.163	-
	VC	3.353	.675	5.0	1.0				
Discourse Management	F2F	3.629	.551	5.0	2.0	.017	.340	.735	-
	VC	3.612	.513	5.0	2.0				
Pronunciation	F2F	3.629	.464	5.0	2.5	.129	3.236	.002	.424
	VC	3.500	.496	5.0	2.0				
Interactive Communication	F2F	3.819	.493	5.0	2.5	.112	2.350	.022	.308
	VC	3.707	.496	5.0	2.0				
Global Achievement	F2F	3.793	.682	5.0	1.5	.043	.798	.428	-
	VC	3.750	.696	5.0	2.0				

The descriptive statistics indicate that the mean scores awarded for the four analytic and one holistic criteria in B2 were slightly higher in the F2F test condition than those in the VC test condition, whereas any systematic patterns in differences were not found in either A2 for Schools or B1 for Schools. In any case, the actual differences were from almost negligible to very small (ranging between 0.016 and 0.085 of a band for A2 for Schools, between 0.008 and 0.063 for B1 for Schools, and between 0.017 and 0.129 for B2). According to paired-samples t-tests, these minimal differences were found to be not statistically significant either in A2 for Schools or B1 for Schools, but two of the differences in B2 test scores were statistically significant: for Pronunciation (*t*(57)=3.236, *p*=0.002) and for Interactive Communication (*t*(57)=2.350, *p*=0.022). The effect sizes of these statistically significant differences were small (Cohen's *d*=0.424 and 0.308, respectively), however, indicating that the differences can be considered trivial.

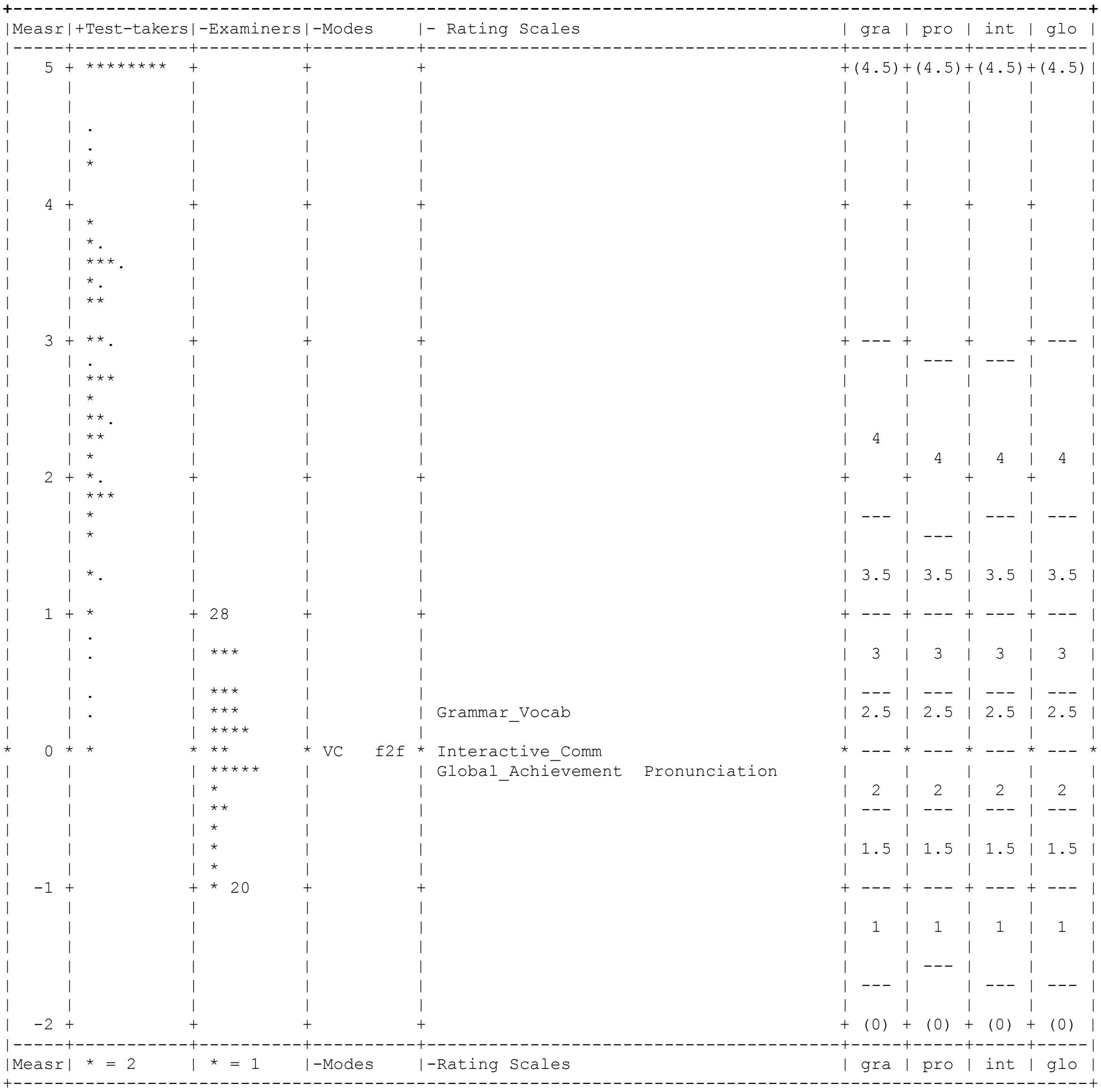
These CTT analyses are based on the assumption that any differences in rater severity have been controlled so that score differences would be mainly due to differences in test-taker performance and/or delivery mode. To complement the findings from CTT analysis, MFRM analyses that factor in rater severity were conducted.

Many-facet Rasch Measurement (MFRM) analysis

Two MFRM analyses were carried out. First, to obtain an overall picture of relative difficulty among the factors, a partial credit model analysis was conducted using four facets for score variance: test-takers, examiners, delivery modes, and rating scales.

Four-facet analysis with partial credit model

All the facets were measured by a common unit (logit) labelled as measure ('measr'), making it possible to directly compare the facets along the same scale. A visual comparison of the four facets and the elements within each facet can be found in the 'vertical rulers' (Figures 7, 8, and 9). The examiner, mode and rating scales facets are negatively scaled, placing the harsher examiners and the more difficult delivery modes and rating scales towards the top.



```
+-----------------------------------------------------------------------------------------------------------+
|Measr|+Test-takers|-Examiners|-Modes     |- Rating Scales                          | gra | pro | int | glo |
|-----+------------+----------+-----------+-----------------------------------------+-----+-----+-----+-----|
|   5 + ********   +          +           +                                         +(4.5)+(4.5)+(4.5)+(4.5)|
|     |            |          |           |                                         |     |     |     |     |
|     |            |          |           |                                         |     |     |     |     |
|     | .          |          |           |                                         |     |     |     |     |
|     | .          |          |           |                                         |     |     |     |     |
|     | *          |          |           |                                         |     |     |     |     |
|     |            |          |           |                                         |     |     |     |     |
|   4 +            +          +           +                                         +     +     +     +     |
|     | *          |          |           |                                         |     |     |     |     |
|     | *.         |          |           |                                         |     |     |     |     |
|     | ***.       |          |           |                                         |     |     |     |     |
|     | *.         |          |           |                                         |     |     |     |     |
|     | **         |          |           |                                         |     |     |     |     |
|     |            |          |           |                                         |     |     |     |     |
|   3 + **.        +          +           +                                         + --- +     +     + --- |
|     | .          |          |           |                                         |     | --- | --- |     |
|     | ***        |          |           |                                         |     |     |     |     |
|     | *          |          |           |                                         |     |     |     |     |
|     | **.        |          |           |                                         |     |     |     |     |
|     | **         |          |           |                                         |  4  |     |     |     |
|     | *          |          |           |                                         |     |  4  |  4  |  4  |
|   2 + *.         +          +           +                                         +     +     +     +     |
|     | ***        |          |           |                                         |     |     |     |     |
|     | *          |          |           |                                         | --- |     | --- | --- |
|     | *          |          |           |                                         |     | --- |     |     |
|     |            |          |           |                                         |     |     |     |     |
|     | *.         |          |           |                                         | 3.5 | 3.5 | 3.5 | 3.5 |
|     |            |          |           |                                         |     |     |     |     |
|   1 + *          + 28       +           +                                         + --- + --- + --- + --- |
|     | .          |          |           |                                         |     |     |     |     |
|     | .          | ***      |           |                                         |  3  |  3  |  3  |  3  |
|     |            |          |           |                                         |     |     |     |     |
|     | .          | ***      |           |                                         | --- | --- | --- | --- |
|     | .          | ***      |           | Grammar_Vocab                           | 2.5 | 2.5 | 2.5 | 2.5 |
|     |            | ****     |           |                                         |     |     |     |     |
*   0 * *          * **       * VC    f2f * Interactive_Comm                        * --- * --- * --- * --- *
|     |            | *****    |           | Global_Achievement  Pronunciation       |     |     |     |     |
|     |            | *        |           |                                         |  2  |  2  |  2  |  2  |
|     |            | **       |           |                                         | --- | --- | --- | --- |
|     |            | *        |           |                                         |     |     |     |     |
|     |            | *        |           |                                         | 1.5 | 1.5 | 1.5 | 1.5 |
|     |            | *        |           |                                         |     |     |     |     |
|  -1 +            + * 20     +           +                                         + --- + --- + --- + --- |
|     |            |          |           |                                         |     |     |     |     |
|     |            |          |           |                                         |  1  |  1  |  1  |  1  |
|     |            |          |           |                                         |     |     |     |     |
|     |            |          |           |                                         |     | --- |     |     |
|     |            |          |           |                                         | --- |     | --- | --- |
|     |            |          |           |                                         |     |     |     |     |
|  -2 +            +          +           +                                         + (0) + (0) + (0) + (0) |
|-----+------------+----------+-----------+-----------------------------------------+-----+-----+-----+-----|
|Measr| * = 2      | * = 1    |-Modes     |-Rating Scales                           | gra | pro | int | glo |
+-----------------------------------------------------------------------------------------------------------+
```

Figure 7: Vertical rulers of A2 for Schools (four-facet analysis with partial credit model)

```
+---------------------------------------------------------------------------------------------------------------+
|Measr|+Test-takers|-Examiners|-Modes    |-Rating Scales                          | gra | dis | pro | int | glo |
|-----+------------+----------+----------+----------------------------------------+-----+-----+-----+-----+-----|
|   5 + *********. +          +          +                                        +(4.5)+(4.5)+(4.5)+(4.5)+(4.5)|
|     |            |          |          |                                        |     |     |     |     |     |
|     |            |          |          |                                        |     |     |     |     |     |
|     |            |          |          |                                        |     |     |     |     |     |
|     |            |          |          |                                        |     |     |     |     |     |
|     | *.         |          |          |                                        |     |     |     |     |     |
|     |            |          |          |                                        |     |     |     |     |     |
|   4 + *          +          +          +                                        +     +     +     +     +     |
|     | .          |          |          |                                        |     |     |     |     |     |
|     | **.        |          |          |                                        |     |     |     |     |     |
|     | ***        |          |          |                                        |     |     |     |     |     |
|     | *          |          |          |                                        |     |     |     |     |     |
|     | *.         |          |          |                                        |     |     |     |     | --- |
|     | .          |          |          |                                        | --- |     | --- |     |     |
|   3 + .          +          +          +                                        +     + --- +     + --- +     |
|     | **         |          |          |                                        |     |     |     |     |     |
|     | *          |          |          |                                        |     |     |     |     |     |
|     | **.        |          |          |                                        |     |     |     |     |     |
|     | ****       |          |          |                                        |  4  |     |     |     |  4  |
|     | **         |          |          |                                        |     |  4  |  4  |     |     |
|     | *          |          |          |                                        |     |     |     |  4  |     |
|   2 + **         +          +          +                                        +     +     +     +     +     |
|     | **         |          |          |                                        | --- |     | --- |     | --- |
|     | **.        |          |          |                                        |     | --- |     | --- |     |
|     | **.        |          |          |                                        |     |     |     |     | 3.5 |
|     | .          |          |          |                                        | 3.5 | 3.5 | 3.5 |     |     |
|     | .          |          |          |                                        |     |     |     | 3.5 |     |
|     | *.         |          |          |                                        | --- |     | --- |     | --- |
|   1 +            + 28       +          +                                        +     + --- +     + --- +     |
|     | .          |          |          |                                        |  3  |     |  3  |     |  3  |
|     |            | **       |          |                                        |     |  3  |     |  3  |     |
|     |            | *        |          |                                        | --- |     | --- |     | --- |
|     |            | ****     |          |                                        |     | --- |     | --- |     |
|     |            | ***      |          |                                        | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 |
|     |            | **       |          | Discourse           Grammar_Vocab      |     |     |     |     |     |
*   0 *            * **       * VC   f2f * Interactive_Comm    Pronunciation      * --- * --- * --- * --- * --- *
|     |            | ******   |          |                                        |     |     |     |     |     |
|     |            | **       |          | Global_Achievement                     |  2  |  2  |  2  |  2  |  2  |
|     |            | *        |          |                                        |     | --- |     | --- |     |
|     |            | **       |          |                                        | --- |     | --- |     | --- |
|     |            | *        |          |                                        |     | 1.5 |     | 1.5 |     |
|     |            | *        |          |                                        | 1.5 |     | 1.5 |     | 1.5 |
|  -1 +            +          +          +                                        +     + --- +     + --- +     |
|     |            | 29       |          |                                        | --- |     | --- |     | --- |
|     |            |          |          |                                        |     |     |     |  1  |     |
|     |            |          |          |                                        |  1  |  1  |  1  |     |     |
|     |            |          |          |                                        |     |     |     |     |  1  |
|     |            |          |          |                                        |     | --- |     | --- |     |
|     |            |          |          |                                        | --- |     | --- |     | --- |
|  -2 +            +          +          +                                        + (0) + (0) + (0) + (0) + (0) |
|-----+------------+----------+----------+----------------------------------------+-----+-----+-----+-----+-----|
|Measr| * = 2      | * = 1    |-Modes    |-Rating Scales                          | gra | dis | pro | int | glo |
+---------------------------------------------------------------------------------------------------------------+
```

Figure 8: Vertical rulers of B1 for Schools (four-facet analysis with partial credit model)

```
+------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Test-takers|-Examiners     |-Modes|-Rating Scales                                              | gra | dis | pro | int | glo |
|-----+------------+---------------+------+------------------------------------------------------------+-----+-----+-----+-----+-----|
|   5 +            +               +      +                                                            +(4.5)+(4.5)+(4.5)+(4.5)+(4.5)|
|     |            |               |      |                                                            |     |     |     |     |     |
… [sections of output omitted]
|     |            |               |      |                                                            | --- | --- |     | --- |     |
|     |            |               |      |                                                            |     |     |     |     |     |
|   3 + *          +               +      +                                                            +     +     +     +     + --- |
|     | *          |               |      |                                                            |     |     |     |     |     |
|     | ***        |               |      |                                                            |     |     |     |     |     |
|     | *          |               |      |                                                            |     |     |  4  |     |     |
|     | *****      |               |      |                                                            |  4  |  4  |     |     |     |
|     |            |               |      |                                                            |     |     |     |  4  |     |
|     | ***        |               |      |                                                            |     |     |     |     |  4  |
|     | *          |               |      |                                                            |     |     |     |     |     |
|   2 + ****       +               +      +                                                            +     +     + --- +     +     |
|     | ******     |               |      |                                                            | --- | --- |     | --- |     |
|     | *******    |               |      |                                                            |     |     |     |     | --- |
|     | ****       |               |      |                                                            |     |     | 3.5 |     |     |
|     | ******     |               |      |                                                            | 3.5 | 3.5 |     | 3.5 |     |
|     | ***        |               |      |                                                            |     |     |     |     | 3.5 |
|     | ****       |               |      |                                                            |     |     | --- |     |     |
|     | *          |               |      |                                                            | --- | --- |     | --- |     |
|   1 + *          +               +      +                                                            +     +     +     +     + --- |
|     | *          |               |      |                                                            |  3  |  3  |  3  |  3  |     |
|     | *          |               |      |                                                            |     |     |     |     |  3  |
|     |            | 5             |      |                                                            |     |     | --- |     |     |
|     | **         | *             |      |                                                            | --- | --- |     | --- | --- |
|     |            | *  *          |      |                                                            |     |     |     |     |     |
|     | *          | *  *          |      | Grammar_Vocab                                              | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 |
|     |            | *  *  *  *  * | VC   |                                                            |     |     |     |     |     |
*   0 *            * *             *      * Discourse          Global_Achievement  Pronunciation       * --- * --- * --- * --- * --- *
|     |            | *             | f2f  |                                                            |     |     |     |     |     |
|     |            | *  *          |      | Interactive_Comm                                           |  2  |  2  |  2  |  2  |  2  |
|     |            | *             |      |                                                            |     |     |     |     |     |
|     |            | *             |      |                                                            | --- | --- |     | --- | --- |
|     |            | *             |      |                                                            |     |     | --- |     |     |
|     |            |               |      |                                                            |     |     |     |     | 1.5 |
|     |            | 14            |      |                                                            | 1.5 | 1.5 | 1.5 | 1.5 |     |
|  -1 +            +               +      +                                                            + (0) + (0) + (0) + (0) + (0) |
|-----+------------+---------------+------+------------------------------------------------------------+-----+-----+-----+-----+-----|
|Measr| * = 1      | * = 1         |-Modes|-Rating Scales                                              | gra | dis | pro | int | glo |
+------------------------------------------------------------------------------------------------------------------------------------+
```

Figure 9: Vertical rulers of B2 (four-facet analysis with partial credit model)

Going from left to right in the figures, the examiners showed different levels of harshness/leniency, as seen in the spread in logit values under 'Examiners'. Examiner 28 is most harsh for both A2 for Schools and B1 for Schools, and Examiner 20 for A2 for Schools and Examiner 29 for B1 for Schools are most lenient. In B2, Examiner 5 is most harsh, and Examiner 15 is most lenient. However, they are approximately within one logit apart from zero, which indicates that the examiners did not differ too much in their rating severity.

The next column ('Modes') in the figures shows that the VC and F2F modes were similar in difficulty, as also noted in Tables 7, 8, and 9 respectively. The 'Scales' column shows that the Assessment and Global Achievement criteria were of similar difficulty as well for all three tests.

We now turn to the measurement reports for the facet of examiners and test delivery mode (the main focus of comparison in the current study). The difficulty of the elements in each facet is shown in terms of the logit scale (Measure) and Fair Averages which indicate expected average raw scores, converted from the Measures, on the rating scale. Tables 4, 5, and 6 provide information about the examiner severity, consistency and agreement, and indicate that overall, the examiners for all three tests displayed some levels of variability, but not to a worrying extent. In terms of differences in severity, the 'Measure' column in Table 4 indicates that the difference between the most lenient and harshest examiner, Examiners 20 and 28 respectively, for A2 for Schools was 2.03 logits, which equates to less than one band (0.59 band based on Observed Average scores; 0.71 band based on Fair Average). Second, the difference between the most lenient examiner, Examiner 29, and the harshest, Examiner 28, for B1 for Schools was 2.22 (see Table 5), which also equates to less than one band (0.97 band based on Observed Average scores; 0.77 band based on Fair Average). Third, the difference between the most lenient and harshest (Examiner 14 and 5, respectively) for B2 was 1.46 logits (see Table 6), which again equates to less than one band (0.79 band based on Observed Average scores; 0.78 band based on Fair Average).

A further measure of interest in Tables 4 to 6 is the 'Infit Mean Square' (Infit MnSq): its values are commonly used as a measure of examiner consistency – or 'fit' in terms of meeting the assumptions of the Rasch model. Infit values ranging between 0.5 and 1.5 (or 1.2 as the more stringent threshold) are considered 'productive for measurement'; those below 0.5 are 'less productive for measurement, but not degrading' (Wright and Linacre 1994) and 'overfitting'; those above 1.2 are 'misfitting' and indicative of rater inconsistency. Low mean-square (MnSq) overfit values indicate that the scores awarded by each examiner can be accurately predicted from each other (intra-rater reliability) and do not display an expected level of variance; they are not indicative of problematic ratings. As can be seen, most of the Infit values reported in Tables 4 to 6 are under 0.5; and none are over the most stringent limit, 1.2, indicating that there are no misfitting raters and all raters displayed an acceptable level of consistency. The issue of overfitting ratings was inevitable in that most of the test-takers for the current study were at the higher end of the rating scale (since they were preparing to take the test) and therefore showed relatively low variability of marks. Tables 4 to 6 provide evidence that the examiners showed acceptable degrees of consistency.

In terms of examiner agreement, there was moderately high inter-rater reliability, as seen in the percentage of exact rater agreement for B2, which was 36.1%, and higher than the model expected level of 25.8% (Table 6). A similar pattern was also found in the other tests, exact 36.5% vs. expected 34.6% for A2 for Schools (Table 4) and exact 33.5% vs. expected 28.8% for B1 for Schools (Table 5). Such slight differences in agreement are normal for trained examiners and suggest that examiners rated to an acceptable degree.

Considering the findings in Tables 4 to 6, we can conclude that the examiners in this study showed acceptable reliability. These findings also provide evidence that the score findings underpinning the research question in this study are based on a robust set of scores.

Table 4: Examiner measurement report of A2 for Schools (four-facet analysis with partial credit model)

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Examiner 20	-1.04	.49	4.33	4.36	.56
Examiner 8	-.93	.24	3.99	4.35	.32
Examiner 7	-.86	.33	4.23	4.34	.47
Examiner 5	-.73	.27	3.94	4.31	.48
Examiner 25	-.57	.42	4.21	4.28	.49
Examiner 4	-.44	.25	4.05	4.26	.52
Examiner 23	-.36	.36	4.22	4.24	.42
Examiner 3	-.35	.37	4.15	4.24	.38
Examiner 14	-.21	.25	3.64	4.20	.15
Examiner 24	-.17	.34	4.15	4.19	.48
Examiner 10	-.15	.28	4.13	4.18	.68
Examiner 6	-.14	.35	4.09	4.18	.35
Examiner 22	-.10	.25	4.11	4.17	.44
Examiner 9	.05	.22	4.11	4.12	.63
Examiner 2	.06	.33	4.02	4.12	.31
Examiner 29	.11	.29	4.11	4.10	.37
Examiner 15	.13	.26	3.90	4.09	.63
Examiner 13	.20	.31	4.00	4.07	.78
Examiner 17	.21	.29	4.03	4.06	.89
Examiner 27	.23	.21	3.87	4.05	.26
Examiner 26	.24	.26	3.96	4.05	.42
Examiner 21	.30	.27	4.08	4.03	.86

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Examiner 12	.38	.24	3.97	3.99	1.19
Examiner 18	.49	.23	4.06	3.94	.85
Examiner 1	.50	.25	3.99	3.94	.46
Examiner 11	.71	.20	3.88	3.83	.53
Examiner 19	.71	.23	3.86	3.83	1.10
Examiner 16	.78	.20	3.93	3.79	.33
Examiner 28	.99	.20	3.74	3.65	.62

Model, Sample: RMSE .29 Adj (True) S.D.: .44 Separation: 1.49 Strata: 2.33 Reliability (not inter-rater): .69

Inter-rater agreement opportunities: 3412 Exact agreements: 1244.5 = 36.5% Expected: 1178.9 = 34.6%

Table 5: Examiner measurement report of B1 for Schools (four-facet analysis with partial credit model)

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Examiner 29	-1.16	.61	4.34	4.36	.81
Examiner 22	-.91	.29	4.23	4.33	.36
Examiner 8	-.76	.22	4.18	4.30	.28
Examiner 24	-.61	.26	4.18	4.27	.53
Examiner 7	-.59	.17	4.05	4.27	.22
Examiner 21	-.45	.21	4.15	4.24	.76
Examiner 5	-.23	.18	3.84	4.19	.37
Examiner 6	-.22	.18	4.00	4.18	.41
Examiner 18	-.21	.23	4.10	4.18	.64
Examiner 15	-.19	.19	4.10	4.18	.53
Examiner 20	-.19	.23	4.07	4.17	.49
Examiner 4	-.17	.19	3.90	4.17	.29
Examiner 14	-.11	.58	4.21	4.13	.28
Examiner 9	-.08	.23	3.90	4.14	.91
Examiner 2	.-03	.22	3.90	4.13	.47
Examiner 23	-.01	.18	4.07	4.12	.69
Examiner 10	.12	.22	3.75	4.07	.43
Examiner 19	.19	.20	3.78	4.04	.74
Examiner 3	.24	.17	3.75	4.01	.60
Examiner 13	.27	.16	3.84	4.01	.32
Examiner 1	.28	.12	3.69	4.01	.33

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Examiner 25	.37	.20	3.69	3.97	.33
Examiner 16	.37	.17	3.89	3.98	.22
Examiner 26	.44	.19	3.67	3.94	.37
Examiner 27	.49	.18	3.65	3.92	.32
Examiner 17	.59	.18	3.68	3.86	.42
Examiner 11	.72	.18	3.53	3.81	.40
Examiner 12	.78	.13	3.54	3.76	.24
Examiner 28	1.06	.16	3.37	3.59	.44

Model, Sample: RMSE .25 Adj (True) S.D.: .46 Separation: 1.84 Strata: 2.79 Reliability (not inter-rater): .77

Inter-rater agreement opportunities: 7191.5 Exact agreements: 2410.5 = 33.5% Expected: 2072.2 = 28.8%

Table 6: Examiner measurement report of B2 (four-facet analysis with partial credit model)

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Examiner 14	-.85	.16	3.99	4.07	.47
Examiner 15	-.58	.20	3.89	3.99	.47
Examiner 7	-.47	.19	3.92	3.93	.33
Examiner 4	-.41	.16	3.89	3.93	.44
Examiner 13	-.31	.18	3.84	3.86	.34
Examiner 19	-.28	.18	3.83	3.84	.29
Examiner 1	-.19	.16	3.74	3.79	.44
Examiner 3	-.02	.14	3.64	3.70	.33
Examiner 8	.09	.15	3.63	3.63	.30
Examiner 11	.12	.11	3.55	3.62	.34
Examiner 16	.12	.12	3.58	3.61	.30
Examiner 10	.16	.16	3.59	3.60	.29
Examiner 12	.17	.15	3.57	3.59	.35
Examiner 2	.26	.16	3.51	3.53	.31
Examiner 18	.31	.14	3.57	3.49	.35
Examiner 6	.37	.11	3.24	3.46	.31
Examiner 17	.42	.15	3.42	3.44	.38
Examiner 9	.46	.14	3.39	3.39	.32
Examiner 5	.61	.12	3.20	3.29	.51

Model, Sample: RMSE .15 Adj (True) S.D.: .36 Separation: 2.35 Strata: 3.47 Reliability (not inter-rater): .85

Inter-rater agreement opportunities: 4568.5 Exact agreements: 1648 = 36.1% Expected: 1177.7 = 25.8%

Now turning to the difference in scores between the two test modes overall, Tables 7, 8, and 9 indicate that the VC delivery mode was slightly harder than the F2F mode in all three tests (F2F: -.07 logits, VC: .07 for A2 for Schools; F2F: -.02, VC: .02 for B1 for Schools; F2F: -.08, VC: .08 for B2). However, the raw score difference was minimal based on Fair Average scores: 4.16 (F2F) and 4.11 (VC) for A2 for Schools, 4.12 (F2F) and 4.11 (VC) for B1 for Schools, and 3.74 (F2F) and 3.64 (VC) for B2. From the perspective of Observed Average scores, the raw score difference for A2 for Schools was slightly larger, 4.17 (F2F) and 3.90 (VC) but not to a worrying degree. The difference for the other tests was still minimal: 3.93 (F2F) and 3.86 (VC) for B1 for Schools and 3.67 (F2F) and 3.57 (VC) for B2.

Table 7: Test delivery measurement report of A2 for Schools (four-facet analysis with partial credit model)

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
F2F	-.07	.09	4.17	4.16	.61
VC	.07	.06	3.90	4.11	.54

Table 8: Test delivery measurement report of B1 for Schools (four-facet analysis with partial credit model)

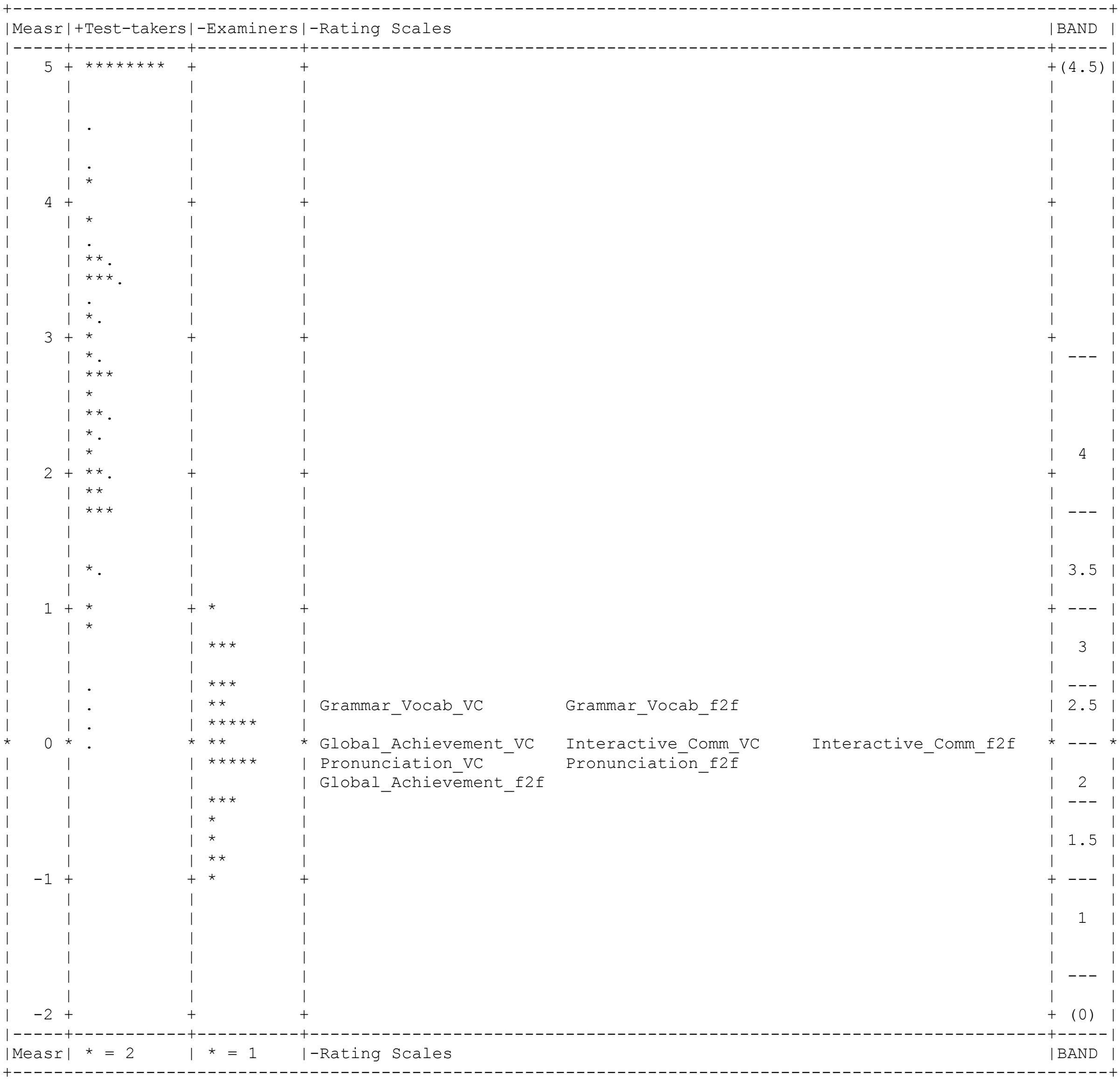
	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
F2F	-.02	.05	3.93	4.12	.41
VC	.02	.05	3.86	4.11	.43

Table 9: Test delivery measurement report of B2 (four-facet analysis with partial credit model)

	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
F2F	-.08	.05	3.67	3.74	.31
VC	.08	.05	3.57	3.64	.39

Three-facet analysis with rating scale model

Additionally to the four-facet analysis, three-facet analyses with the facets of test-takers, examiners and rating scales were conducted to investigate the performance of each rating scale in each mode separately. This led to a total of eight elements for A2 for Schools and 10 elements for B1 for Schools and B2 in the rating scale facet: F2F Grammar and Vocabulary, VC Grammar and Vocabulary, F2F Discourse Management (B1 for Schools and B2 only), VC Discourse Management (B1 for Schools and B2 only), F2F Pronunciation, VC Pronunciation, F2F Interactive Communication, VC Interactive Communication, F2F Global Achievement, and VC Global Achievement. The results of the three-facet analyses are visually presented in Figure 10 for A2 for Schools, Figure 11 for B1 for Schools, and Figure 12 for B2.

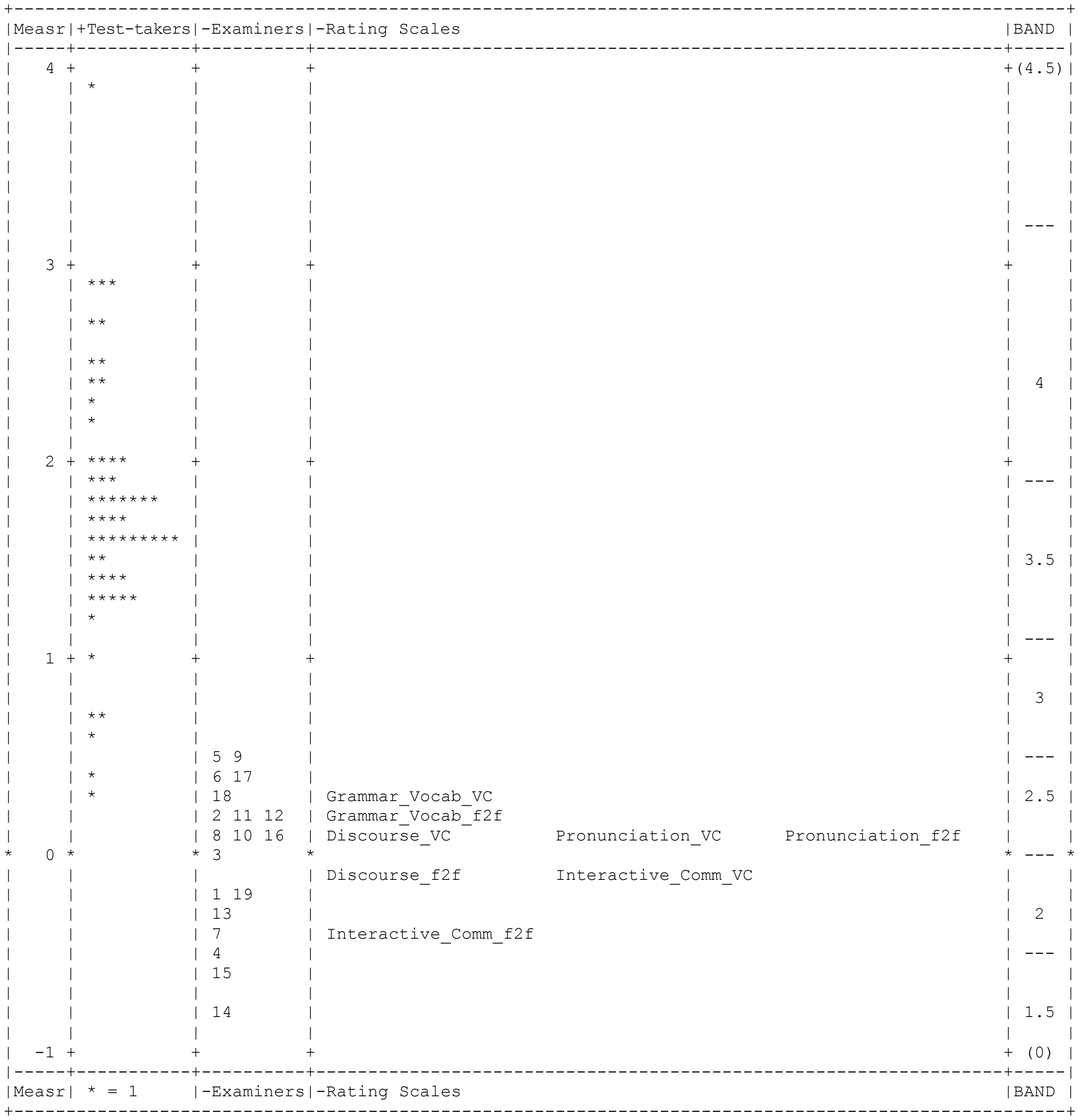


```
+--------------------------------------------------------------------------------------------------------------+
|Measr|+Test-takers|-Examiners|-Rating Scales                                                            |BAND |
|-----+------------+----------+--------------------------------------------------------------------------+-----|
|   5 + ********   +          +                                                                          +(4.5)|
|     |            |          |                                                                          |     |
|     |            |          |                                                                          |     |
|     | .          |          |                                                                          |     |
|     |            |          |                                                                          |     |
|     | .          |          |                                                                          |     |
|     | *          |          |                                                                          |     |
|   4 +            +          +                                                                          +     |
|     | *          |          |                                                                          |     |
|     | .          |          |                                                                          |     |
|     | **.        |          |                                                                          |     |
|     | ***.       |          |                                                                          |     |
|     | .          |          |                                                                          |     |
|     | *.         |          |                                                                          |     |
|   3 + *          +          +                                                                          +     |
|     | *.         |          |                                                                          | --- |
|     | ***        |          |                                                                          |     |
|     | *          |          |                                                                          |     |
|     | **.        |          |                                                                          |     |
|     | *.         |          |                                                                          |     |
|     | *          |          |                                                                          |  4  |
|   2 + **.        +          +                                                                          +     |
|     | **         |          |                                                                          |     |
|     | ***        |          |                                                                          | --- |
|     |            |          |                                                                          |     |
|     |            |          |                                                                          |     |
|     | *.         |          |                                                                          | 3.5 |
|     |            |          |                                                                          |     |
|   1 + *          + *        +                                                                          + --- |
|     | *          |          |                                                                          |     |
|     |            | ***      |                                                                          |  3  |
|     |            |          |                                                                          |     |
|     | .          | ***      |                                                                          | --- |
|     | .          | **       | Grammar_Vocab_VC         Grammar_Vocab_f2f                               | 2.5 |
|     | .          | *****    |                                                                          |     |
*   0 * .          * **       * Global_Achievement_VC    Interactive_Comm_VC      Interactive_Comm_f2f   * --- *
|     |            | *****    | Pronunciation_VC         Pronunciation_f2f                               |     |
|     |            |          | Global_Achievement_f2f                                                   |  2  |
|     |            | ***      |                                                                          | --- |
|     |            | *        |                                                                          |     |
|     |            | *        |                                                                          | 1.5 |
|     |            | **       |                                                                          |     |
|  -1 +            + *        +                                                                          + --- |
|     |            |          |                                                                          |     |
|     |            |          |                                                                          |  1  |
|     |            |          |                                                                          |     |
|     |            |          |                                                                          |     |
|     |            |          |                                                                          | --- |
|     |            |          |                                                                          |     |
|  -2 +            +          +                                                                          + (0) |
|-----+------------+----------+--------------------------------------------------------------------------+-----|
|Measr| * = 2      | * = 1    |-Rating Scales                                                            |BAND |
+--------------------------------------------------------------------------------------------------------------+
```

Figure 10: Vertical rulers of A2 for Schools (three-facet analysis with rating scale model)

```
+-----------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Test-takers|-Examiners|-Rating Scales                                                                                                         |BAND |
|-----+------------+----------+-----------------------------------------------------------------------------------------------------------------------+-----|
|   5 + *********. +          +                                                                                                                       +(4.5)|
|     |            |          |                                                                                                                       |     |
|     |            |          |                                                                                                                       |     |
|     |            |          |                                                                                                                       |     |
|     |            |          |                                                                                                                       |     |
|     | .          |          |                                                                                                                       |     |
|     | *          |          |                                                                                                                       |     |
|   4 + .          +          +                                                                                                                       +     |
|     | .          |          |                                                                                                                       |     |
|     | ***        |          |                                                                                                                       |     |
|     |            |          |                                                                                                                       |     |
|     | ***        |          |                                                                                                                       |     |
|     | **         |          |                                                                                                                       |     |
|     | .          |          |                                                                                                                       |     |
|   3 + *          +          +                                                                                                                       + --- |
|     | .          |          |                                                                                                                       |     |
|     | **.        |          |                                                                                                                       |     |
|     | .          |          |                                                                                                                       |     |
|     | ****.      |          |                                                                                                                       |     |
|     | **.        |          |                                                                                                                       |  4  |
|     | *          |          |                                                                                                                       |     |
|   2 + **.        +          +                                                                                                                       +     |
|     | *.         |          |                                                                                                                       |     |
|     | **         |          |                                                                                                                       | --- |
|     | ***        |          |                                                                                                                       |     |
|     | *.         |          |                                                                                                                       | 3.5 |
|     | .          |          |                                                                                                                       |     |
|     | *          |          |                                                                                                                       |     |
|   1 + .          + *        +                                                                                                                       + --- |
|     | .          |          |                                                                                                                       |     |
|     |            | **       |                                                                                                                       |  3  |
|     |            | *        |                                                                                                                       |     |
|     |            | ****     |                                                                                                                       | --- |
|     |            | ***      | Grammar_Vocab_VC                                                                                                      | 2.5 |
|     |            | **       | Grammar_Vocab_f2f                                                                                                     |     |
*   0 *            * ***      * Discourse_VC            Discourse_f2f           Global_Achievement_f2f  Pronunciation_VC        Pronunciation_f2f     * --- *
|     |            | *****    | Global_Achievement_VC   Interactive_Comm_VC     Interactive_Comm_f2f                                                  |     |
|     |            | **       |                                                                                                                       |  2  |
|     |            | *        |                                                                                                                       | --- |
|     |            | **       |                                                                                                                       |     |
|     |            | *        |                                                                                                                       | 1.5 |
|     |            | *        |                                                                                                                       |     |
|  -1 +            +          +                                                                                                                       + --- |
|     |            | *        |                                                                                                                       |     |
|     |            |          |                                                                                                                       |     |
|     |            |          |                                                                                                                       |  1  |
|     |            |          |                                                                                                                       |     |
|     |            |          |                                                                                                                       | --- |
|     |            |          |                                                                                                                       |     |
|  -2 +            +          +                                                                                                                       + (0) |
|-----+------------+----------+-----------------------------------------------------------------------------------------------------------------------+-----|
|Measr| * = 2      | * = 1    |-Rating Scales                                                                                                         |BAND |
+-----------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Figure 11: Vertical rulers of B1 for Schools (three-facet analysis with rating scale model)



```
+----------------------------------------------------------------------------------------------------+
|Measr|+Test-takers|-Examiners|-Rating Scales                                                  |BAND |
|-----+-----------+----------+-----------------------------------------------------------------+-----|
|   4 +           +          +                                                                 +(4.5)|
|     | *         |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 | --- |
|     |           |          |                                                                 |     |
|   3 +           +          +                                                                 +     |
|     | ***       |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     | **        |          |                                                                 |     |
|     |           |          |                                                                 |     |
|     | **        |          |                                                                 |     |
|     | **        |          |                                                                 |  4  |
|     | *         |          |                                                                 |     |
|     | *         |          |                                                                 |     |
|     |           |          |                                                                 |     |
|   2 + ****      +          +                                                                 +     |
|     | ***       |          |                                                                 | --- |
|     | *******   |          |                                                                 |     |
|     | ****      |          |                                                                 |     |
|     | ********* |          |                                                                 |     |
|     | **        |          |                                                                 | 3.5 |
|     | ****      |          |                                                                 |     |
|     | *****     |          |                                                                 |     |
|     | *         |          |                                                                 |     |
|     |           |          |                                                                 | --- |
|   1 + *         +          +                                                                 +     |
|     |           |          |                                                                 |     |
|     |           |          |                                                                 |  3  |
|     | **        |          |                                                                 |     |
|     | *         |          |                                                                 |     |
|     |           | 5 9      |                                                                 | --- |
|     | *         | 6 17     |                                                                 |     |
|     | *         | 18       | Grammar_Vocab_VC                                                | 2.5 |
|     |           | 2 11 12  | Grammar_Vocab_f2f                                               |     |
|     |           | 8 10 16  | Discourse_VC         Pronunciation_VC      Pronunciation_f2f    |     |
*   0 *           * 3        *                                                                 * --- *
|     |           |          | Discourse_f2f        Interactive_Comm_VC                        |     |
|     |           | 1 19     |                                                                 |     |
|     |           | 13       |                                                                 |  2  |
|     |           | 7        | Interactive_Comm_f2f                                            |     |
|     |           | 4        |                                                                 | --- |
|     |           | 15       |                                                                 |     |
|     |           |          |                                                                 |     |
|     |           | 14       |                                                                 | 1.5 |
|     |           |          |                                                                 |     |
|  -1 +           +          +                                                                 + (0) |
|-----+-----------+----------+-----------------------------------------------------------------+-----|
|Measr| * = 1     |-Examiners|-Rating Scales                                                  |BAND |
+----------------------------------------------------------------------------------------------------+
```

Figure 12: Vertical rulers of B2 (three-facet analysis with rating scale model)

As in the four-facet analysis, the examiners did not vary much in terms of their rating severity. The analytic criteria and holistic score (Global Achievement) in the VC mode were similar to or slightly more difficult than those in the F2F mode; all are centred around zero in logit and do not indicate major differences in the difficulty levels across all the rating scales.

Tables 10, 11 and 12 present measurement reports for the rating scales in each mode for all three tests. In terms of rater consistency, as in the four-facet analysis, most of the infit values were under 0.5, but none were over 1.5 or the more stringent 1.2, indicating no misfitting items in any facet. As visually identified in Figures 10, 11, and 12, all individual criteria in the Assessment Scales and the Global Achievement scale were centered around zero, and the raw score differences in terms of the Fair Average scores were very small. The largest differences in the Fair Average scores for all the tests are minor: a difference of 0.09 of a band for Global Achievement in A2 for Schools (4.22/F2F and 4.13/VC), a difference of 0.04 of a band for Grammar and Vocabulary in B1 for Schools (4.05/F2F and 4.01/VC), and a difference of 0.17 of a band for Interactive Communication in B2 (3.91/F2F and 3.74/VC). From the perspective of the Observed Average scores, the largest difference, 0.31, is found in the Grammar and Vocabulary score for A2 for Schools (4.12/F2F and 3.81/VC), but all the other differences are minor, the largest being -0.15 of a band for Global Achievement in B1 for Schools and 0.16 of a band for Interactive Communication for B2 as found in the Fair Average score difference.

Table 10: Rating scale measurement report of A2 for Schools (three-facet analysis)

Rating category	Test mode	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Grammar and Vocabulary	F2F	.24	.14	4.12	4.05	.53
	VC	.31	.10	3.81	4.02	.49
Pronunciation	F2F	-.13	.17	4.21	4.18	.68
	VC	-.15	.12	3.99	4.18	.55
Interactive Communication	F2F	-.05	.16	4.17	4.15	.57
	VC	.07	.11	3.90	4.11	.64
Global Achievement	F2F	-.29	.29	4.17	4.22	.74
	VC	.01	.19	3.98	4.13	.30

Table 11: Rating scale measurement report of B1 for Schools (three-facet analysis)

Rating category	Test mode	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Grammar and Vocabulary	F2F	.15	.10	3.88	4.05	.39
	VC	.24	.09	3.78	4.01	.40
Discourse Management	F2F	.06	.11	3.89	4.09	.35
	VC	.06	.10	3.83	4.08	.38
Pronunciation	F2F	-.01	.11	3.94	4.11	.41
	VC	.03	.10	3.86	4.10	.46
Interactive Communication	F2F	-.21	.12	4.00	4.17	.50
	VC	-.10	.10	3.90	4.14	.43
Global Achievement	F2F	-.07	.22	3.91	4.13	.38
	VC	-.15	.20	4.06	4.15	.40

Table 12: Descriptive statistics and paired-samples t-tests of test scores of B2 (*N*=58)

Rating category	Test mode	Measure	Model S.E.	Observed Average	Fair (M) Average	Infit MnSq
Grammar and Vocabulary	F2F	.20	.10	3.52	3.58	.30
	VC	.32	.09	3.44	3.50	.36
Discourse Management	F2F	-.07	.10	3.68	3.74	.25
	VC	.12	.09	3.56	3.63	.38
Pronunciation	F2F	.10	.10	3.59	3.64	.29
	VC	.12	.09	3.56	3.63	.35
Interactive Communication	F2F	-.39	.11	3.85	3.91	.27
	VC	-.07	.10	3.69	3.74	.46
Global Achievement	F2F	-.19	.19	3.75	3.81	.52
	VC	-.15	.18	3.73	3.79	.61

Concluding remarks

To summarise, considering the findings from the CTT, four-facet and three-facet analysis together, the score differences between the F2F and the VC mode are almost negligible, indicating that the two test delivery modes can be considered comparable in their difficulty level and providing strong evidence to support the comparability of the two delivery modes with regard to test scores. To further strengthen the comparability argument, it is indispensable to examine from other angles such as examiner behaviour and linguistic features elicited in test-taker language, which are dealt with in the other articles of this special issue.

References

Bernstein, J, van Moere, A and Cheng, J (2010) Validating automated speaking tests, *Language Testing* 27 (3), 355–377.

Chapelle, C A and Lee, H (2022) Validation of second language spoken assessments, in Haug, T, Mann, W and Knoch, U (Eds) *The Handbook of Language Assessment Across Modalities*, Oxford: Oxford University Press, 273–284.

Clark, J D and Hooshmand, D (1992) "Screen-to-screen" testing: An exploratory study of oral proficiency interviewing using video conferencing, *System* 20 (3), 293–304.

Craig, D A and Kim, J (2010) Anxiety and performance in video-conferenced and face-to-face oral interviews, *Multimedia-Assisted Language Learning* 13 (3), 9–32.

IBM Corp. (2017) *IBM SPSS Statistics for Windows, Version 25.0*, Armonk: IMB Corp.

Kiddle, T and Kormos, J (2011) The effect of mode of response on a semidirect test of oral proficiency, *Language Assessment Quarterly* 8 (4), 342–360.

Kim, J and Craig, D A (2012) Validation of a video-conferenced speaking test, *Computer-Assisted Language Learning*, 25 (3), 257–275.

Linacre, J M (2020) *Facets computer program for many-facet Rasch measurement, Version 3.83.3*, Oregon: Winsteps.com.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2021) Video-conferencing speaking tests: Do they measure the same construct as face-to-face tests?, *Assessment in Education: Principles, Policy & Practice* 28 (4), 369–388.

Stansfield, C and Kenyon, D (1992) Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview, *System* 20 (3), 347–364.

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press.

Wright, B and Linacre, M (1994) *Reasonable mean-square fit values*, available online: www.rasch.org/rmt/rmt83b.htm

Investigating language functions elicited by in-person and online paired speaking tests

Andrew Mullooly, Capabilities, Cambridge University Press & Assessment
Nicholas Glasson, Research, Cambridge University Press & Assessment

Introduction

Contemporary research into speaking test comparability and broader validity arguments are rarely based on score investigations or reported stakeholder perceptions alone. Since van Lier's exhortation to become 'serious students of natural conversation' (1989:506), explorations of what constitutes 'validity' have increasingly taken account of the language elicited from test-takers. For example, Weir's (2005) delineating of context validity includes functional language as a key component. Arguments have been made for broadening the breadth of insights in this regard (O'Loughlin 2001, Zhou 2015) to go 'beyond scores' (Nakatsuhara, Inoue, Berry and Galaczi 2021:371). As Nakatsuhara et al point out, scores are but one window via which evidence is obtained. Even where scores or perceptions are comparable, it is important to ascertain whether differences exist in the language one test elicits versus another.

Relatively few studies have compared the language elicited by in-person (or face-to-face (F2F)) and video-call (VC) modes. However, what evidence there is suggests there are aspects of the VC mode – such as lags in audio-visual transmission (or 'latency'; see Seuren, Wherton, Greenhalgh and Shaw 2021) or limited visual access – that are implicated in turn-taking. Cooke (2015), adopting a conversation analytic approach, observed less smooth speaker transitions, increased use of clarification requests and reformulations in a VC version of IELTS Speaking when compared to its F2F counterpart. Similarly, Jin and Zhang (2016) found in their discourse analytic study of online paired tests that a lack of visual access and fewer audible cues led to a greater incidence of overlapped turns between speakers, which in turn impacted score reliability. Indeed, most VC studies have highlighted how mode is influential in both the nature of turn initiation/closure and the frequency of repair operations such as clarification requests and requesting repetition (Jin and Zhang 2016, Nakatsuhara et al 2017, Ockey, Gu and Keehner 2017, Zhang and Jin 2021). These potential differences highlight the importance of ensuring different delivery modes are comparable rather than identical. This is particularly important in the context of wider operationalising of a test delivery mode so that aspects like examiner training take adequate account of any differences.

While extremely valuable to validation activities, conversation or discourse analytic approaches of the kind described above are very labour-intensive. In conversation analysis particularly there is often a reluctance to quantify findings (Schegloff 1993 – although see Stivers 2015) which in turn limits the generalisability of findings, something which large-scale test research rightly requires. A way 'around' this conflict of micro versus macro insights is found in the use of checklists or other coding frames that, while derived from an empirical base of interactional features, provide a practical way of quantifying language output.

O'Sullivan, Weir and Saville (2002) detail the initial development and use of the observation checklist used in the present study. In keeping with their original aspiration of the checklist being operationalised *mutatis mutandis* for a broader range of speaking tests, it has been adopted for a number of speaking test validation initiatives with appropriate modifications. It has been adapted for the examiner-candidate format IELTS test (Brooks 2003), used in validation studies of non-Cambridge exams (Inoue and Nakatsuhara 2021) and, importantly for the current study, utilised for studies of VC tests (Nakatsuhara et al 2016, 2017). The present study set out to identify any differences in language functions elicited in the test-taker samples across the VC and F2F modes of delivery for A2 Key for Schools, B1 Preliminary for Schools and B2 First (henceforth A2, B1 and B2).

Methodology

Given that previous research has mostly explored interlocutor-led interactions and monological tasks in VC assessment settings (notably IELTS studies), the focus of this research strand was on coding the interactive parts of the tests, where candidates engage in peer-peer discussion. For A2, this is the discussion task (Part 2, Phase 1) and examiner-led follow-up task (Part 2, Phase 2). For B1 and B2, this is the discussion task (Part 3) and examiner-led follow-up task (Part 4).

Data

For the initial B2 trial, language function analysis of candidate speech was carried out on 56 recordings (28 tests for each mode with the same test-taker pairs) to see whether VC and F2F delivery modes elicited comparable language functions from test-takers. For the later A2 and B1 for Schools trials, the sample consisted of 16 test instances (eight F2F and eight VC tests), with the same test-taker pairings and examiners for both.

Procedure

Following a familiarisation session, both transcripts from a single pair of B2 test-takers were coded individually by two researchers using the same modified version of O'Sullivan et al's (2002) observation checklist. The researchers then discussed any differences in how they had coded and agreed a standardised approach. The checklist was reviewed to consider whether this initial coding suggested any additional modifications, particularly in light of the paired format of the test.

As a result of this review, new codes were added, primarily to help distinguish between utterances that would otherwise have been captured together under existing broader categories. For example, Agreeing (prompted) (AGRP) and Disagreeing (prompted) (DISP) were included alongside the existing Agreeing (AGR) and Disagreeing (DIS) categories, to show where a test-taker's expression of agreement or disagreement with their partner was directly elicited by an examiner's scripted follow-up prompt.

Similarly, to allow for more explicit defining of turn allocation types, Nominating self (NOMS) and Nominating other (NOMO) were added to clarify the nature of nomination. Another new category, Asking for repetition (AREP) was introduced to differentiate this from Asking for clarification (CLAR). Such requests could be directed either at a test partner or towards the examiner. Adjusting equipment (ADJ), Reporting technical problem (REPP) and Solving technical problem (SOL) were also added to account for technical issues in the VC mode.

After discussion on the coding of the two initial transcripts, both researchers individually recoded these and each separately coded the same four additional transcripts, adopting a similar approach to recent test discourse studies comparing test modes (see Nakatsuhara et al 2017). The resulting double-coding reliability demonstrated an overall percentage of 96.55% agreement. This was considered high enough for the remaining transcripts to be single-coded.

For the later A2 and B1 study, the same approach of familiarisation followed by double and then single coding was taken. In this instance, the subsequent language function analysis was carried out on a subset of 32 recordings (16 test instances, eight F2F and eight VC tests for the same test-taker pairs per level).

Discussion was informed via coder-agreement metrics (using NVivo's data analysis tools), allowing the researchers to focus review deliberations on areas where they had shown less agreement, thereby enhancing levels of coder agreement in a more targeted manner. As with the B2 study, similarly high levels of agreement were reached before proceeding to single coding.

Two further new codes were added during the A2 and B1 studies to cover stretches of language that were not codable as specific functions. Unintelligible (UNI) was used to indicate instances of unintelligible speech, particularly where the main idea of an utterance was not clear, and Breakdown (BRK) to indicate instances where communication broke down, specifically, where a candidate had to resort to their first language in order to continue. These two codes were partly a reflection of the lower proficiency level. The use of UNI also related to the audio quality of certain recordings. For ease of comparison, where present in the sample, UNI and BRK are included in all charts showing distribution of linguistic functions at A2 and B1 level. Given these two codes do not relate to language functions, they are not included in the consideration of significant statistical differences. A full list of the codes used across the different trials is provided in the Appendix.

While checklists provide a convenient, adaptable tool for research into larger speaking test data sets, Green (2012) notes that checklists cannot capture features of talk-in-interaction like turn length, and nor can 'speech acts … be realised through different exponents at different levels' (Green 2012:37). For this reason, a subset of the transcribed data was analysed in more detail (Glasson, Mullooly and Galaczi 2022).

For the B2 sample only, as well as candidate speech, examiner speech was also looked at for Part 4 of the test, where there is freedom for examiners to choose which questions to address to individual candidates or both candidates. As this was not a main focus of the study, it was only single coded.

Analysis

For each test part focused on in this study an initial descriptive picture of language function use was drawn from the code data. These outputs (shown here as bar graphs) illustrate the percentage of candidates using the interactional feature (at least once) by mode of delivery. While useful for getting a general impression of the relative prominence of features by mode, these do not tell us about the more specific mean differences of feature use by candidates across the two conditions (F2F and VC) nor whether these might be meaningful differences. The data in this case were non-normally distributed so a Wilcoxon's Signed Rank test was used to compare the mean frequency of each language function in the two modes (SPSS 25, IBM Corp. 2017). This approach allowed for the inclusion of mean comparisons alongside significance and effect size measures to get a better understanding of the linguistic output of candidates.

Where any potentially significant differences were found, a correction for multiple comparisons was made. It should be noted that employing such a correction increases the risk of a Type II error. That is, concluding that there is no effect in the sampled population when there is in reality (see Dancey and Reidy 2011). Additionally, correcting for multiple comparisons is less desirable when false negatives could be more costly (see McDonald (2014) for a discussion of these issues). As this study was exploratory, and there would be implications for ignoring potential differences in language use between test modes, the differences that were detected prior to correction will still be discussed. Results and discussions are primarily focused on features which indicated a statistically significant difference (prior to correction).

Results

Results are reported in two sections. The first is focused on the peer-peer, paired discussion and decision-making tasks of A2, B1 and B2 (i.e., where the examiner withdraws from the interaction). The second is focused on an analysis of the follow-up examiner-led sections of these speaking tests (i.e., where the examiner provides a series of questions related to the earlier discussion topics as prompts for candidate responses).

First, an overview is provided of the distribution of linguistic functions across modes, then inferential statistics (mean comparisons) are provided for instances where differences were potentially significant. This is then followed by transcriptions of relevant linguistic functions as further exemplification of use.

Paired tasks

A2 Part 2 Phase 1

Figure 1 shows the percentage of A2 test-takers using functions at least once in Part 2 Phase 1 of the test. For both modes, Part 2 Phase 1 was the part of the test that elicited the widest range of functions – this is perhaps unsurprising since is it the phase of the test where test-takers are required to take control of the interaction. The most commonly used linguistic functions of Justifying (JUST), Expressing preference (PREF), Expressing an opinion (EOP) and Changing topic (CHAN) were all present for similar percentages of test-takers in both modes.

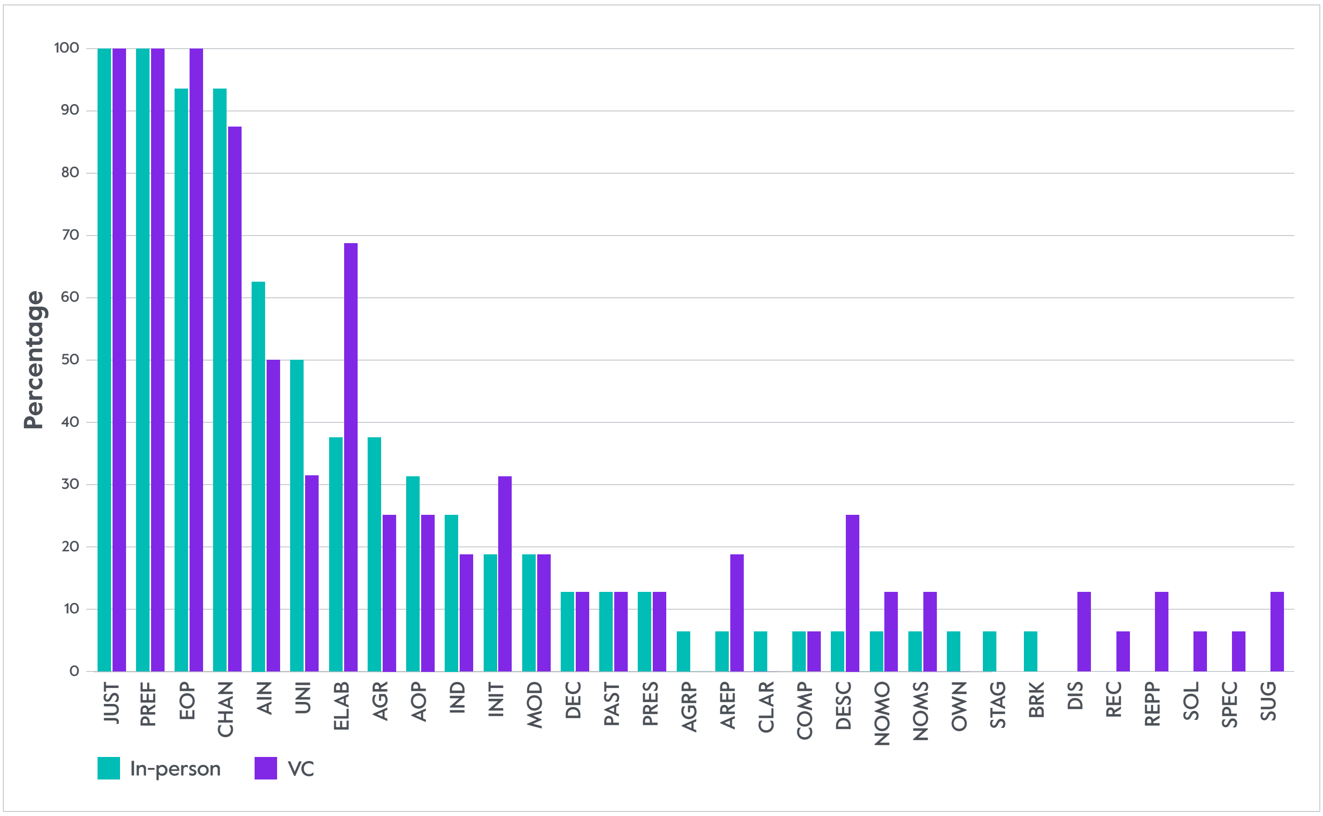


Figure 1: Distribution of linguistic functions in A2 Part 2 Phase 1 task (both modes)

While the percentages shown in Figure 1 highlighted some differences in the percentages of candidates using specific functions, for example Elaborating (ELAB), at least once, the mean comparisons indicated there were no significant statistical differences in linguistic output between the two modes for A2 Part 2 Phase 1. Put simply, while overall more candidates elaborated more in the VC condition, further analysis suggests that this was not meaningfully different (note: samples were small in this case, so percentages could be seen to exaggerate differences).

B1 Part 3

Figure 2 shows the percentage of B1 test-takers using functions at least once in Part 3 of the test. As with Part 2 Phase 1 of the A2 test, this is the part of the test that requires test-takers to maintain the discussion between themselves and as such had the widest range of functions in both modes, with a similar focus on expressing and justifying opinions, alongside the interactional functions of Agreeing (AGR) and Asking for opinion (AOP).

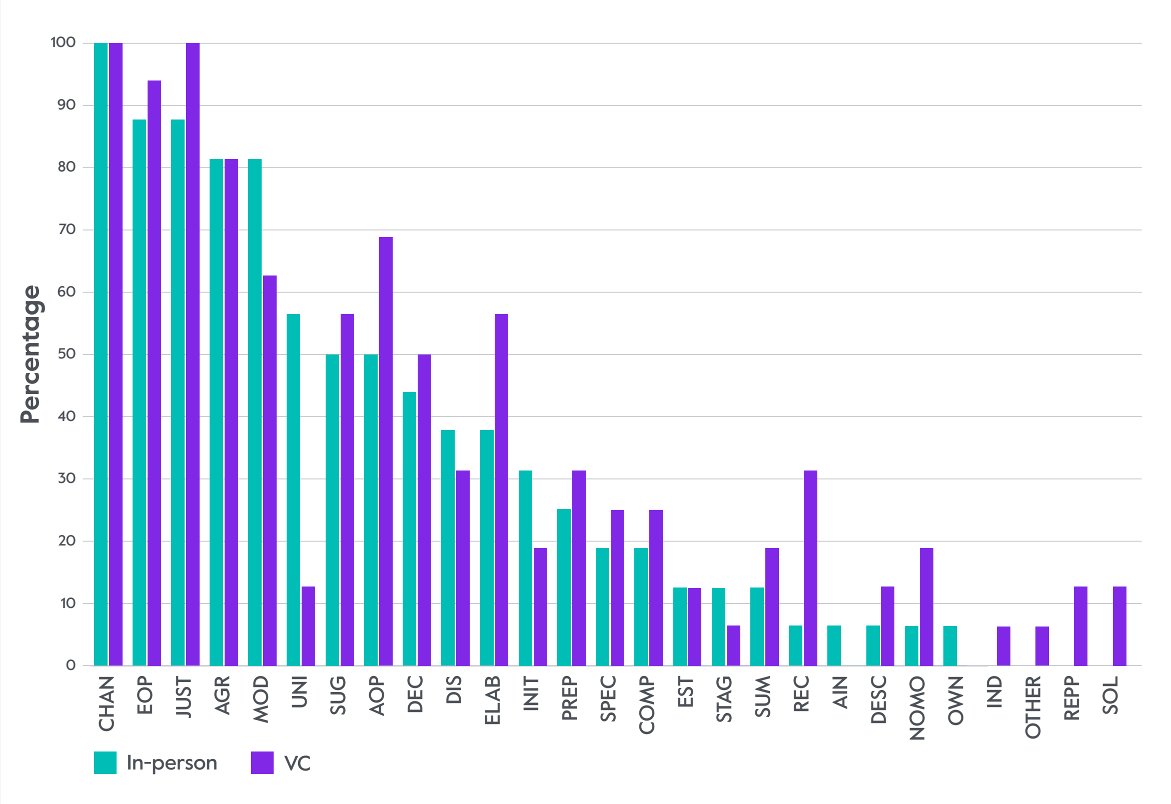


Figure 2: Distribution of linguistic functions in B1 Part 3 task (both modes)

At this higher level, there was increased evidence of test-takers modifying their partner's utterances across both modes but particularly in the F2F mode of the test (81.25%, or 13 out of 16 test-takers, compared to 62.5%, or 10 test-takers in the VC mode). The definition of Modifying (MOD) covered modifying, commenting on or adding to the arguments or comments made by their test partner. The following examples show how test-taker C3 modified opinions expressed by C4 in both the VC and F2F Part 3 of the test.

Extract B1.1: Centre MX1 – F2F mode

1	C4:	OK (.) so I don't disagree with you (.)
2		but I can also say that washing a car is also an
3		activity that's not every day you can do (.) so
4		it can be quite funny or not a boring activity to
5		do while he is not doing anything.
6→	C3:	I do not agree because washing a car is like very
7		boring and give him no fun if you don't have people to
8		enjoy (.) to wash the car (.) would you help someone else
9		that wash the car with you (.) it will be funny but alone
10		it would be like boring.

Extract B1.2: Centre MX1 – VC mode

1	C4:	... so maybe what about the earrings (.) it's like not
2		a thing you use like every day (.) but it's a thing you
3		can use for special days (.) for maybe a party (.) and
4		you can show to others where you came from or something you
5		take from another place
6→	C3:	OK (.) you got a point because it is (.) yeah as you
7		mentioned (.) it is not commonly to use it but it is like
8		safe (.) and I think that it is a very good option.

The important point in both modes was the engagement in another speaker's utterance – be it to disagree or agree with a section of prior talk – and how this shows some evidence of 'substantive recipiency' (Waring 2002) i.e., displays of understanding, across both modes.

There were no significant differences in linguistic functions in B1 Part 3. It is worth noting that Unintelligible (UNI) speech was over seven times as frequent for the F2F mode than in the VC mode of the test. However, this was felt to largely reflect the challenge in transcribing the recordings and was due to a combination of the use of face masks by participants and lack of specialist microphones in the test room. The fact that unintelligible speech was not accompanied by increased requests for repetition indicates that the test-takers themselves either understood each other or opted to let immediate interactional issues pass (Firth 1996, Hüttner 2014). This can be seen in the following example, in which test-taker C14 picks up on a point made by their test partner C13, despite a relevant section of the recording being coded as unintelligible.

Extract B1.3: Centre RO1 – F2F mode

C13: I think that is a great activity (.) but it can exhaust
you and it might be a lot to watch
[unintelligible 00:06:56]
So (.) maybe talking to his friends?
C14: Yeah I think if he is bored (.) that would be a good
activity (.) to talk with his friends.
So (.) do we agree on catching up with his friends?

The impact of face masks on F2F communication could however be a point of interest for future studies.

B2 Part 3

Figure 3 shows the distribution of linguistic functions in the B2 Part 3 collaborative Speaking task across both modes of delivery. The most commonly used linguistic functions of Expressing an opinion (EOP), Justifying (JUST), Changing topic (CHAN) and Agreeing (AGR) were all similarly present in both modes – thus mirroring the findings for A2 and B1 levels. In line with the higher level of the candidates and more abstract nature of the task, there is a greater prevalence of Speculating (SPEC) than at the lower levels. The percentage of test-takers modifying their partner's speech was again higher for the F2F mode of the test but the percentage difference with the VC mode was less pronounced than in the B1 sample.

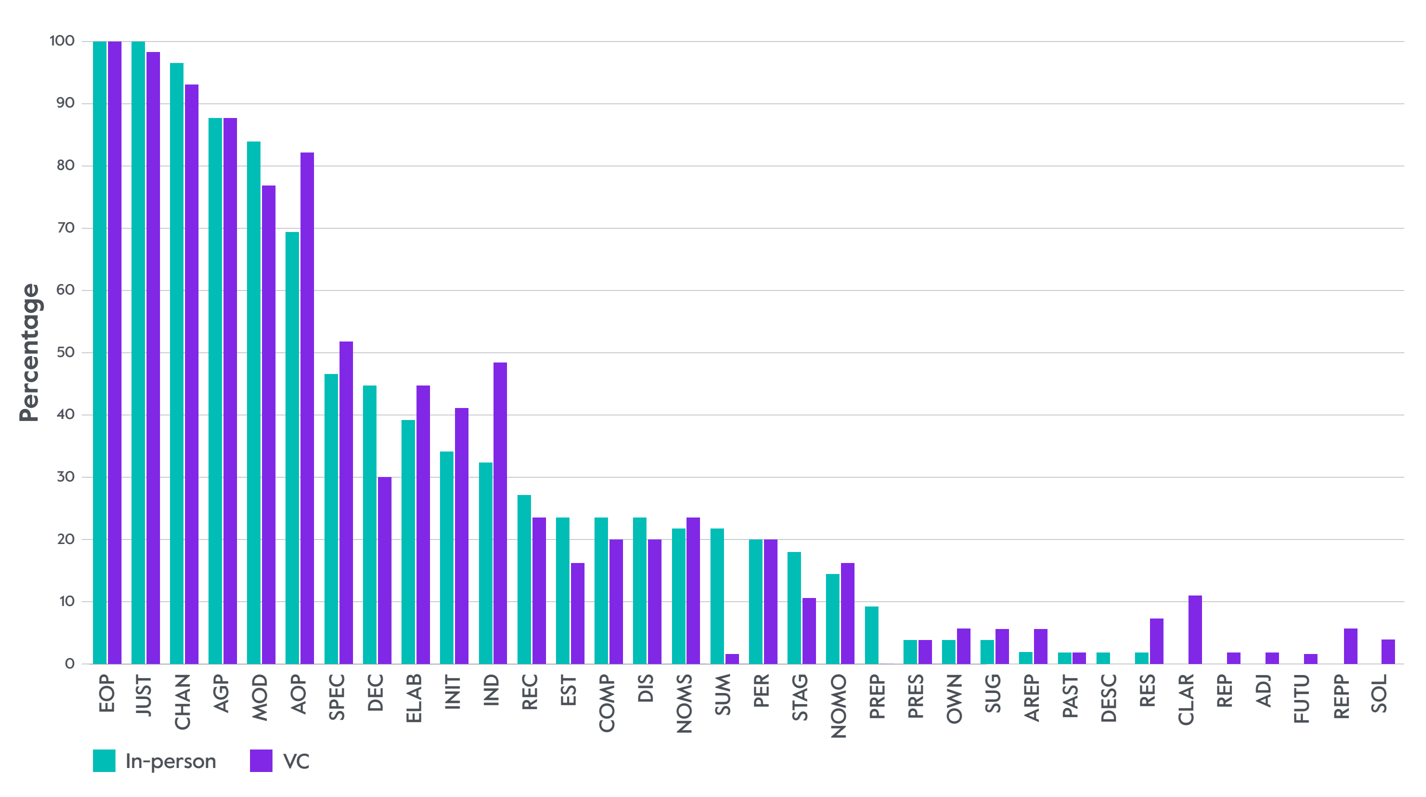


Figure 3: Distribution of linguistic functions in B2 Part 3 task (both modes)

Table 1 outlines the differences in linguistic functions that were highlighted as potentially statistically significant. Asking for opinion (AOP) and Asking for clarification (CLAR) had a higher frequency in the VC test mode. Expressing preference (PREF), Deciding (DEC), Modifying (MOD) and Summarising (SUM) were more frequent in the F2F mode. B2 Part 3 had the most statistically significant differences of all test parts across the three different levels, but effect sizes were small for all of these comparisons. Additionally, Bonferroni corrections indicated these differences were non-significant (critical value = 0.0014).

Table 1: Significant differences in linguistic output in B2 Part 3

Function	Test mode	Mean	SD	Z (df=55)	Sig. (2-tailed)	Effect size – r	Magnitude of effect
Expressing preference (PREF)	F2F	0.09	0.288	-2.236	0.025	-0.211	Small
	VC	0.00	0.000				
Asking for opinion (AOP)	F2F	1.07	0.892	-2.078	0.038	-0.196	Small
	VC	1.34	0.859				
Asking for clarification (CLAR)	F2F	0.00	0.000	-2.449	0.014	-0.231	Small
	VC	0.11	0.312				
Deciding (DEC)	F2F	0.50	0.603	-1.98	0.048	-0.187	Small
	VC	0.30	0.464				
Modifying (MOD)	F2F	1.59	1.108	-1.968	0.049	-0.186	Small
	VC	1.29	0.948				
Summarising (SUM)	F2F	0.21	0.414	-3.051	0.002	-0.288	Small
	VC	0.02	0.134				

Part 3 of B2 is divided into separate discussion and decision phases. Despite the differences mentioned above, both modes showed clear evidence of test-takers interacting with each other throughout these two phases. The interaction patterns were closer to established definitions of collaborative exchanges than parallel or asymmetric interaction (Galaczi 2008). In Extract B2.1, for example, we can see how in Line 5 in the first extract C2 picks up the idea of ‘computer games’ and appraises it. Similarly, in Line 9, C1 provides a hedged summing up of deliberations that is the product of both speakers’ contributions.

Extract B2.1: Centre A – F2F mode

1 C2: For young people probably celebrity visits and
2 discounts in my opinion are the advantage for shops.
3 C1: Yes (.) but also what do you think about computer
4 games? Young children can enjoy playing.
5→ C2: Yes (.) probably it could be a good idea (.)
6 I'm not very interested in it (.) but probably for
7 children and young people in general it would be
8 interesting to have them in a shop.
9→ C1: So (.) I guess (.) the computer games and discounts
10 for students.

In Extract B2.2 we can also see how in Line 3 C8 opens talk by eliciting opinion from their peer. Interestingly, while it was a point of difference between modes, the higher incidence of Asking for opinion (AOP) in the VC mode suggests evidence of mutuality and equality (Galaczi 2008) where candidates 'share the opportunities and responsibility' for topic initiation and development (Nakatsuhara 2013:35).

Extract B2.2: Centre B – VC mode

1 EX: Thank you (.) Now you have about a minute to decide
2 which two ideas would be most popular with young people.
3→ C8: Do you have anything about the café?
4 C7: I think the coffee (.) the café is the best one because
5 I think that as you said it attracts a lot of people and
6 also like tourists to visit the shop or the area.
7 I think the other would be discounts for students that
8 we didn't say it because it will attract a lot of students.
9→ C8: Yeah (.) but not all young people study I think it would be
10 better free Wi-Fi or celebrity visits

Examiner-led questions

A2 Part 2 Phase 2

On face value there seem to be some noteworthy differences in the percentages of test-takers demonstrating certain functions between the two modes (Figure 4). Both Comparing (COMP) and Expressing an opinion (EOP) were demonstrated by more test-takers in the F2F mode of the test. For Comparing (COMP), this may have partly been a reflection of the small sample size, with a relatively low overall number of instances across both modes (10 for the F2F test and six for VC).

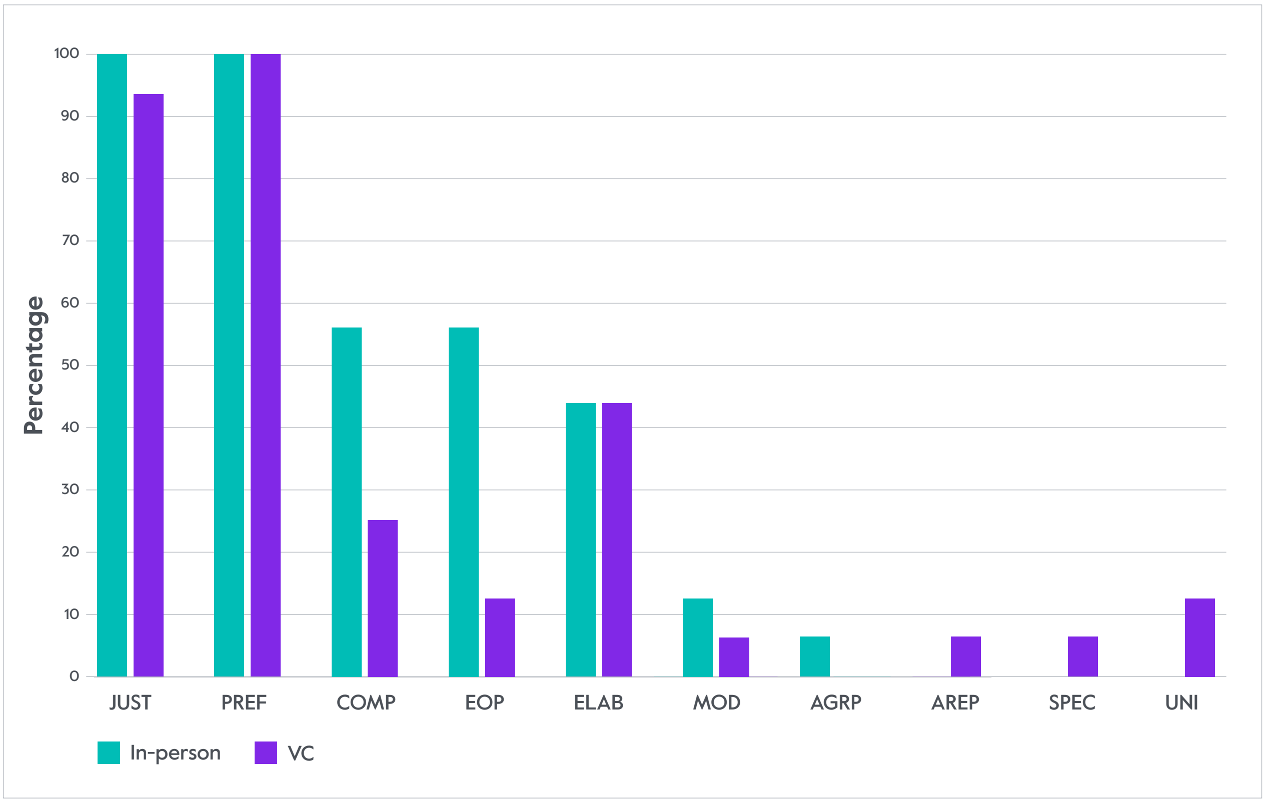


Figure 4: Distribution of linguistic functions in A2 Part 2 Phase 2 (both modes)

There was a significant statistical difference for Expressing an opinion (EOP) with a large effect size. However, it is worth looking at the mean totals for EOP alongside those of the other function in which a significant statistical difference was observed, Expressing preference (PREF). When we do so, we can see that EOP was more frequent in the F2F mode (a Mean of 1.13 compared to 0.25 for the VC mode), with the reverse pattern true for PREF, with a Mean of 4.00 for the VC mode, compared to 2.88 for the F2F test. Given the closeness of these two informational functions, it seems fair to view the performance elicited by this part of the test as largely comparable across the modes.

Table 2: Significant differences in linguistic output in A2 Part 2 Phase 2

Function	Test mode	Mean	SD	Z (df=55)	Sig. (2-tailed)	Effect size – r	Magnitude of effect
Expressing an opinion (EOP)	F2F	1.13	0.991	-2.070	0.038	-0.52	Large
	VC	0.25	0.463				
Expressing preference (PREF)	F2F	2.88	0.991	-2.121	0.034	-0.53	Large
	VC	4.00	0.000				

The potential conflation of these two informational functions was confirmed by a further Wilcoxon's Signed Rank test which indicated that when the two codes were combined, the difference between the two modes were non-significant (p=.157). Similarly, a Bonferroni correction also indicated these differences were not significant.

In sum, the proximity of the functions made inferential measures less reliable and this is a factor for future discourse analytic studies to bear in mind.

B1 Part 4

Looking at Figure 5, the fact that more test-takers justified (JUST) than expressed an opinion (EOP) in both modes of the test may seem slightly incongruous, with the expectation being that any justification would follow an already expressed opinion. This pattern is likely because some of the initial opinions were captured not as Expressing an opinion (EOP) but instead under Expressing preference (PREF).

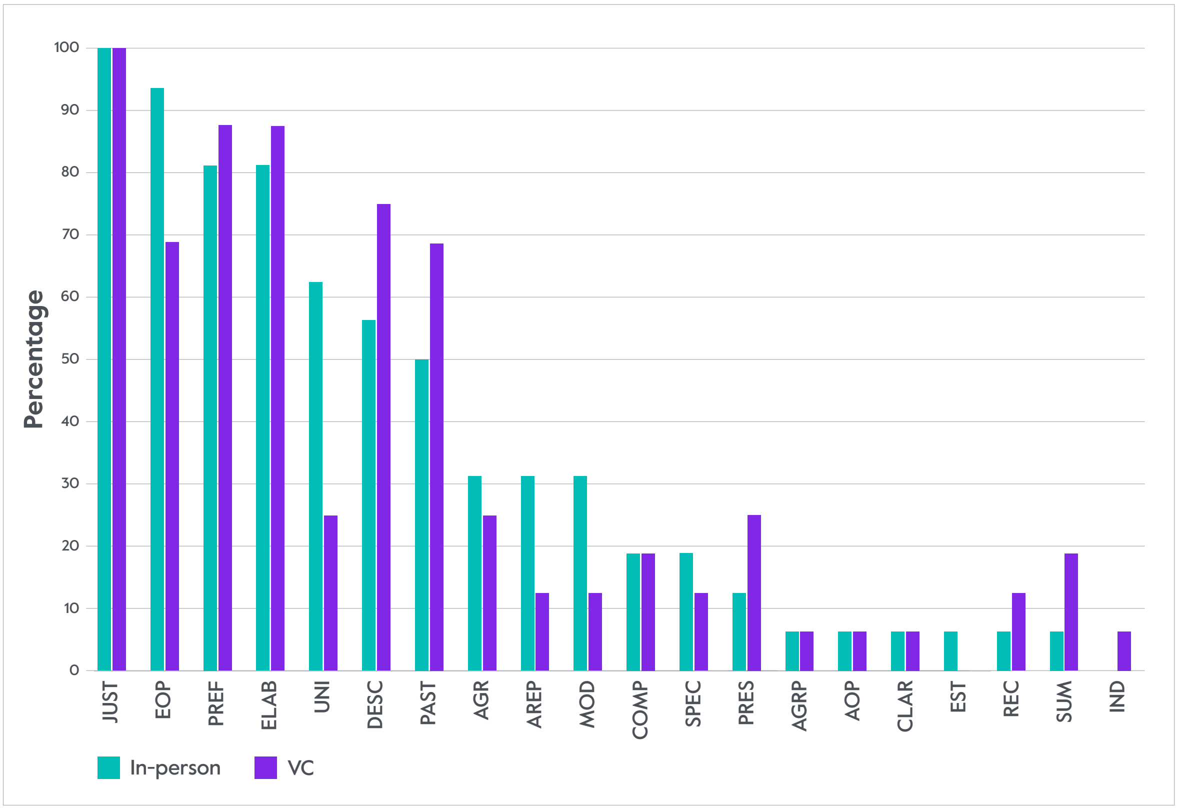


Figure 5: Distribution of linguistic functions in B1 Part 4 (both modes)

Despite more test-takers Expressing preference (PREF) in the VC mode, the mean figures showed that both PREF and Expressing an opinion (EOP) were more frequent in the F2F mode of the test. However, as with Part 3 of the test, there were no significant statistical differences observed in B1 Part 4. There was again a notably higher frequency of Unintelligible (UNI) speech for the F2F mode which was felt to be at least partly attributable to the difficulty in transcribing the recordings. Although more than twice as many test-takers asked for repetition (AREP) at least once during Part 4 of the F2F test (31.25% compared to 12.5% in the VC mode), the small sample size meant that this actually only equated to two more requests in total, again suggesting that comprehension between test participants was not significantly impacted. Asking for repetition (AREP) did not show a significant statistical difference across modes.

B2 Part 4

An overview of the distribution of linguistic functions in Part 4 of the B2 test is provided in Figure 6.

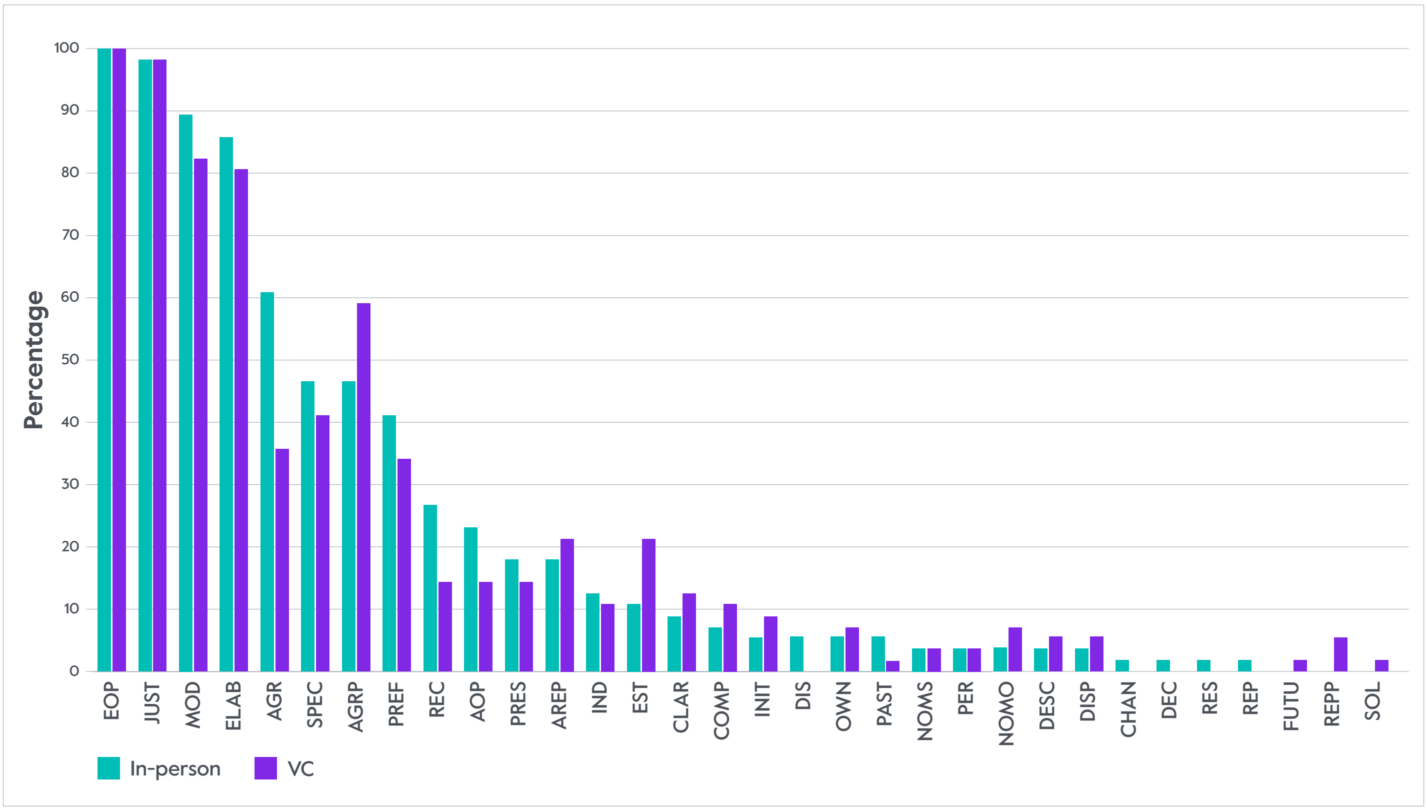


Figure 6: Distribution of linguistic functions in B2 Part 4 (both modes)

For Part 4 of B2, Agreeing (AGR), Agreeing (prompted) (AGRP), Expressing an opinion (EOP) and Reciprocating (REC) all showed statistically significant differences, of which AGR and EOP also had a medium effect size (Table 3). AGR and REC were demonstrated by more candidates in the F2F version of the test, whereas AGRP was used by more candidates in VC mode. All test-takers expressed opinions during Part 4 of both their tests, but the frequency was higher F2F than for VC tests. Bonferroni corrections indicated these differences were non-significant (critical value $p < 0.0016$) in the case of AGRP and REC but AGR and EOP were significant.

Table 3: Significant differences in linguistic output in B2 Part 4

Function	Test mode	Mean	SD	Z (df=55)	Sig. (2-tailed)	Effect size – r	Magnitude of effect
Agreeing (AGR)	F2F	1.57	1.757	-3.677	0.000	-0.347	Medium
	VC	0.57	0.850				
Agreeing (prompted (AGRP)	F2F	0.54	0.631	-2.237	0.025	-0.211	Small
	VC	0.82	0.811				
Expressing an opinion (EOP)	F2F	4.29	1.692	-3.513	0.000	-0.332	Medium
	VC	3.34	1.049				
Reciprocating (REC)	F2F	0.48	1.079	-2.485	0.013	-0.235	Small
	VC	0.14	0.353				

To what extent examiner behaviours contributed to these differences is an important question, as it might prove possible to help mitigate against them through additional specialised training for examiners delivering the VC version of the test. It is worth noting that the aim of any examiner training should not be to make the VC mode a clone of the F2F one. The language elicited in both modes can be slightly different, but still comparable.

Examiner speech

Within the constraints of the B2 First Interlocutor Frame, there is freedom for examiners to choose which questions to address to individual candidates or both candidates during Part 4 of the test. Attention was therefore given to any ways in which the framing of questions by examiners across the delivery modes appeared to impact on the language functions and nature of responses elicited in Part 4 of the test.

Examiner questions that did not nominate an individual candidate by name were more likely to elicit unprompted responses from both candidates, and thereby for candidates to demonstrate Agreeing (AGR), and potentially Reciprocating (REC), rather than Agreeing (prompted) (AGRP). Examiners nominated individual candidates by name in 76% of B2 VC Part 4 questions, compared to just 45% of F2F questions. A common pattern in VC delivery was for the examiner to first nominate an individual and then use one of the three available prompts from the interlocutor frame to direct a follow-up to the other candidate by name. Illustrations of these different examiner behaviours can be seen in the two excerpts below:

Extract B2.3: Centre C – F2F mode

1	Examiner:	Some people think we buy too many things these days
2		(.) what do you think?
3	C3:	I agree with that (.) do you?
4	C4:	Yes (.) of course.
5		I think that there are too many clothes and people
6		don't really know what they like.

Extract B2.3: Centre C – VC mode

1	Examiner:	[C3 name] is it a good idea for a country to spend
2		a lot of money building sports facilities?
3	C3:	I think so (.) because I think that sports are
4		necessary for health and I think that it is a good
5		way to make people do more sports.
6→	Examiner:	Do you agree [C4 name]?
7	C4:	Yes (.) totally (.) I think they are a part of life ...

There is evidence however that differences in examiner turn allocation alone do not fully account for variation in interaction patterns and levels of functions elicited from test-takers across the two modes. Where the same question type was present in both F2F and VC tests, it was generally more likely to result in a response from both candidates in the F2F mode. For example, when an initial question asking for opinion was not directed by the examiner to anyone by name, it resulted in responses by both test-takers in 79% of F2F instances (53 out of 67 times) but only 67% of the time in VC mode (20 out of 30).

In order to maximise the chance of obtaining satisfactory audio quality, cameras for F2F recordings were positioned close to candidates, meaning that examiners were generally not shown in shot. It is not possible therefore to be certain as to what extent eye contact and gestures were employed by all examiners either as a means of nominating individuals to speak or encouraging both test-takers to do so. However, at one centre where the examiner could be seen on all recordings, clear eye contact was made with both test-takers on all 23 occasions when the examiner asked an initial question asking for opinion without referring to either test-taker by name. In most instances (16 out of 23), this was coupled with a gesture of moving both hands together. This should be viewed as a likely significant contributing factor in both test-takers speaking without further prompting for 20 out of the 23 questions (87%). If replicated across other examiners' behaviour, it would illustrate the role of examiners' non-verbal communication in guiding Part 4 F2F interaction patterns.

The VC mode offered much less potential for such non-verbal communication. Examiners and test-takers were typically visible from around shoulder height. The same examiner who had been in shot for all F2F tests only attempted a similar hand gesture once in the seven occasions they asked questions asking for opinion without nominating test-takers in the VC mode.

Nominating both test-takers together was employed infrequently as a means of turn allocation by examiners in the VC mode (and not at all for F2F tests) but resulted in responses from both test-takers six out of the eight times it was used.

Discussion and conclusions

This research strand was focused on identifying differences in language functions elicited in the test-taker language across the VC and F2F modes of delivery.

Across the three exam levels and modes, the greatest difference in linguistic function use was at the B2 level – notably, in the case of Agreeing (AGR) and Expressing an opinion (EOP) in the examiner-led Part 4 where differences remained statistically significant after correction (see Table 3). As mentioned, this perhaps underscores the importance of the examiner role and examiner training for VC testing. At the A2 and B1 levels, differences were more marginal and could be explained by issues of category proximity (i.e., codes which are harder to consistently differentiate) – a methodological issue which future studies of this kind should consider alongside the potential impact of face masks and social distancing on capturing data.

In previous studies focusing on comparability of language functions used in VC and F2F versions of the single-candidate format IELTS Speaking test, Asking for clarification (CLAR) was found to be the function with most marked differences in use under the two test modes (Nakatsuhara et al 2017). Given the nature of IELTS as a multilevel test, with a candidature that covers a wider range of CEFR levels, direct comparisons between the different tests and cohorts should be treated with caution. There does not appear however to have been such a pronounced difference in asking for clarification across the two modes for either the B2 study or the later A2 and B1 trials. Looking at the B2 results, a significant difference (with a small effect size) was found between the VC and F2F modes in Part 3 of the test, but just over 10% of candidates asked for clarification in this part of the test in VC mode.

The lower percentages of test-takers asking for clarification can, in part at least, likely be attributed to the introduction of a new category, Asking for repetition (AREP). In Part 4 in particular, this was used by 21.4% of test-takers in the VC mode and 17.9% for F2F tests. The need for social distancing and mask wearing may have contributed to some of these requests for repetition in the F2F mode. Even when the two categories were combined, asking for clarification (CLAR) or repetition (AREP) was still less prevalent throughout the VC test than might have been anticipated based on previous studies' findings. There are several possible explanations for this. Improvements to the technology may be one contributing factor, another could be familiarity with VC (see the opening article of this issue).

Taken together, these findings provide a hopefully useful snapshot of the terrain similar function-focused comparability studies will need to navigate.

References

Brooks, L (2003) Converting an observation checklist for use with the IELTS Speaking test, *Research Notes* 11, 20–21.

Cooke, S (2015) *Configuring the game of speaking: Interactional competence in the IELTS Oral Proficiency Interview across two modes of response*, unpublished master's dissertation, Lancaster University.

Dancey, C P and Reidy, J (2011) *Statistics without Maths for Psychology* (Fifth edition), Harlow: Pearson Education Limited.

Firth, A (1996) The discursive accomplishment of normality: On 'lingua franca' English and conversation analysis, *Journal of Pragmatics* 26, 237–259.

Galaczi, E (2008) Peer-peer interaction in a speaking test: The case of the first certificate in English examination, *Language Assessment Quarterly* 5 (2), 89–119.

Glasson, N, Mullooly, A and Galaczi, E (2022, March 7–11) *What does beginner interaction look like? Video-call candidate interactions with a focus on speaker continuation and transition*, paper presented at 43rd Language Testing Research Colloquium, Tokyo.

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*, English Profile Studies Volume 2, Cambridge: UCLES/Cambridge University Press.

Hüttner, J (2014) Agreeing to disagree: 'doing disagreement' in assessed oral L2 interactions, *Classroom Discourse* 5 (2), 194–215.

IBM Corp. (2017) *IBM SPSS Statistics for Windows, Version 25.0*, Armonk: IMB Corp.

Inoue, C and Nakatsuhara, F (2021) Validation of a large-scale task-based test: Functional progression in dialogic speaking performance, in Sudharshana, N P and Mukhopadhyay, L (Eds) *Task-Based Language Teaching and Assessment*, Singapore: Springer, 217–247.

Jin, Y and Zhang, L (2016) The impact of test mode on the use of communication strategies in paired discussion, in Yu, G and Jin, Y (Eds) *Assessing Chinese Learners of English*, Basingstoke: Palgrave Macmillan, 61–84.

McDonald, J H (2014) *Handbook of Biological Statistics* (Third edition), Baltimore: Sparky House Publishing.

Nakatsuhara, F (2013) *The Co-construction of Conversation in Group Oral Tests*, Frankfurt am Main: Peter Lang.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2016) *Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery. A preliminary comparison of test-taker and examiner behaviour*, IELTS Partnership Research Papers 1, IELTS Partners: British Council/IDP: IELTS Australia/Cambridge Assessment English.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2021) Video-conferencing speaking tests: do they measure the same construct as face-to-face tests?, *Assessment in Education: Principles, Policy & Practice* 28 (4), 369–388.

O'Loughlin, K J (2001) *The Equivalence of Direct and Semi-direct Speaking Tests*, Studies in Language Testing Volume 13, Cambridge: UCLES/Cambridge University Press.

O'Sullivan, B, Weir, C J and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19 (1), 33–56.

Ockey, G J, Gu, L and Keehner, M (2017) Web-based virtual environments for facilitating assessment of L2 oral communication ability, *Language Assessment Quarterly* 14 (4), 346–359.

Schegloff, E A (1993) Reflections on quantification in the study of conversation, *Research on Language and Social Interaction* 26 (1), 99–128.

Seuren, L M, Wherton, J, Greenhalgh, T and Shaw, S E (2021) Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction, *Journal of Pragmatics* 172, 63–78.

Stivers, T (2015) Coding social interaction: A heretical approach in conversation analysis?, *Research on Language and Social Interaction* 48 (1), 1–19.

van Lier, L (1989) Reeling, writhing, drawing, stretching, and fainting in coils: Oral proficiency interviews as conversation, *TESOL Quarterly* 23 (3), 489–508.

Waring, H Z (2002) Displaying substantive recipiency in seminar discussion, *Research on Language and Social Interaction* 35, 453–479.

Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

Zhang, L and Jin, Y (2021) Assessing interactional competence in the computer-based CET-SET: An investigation of the use of communication strategies, *Assessment in Education: Principles, Policy & Practice* 28 (4), 389–410.

Zhou, Y (2015) Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking, *Language Testing in Asia* 5 (2).

Appendix

Informational functions	
COMP	Comparing
DESC	Describing
ELAB	Elaborating
EOP	Expressing an opinion
FUTU	Providing personal information (Future)
JUST	Justifying
PAST	Providing personal information (Past)
PREF	Expressing preference
PRES	Providing personal information (Present)
SPEC	Speculating
STAG	Staging
SUG	Suggesting
SUM	Summarising
Interactional functions	
AGR	Agreeing
AGRP	Agreeing (prompted)
AIN	Asking for information
AOP	Asking for opinion
AREP	Asking for repetition
CLAR	Asking for clarification
CORR	Correcting
DIS	Disagreeing
DISP	Disagreeing (prompted)
EST	Establishing common ground
IND	Indicating understanding
MOD	Modifying
OTHER	Checking other's understanding
OWN	Checking own understanding
PER	Persuading
REP	Conversational repair
RES	Responding to request for repetition or clarification

Managing the Interaction	
INIT	Initiating
CHAN	Changing topic
DEC	Deciding
NOMO	Nominating other
NOMS	Nominating self
REC	Reciprocating
Technical	
ADJ	Adjusting equipment
REPP	Reporting technical problem
SOL	Solving technical problem
Other	
BRK	Breakdown
NOMI	Nominating individual *(Examiner Only)*
NOMP	Nominating pair *(Examiner Only)*
UNI	Unintelligible

Modified from O'Sullivan et al (2002)

The eye of the stakeholder – perceptions of remote speaking

Nicholas Glasson, Research, Cambridge University Press & Assessment
Amy Devine, Research, Cambridge University Press & Assessment

Introduction

As Craig and Kim (2010) observed over a decade ago, 'videoconferencing has had great potential for years' (2010:9), but it is only in the last few years that more attention has been focused on exploring the operationalising of video-call (VC) technology for assessment purposes. In particular, studies of test-taker and examiner perceptions are still relatively scarce within the field.

One of the major concerns with VC speaking assessment is its dependence on a stable internet connection. For example, in a US and China-focused study of an online speaking test, while perceptions of the VC mode were generally positive, this was notably less the case in China, where poor connectivity led to a much higher incidence of interruptions and unnatural pauses (Ockey, Timpe-Laughlin, Davis and Gu 2019). As those authors point out, 'subtle disruptions, such as delays in the audio signal or dropped video, have the potential to disrupt turn taking and create misunderstandings between speakers, complicating scoring and again impacting inferences regarding test takers' abilities' (Ockey et al 2019:18). A more recent IELTS study highlighted how important the platform is to perceptions of stability and connectivity (Lee, Patel, Lynch and Galaczi 2021) – again with broadly positive test-taker reactions but variance depending on what platforms and equipment were used.

The visual aspect of online interaction also plays an important role in perceptions of the mode. The distance between candidate and monitor, ability to use hand gestures, and lack of eye contact have all been identified as influential factors (Lee et al 2021, Ockey et al 2019, Tsunemoto, Lindberg, Trofimovich and McDonough 2022).

Other studies have also shown how a lack of video can make turn-taking harder and reduce a sense of 'being there' (Jin and Zhang 2016, Ockey, Gu and Keehner 2017). Indeed, one study highlighted participants not only wanted to see their peers clearly but also themselves for fear they would not be visible (Lee et al 2021).

From an examiner perspective, studies have suggested that while examiners can adapt to the VC mode, the modality is seen to impact both the mechanics of test delivery and rating. In the case of IELTS, research highlighted how examiners found the VC mode 'subdued' and requiring more gesturing and physical effort (Lee et al 2021:20). That study indicated how articulation and speed of examiner speech, intonation, gestures, turn-taking, and requests for clarification were all potentially influenced by mode and how poor sound quality can impact on rater confidence in score decisions.

Interestingly, despite the technological challenges VC tests present, studies have quite consistently reported test-takers feel the online mode allows them adequate opportunity to show their language abilities (Kim and Craig 2012, Lee et al 2021, Ockey et al 2019). Studies indicate no consistent test-taker preference for face-to-face (F2F) interaction although some research indicates test-takers preferred examiner-mediated testing online rather than F2F because it reduced anxiety (Kim and Craig 2012, Davis, Timpe-Laughlin, Gu and Ockey 2016). Conversely, other evidence highlights that the distance between test-taker and their monitor, rules prohibiting the use of gestures, and being able to manipulate sound volume can all be sources of anxiety in VC tests (Lee et al 2021). Test anxiety is notably a very fluid and context-dependent variable. In the present study the focus was on the reported perception and causes of test anxiety by test-takers.

Research questions

This article explores two research questions:

- What are the examiners' perceptions of the VC and F2F test modes?
- What are the test-takers' perceptions of the VC and F2F test modes?

Methodology

As noted in the opening article to this issue, the B2 First (henceforth B2) study was conducted from October to December 2020, with the A2 Key/B1 Preliminary (henceforth A2/B1) for Schools study following in February to June 2021. Instruments and approaches used in the B2 study were replicated for the later A2/B1 study with only minor modifications. This strand of the wider remote speaking research involved a mix of online surveys and focus groups.

All VC tests in both studies were conducted via Zoom. All test-takers were asked to provide informed consent. Where a test-taker was under 18, parental consent was obtained.

Data

Examiner surveys

After completing all F2F and VC speaking tests with the participating test-takers from their centre, examiners were asked to complete an examiner feedback questionnaire about their perceptions of the two test delivery modes. Questionnaire topics included gathering background information; perceptions of delivering and rating the tests; examiners' preference of test mode, and the perceived impact of VC test mode on examiners and test-takers; and examiners' suggestions for improving the VC test. The questionnaires consisted of 49 questions for A2/B1 and 52 for B2, including closed-ended, multiple choice, Likert scale and open-ended questions. Every participating examiner completed the questionnaire, each taking approximately 15 minutes to do so.

Examiner focus groups

Examiner focus groups were used in the initial phase of the A2/B1 study. The key themes explored in this session were: general perceptions of the online test; rating, timing and managing materials; comparisons of F2F and VC speaking.

Test-taker surveys

After their second test, test-takers were asked to complete a feedback questionnaire about their perceptions of the two test delivery modes. Questionnaire topics included questions gathering background information, difficulty of each part of the test in the two test modes, perceptions of the audio and video quality in the VC test, test-takers' preference of test mode, and test-takers' suggestions for improving the VC test.

The questionnaire consisted of 42 questions for A2/B1 and 37 for B2, including closed-ended, multiple choice, Likert scale and open-ended questions. The questionnaire took approximately 10 to 15 minutes for test-takers to complete. For the A2/B1 study, a total of 127 test-takers across both levels completed the questionnaire (69% response rate). For the B2 study, a total of 44 test-takers completed the questionnaire (76% response rate).

Structured test-taker interviews

At the end of their second speaking test (F2F or VC depending on the test order allocated to each pair), the interlocutor conducted a brief, structured focus group interview with each pair of test-takers. The interview consisted of five main questions about test-takers' experiences of the two test modes and took approximately 10 minutes. The interlocutor recorded notes summarising test-takers' responses. In some cases, the interlocutor recorded the test-takers' responses verbatim; in others the interlocutor provided a summary of the test-takers' views.

Analysis

Closed-response questions in the examiner and test-taker feedback questionnaires were analysed using descriptive statistics, and where the sample size was sufficient and questionnaire design allowed, Wilcoxon's Signed Rank test was used to compare ratings in the two test modes. IBM SPSS Statistics 25 was used for the descriptive and inferential statistics.

Open-ended responses from the two questionnaires and the notes from the test-taker focus groups were used to interpret the statistical analysis of closed-response questions. Open-ended response data was analysed for recurrent themes using MAXQDA 2020.

Candidate profiles

The B2 study

60 volunteer test-takers took part in this study: 32 in Spain, 24 in Italy and four in Portugal. Of these volunteers, 58 were learners studying towards taking their live B2 exam, with most due to do so within a month of participating in the study. The other two test-takers were 'dummy' candidates from test centres where an uneven number of volunteers meant one participant would otherwise have been left without a test partner – although live B2 Speaking tests can sometimes be taken in a group of three, this study focuses solely on the more typical paired format.

Additional background information was collected through a survey, which was completed by 44 test-takers (excluding dummy candidates). Within that sample of 44, the majority of test-takers were between 16 to 25 years old, with the youngest aged 14/15 years old; 22 were male and 22 female; their first languages included Spanish (50%), Italian (43%), Portuguese (5%) and Romanian (2%).

The A2/B1 study

Phase 1 involved an initial cohort of 12 A2 level test-takers and 16 B1 level test-takers drawn from two centres, one in Spain, the other in Italy. At this point in the study, no candidate data was being routinely collected via a survey, but all participants were learners aiming to take the exam for which they trialled and were all typical for the level in terms of proficiency, age and gender split. For Phase 2, a total 184 volunteer test-takers took part in the wider-scale study; details of this larger trial cohort can be found in Table 1.

Table 1: A2/B1 Phase 2 trial cohort overview

Country	Test-taker total (A2)	Test-taker total (B1)
Spain	26	28
Italy	26	30
Mexico	14	12
Romania	16	16
Vietnam	10	6
Totals	**92**	**92**

All volunteer test-takers were learners studying towards taking their A2 or B1 for Schools exam. Almost all test-takers were planning to take the for Schools version. For A2, 45.6% fell into the 10 to 12 age group, and 54.4% into the 13 to 15 age group. For B1 for Schools, 71.4% fell into the 13 to 15 age group, and 20% in the 16 to 18 age group. Across both A2 and B1 cohorts, 53% were female, 43% male and the remaining 4% were respondents declining to say or identifying as non-binary.

Overall, the demographic information on the trial participants broadly reflects that of the B2, A2 and B1 for Schools[1] populations (according to operational data) in terms of age and gender, making the findings and conclusions drawn broadly generalisable to these test-taker populations.

Examiner profiles

The B2 study

19 certificated B2 Speaking Examiners (SEs) participated in the research. Examiners reported an average of 22.21 years of EFL teaching experience (Standard Deviation (SD) = 10.02, min = 10, max = 45) and an average of 10.58 years' experience as a B2 SE (SD = 8.1, min = 2, max = 33). 10 of the 19 were Team Leaders as well as SEs.

The A2/B1 for Schools study

A total of 29 examiners took part over both phases of trialling, and 27 were involved in a multiple-marking exercise in Phase 2. Phase 1 trials involved four examiners who had prior experience of VC-mode trialling from the earlier B2 study. Three out of the four examiners also participated in Phase 2, during which 22 examiners were involved in both VC and F2F trials, whilst six examiners carried out VC tests only. 16 examiners participated as both interlocutor and assessor in Phase 2, while six acted as assessor only and six were interlocutor only. All examiners were certificated to examine A2 and B1 levels. Additional information was collected through a survey, which was completed by all 29 examiners.

[1] One minor variation is that the 10 to 12 age group made up a smaller percentage of the B1 for Schools participants (8.6%) than operational candidature (24.2%).

All 29 examiners reported more than six years of English Language Teaching experience and a majority (75%) had five or more years' experience as A2 for Schools and B1 for Schools SEs. 10 of the 29 were Team Leaders as well as SEs, and one examiner was a Professional Support Leader.

Results

The results are reported according to the main areas highlighted by relevant literature, and which emerged in the course of these two studies: test delivery, rating performance, audio/visual quality, test difficulty, opportunity to demonstrate English ability, test anxiety, and mode preferences.

Test delivery (examiners)

According to the closed-ended questionnaire responses, examiners' perceptions of test delivery were comparable between the two test modes (Tables 2 to 4). Across all three levels, for both F2F and VC, most examiners agreed or strongly agreed that they felt comfortable delivering the tests as a whole, that delivering each test part was straightforward, and the interlocutor frames were straightforward to manage and use.

There were negligible differences in examiners' mean ratings of delivering the tests in F2F and VC mode. Means on the questionnaire responses were not compared statistically due to the small sample size, but it is worth noting that the VC test was consistently perceived as more challenging to deliver, but by a very small margin. The limited training which participating examiners received prior to the trial is a possible reason for this slight difference, and this may explain the higher SDs for the VC mode, indicating a greater variety in responses from examiners – potentially due to their limited familiarity with the VC mode of delivery. Nevertheless, it is reassuring that even with such limited preparation, examiners perceived the VC mode as straightforward to deliver and very similar to the F2F test.

Table 2: Descriptive statistics for examiner responses related to test delivery for A2 for Schools

A2 Examiner survey items (delivery)	Test mode	N	Min.	Max.	Mean	SD
Overall, I felt comfortable in delivering the A2 Speaking test in the ... format.	F2F	20	4	5	4.90	0.31
	VC	20	3	5	4.45	0.61
I found it straightforward to deliver Part 1 (interview) of the Speaking test in the ... format.	F2F	20	4	5	4.85	0.37
	VC	21	3	5	4.52	0.60
I found it straightforward to deliver Part 2 (collaborative task) of the Speaking test in the ... format.	F2F	20	4	5	4.85	0.37
	VC	21	3	5	4.38	0.60
The interlocutor frame was straightforward to manage and use in the ... format.	F2F	20	4	5	4.80	0.41
	VC	21	3	5	4.48	0.60
1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree						

Table 3: Descriptive statistics for examiner responses related to test delivery for B1 for Schools

B1 Examiner survey items (delivery)	Test mode	N	Min.	Max.	Mean	SD
Overall, I felt comfortable in delivering the B1/B1 for Schools Speaking test in the ... format.	F2F	21	4	5	4.81	0.40
	VC	20	3	5	4.65	0.59
I found it straightforward to deliver Part 1 (interview) of the Speaking test in the ... format.	F2F	20	4	5	4.85	0.37
	VC	20	3	5	4.70	0.57
I found it straightforward to deliver Part 2 (individual task) of the Speaking test in the ... format.	F2F	20	4	5	4.85	0.37
	VC	0[2]	0	0	0	0.00
I found it straightforward to deliver the Part 3 (collaborative task) of the Speaking test in the ... format.	F2F	20	4	5	4.85	0.37
	VC	20	3	5	4.45	0.76
I found it straightforward to deliver Part 4 (discussion) of the Speaking test in the ... format.	F2F	20	4	5	4.85	0.37
	VC	20	3	5	4.55	0.76
The interlocutor frame was straightforward to manage and use in the ... format.	F2F	21	4	5	4.81	0.40
	VC	21	3	5	4.62	0.59
1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree						

Table 4: Descriptive statistics for examiner responses related to test delivery for B2

B2 Examiner survey items (delivery)	Test mode	N	Min.	Max.	Mean	SD
Overall, I felt comfortable in delivering the B2 Speaking test ...	F2F	12	4	5	4.83	0.39
	VC	12	4	5	4.67	0.49
I found it straightforward to deliver Part 1 (interview) of the Speaking test ...	F2F	12	4	5	4.92	0.29
	VC	12	4	5	4.75	0.45
I found it straightforward to deliver Part 2 (long turn) of the Speaking test ...	F2F	12	4	5	4.92	0.29
	VC	12	4	5	4.67	0.49
I found it straightforward to deliver Part 3 (collaborative task) of the Speaking test ...	F2F	12	4	5	4.83	0.39
	VC	12	4	5	4.58	0.51
I found it straightforward to deliver Part 4 (discussion) of the Speaking test ...	F2F	12	4	5	4.92	0.29
	VC	12	4	5	4.67	0.49
The interlocutor frame was straightforward to manage and use ...	F2F	12	5	5	5.00	0.00
	VC	12	4	5	4.75	0.45
1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree						

[2] Due to an error in the survey instrument, this data point was not available for analysis.

Across all three exam levels, open-response items and focus groups highlighted two main reasons for the perception of the VC tests being more challenging to deliver. The first was the limited scope to use gesture or other non-verbal cues to manage the interaction, as one B2 examiner commented:

> *'Prompting and time indication with gesture and eye contact was not possible. ... I am aware that I need to learn a new way to prompt candidates when timing is short and to indicate time management better in order to be able to interrupt speaking comfortably.'*
> (B2 Examiner K)

Another key theme many examiners touched upon was the added challenge of managing materials while delivering the interlocutor frame in VC mode:

> *'... and handling the materials was more difficult, because I had to try to make sure what was on screen was correct, as well as read the instructions while trying to maintain "eye contact"... Fumbling way more [with materials].'*
> (A2/B1 Examiner 20)

In both studies examiners commented that many of the challenges they faced in delivering VC tests could be ameliorated with training, practice and adaptations e.g. headsets and having a second monitor to view the interlocutor frame.

Rating performance (Examiners)

According to the closed-ended questionnaire responses, examiners' perceptions of rating test-taker performance were also comparable in F2F and VC test modes (see Tables 5 to 7). The majority of examiners agreed or strongly agreed that it was comfortable or straightforward to rate test-taker performance in the two test modes. However, consistently there were slightly lower ratings given for ease of rating candidate performance in the VC test mode than in the F2F mode.

Across all survey items and levels there is a higher SD for the VC mode, suggesting a greater variance in responses. The lowest mean values for the VC tests were seen in rating Pronunciation (A2 and B2), Interactive Communication (all three tests) and Global Achievement (A2 and B1) (see Tables 5 to 7). However, at a mean above 4, they were nevertheless rated highly.

Table 5: Descriptive statistics for examiner responses related to rating test-takers' performance for A2 for Schools

A2 Examiner survey items (rating)	Test mode	N	Min.	Max.	Mean	SD
Overall, I felt comfortable rating candidate performance in the … A2 Speaking test.	F2F	21	4	5	4.81	.402
	VC	22	3	5	4.41	.666
I found it straightforward to apply the Grammar and Vocabulary descriptors in the … format.	F2F	19	4	5	4.79	.419
	VC	19	3	5	4.47	.697
I found it straightforward to apply the Pronunciation descriptors in the … format.	F2F	19	3	5	4.47	.612
	VC	19	3	5	4.32	.671
I found it straightforward to apply the Interactive Communication descriptors in the … format.	F2F	19	4	5	4.79	.419
	VC	19	3	5	4.32	.671
I found it straightforward to apply the Global Achievement descriptors in the … format.	F2F	21	4	5	4.71	.463
	VC	19	3	5	4.42	.692
I feel confident about the accuracy of my ratings in the … format.	F2F	22	4	5	4.64	.492
	VC	23	3	5	4.35	.573
1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree						

Table 6: Descriptive statistics for examiner responses related to rating test-takers' performance for B1 for Schools

B1 Examiner survey items (rating)	Test mode	N	Min.	Max.	Mean	SD
Overall, I felt comfortable rating candidate performance in the … B1 Speaking test.	F2F	23	4	5	4.78	.422
	VC	22	3	5	4.45	.671
I found it straightforward to apply the Grammar and Vocabulary descriptors in the … format.	F2F	21	4	5	4.76	.436
	VC	19	3	5	4.68	.582
I found it straightforward to apply the Discourse Management descriptors in the … format.	F2F	21	4	5	4.76	.436
	VC	19	3	5	4.63	.597
I found it straightforward to apply the Pronunciation descriptors in the … format.	F2F	21	4	5	4.57	.507
	VC	19	3	5	4.58	.607
I found it straightforward to apply the Interactive Communication descriptors in the … format.	F2F	21	4	5	4.76	.436
	VC	19	3	5	4.47	.612
I found it straightforward to apply the Global Achievement descriptors in the … format.	F2F	18	4	5	4.78	.428
	VC	18	3	5	4.33	.686
I feel confident about the accuracy of my ratings in the … format.	F2F	23	4	5	4.65	.487
	VC	23	3	5	4.35	.647
1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree						

Table 7: Descriptive statistics for examiner responses related to rating test-takers' performance for B2

B2 Examiner survey items (rating)	Test mode	N	Min.	Max.	Mean	SD
Overall, I felt comfortable rating candidate performance in the … B2 Speaking test.	F2F	19	4	5	4.84	0.375
	VC	19	4	5	4.74	0.452
I found it straightforward to apply the Grammar and Vocabulary descriptors in the … format.	F2F	9	4	5	4.89	0.333
	VC	9	4	5	4.78	0.441
I found it straightforward to apply the Discourse Management descriptors in the … format.	F2F	9	4	5	4.89	0.333
	VC	9	4	5	4.67	0.500
I found it straightforward to apply the Pronunciation descriptors in the … format.	F2F	9	4	5	4.89	0.333
	VC	9	3	5	4.44	0.726
I found it straightforward to apply the Interactive Communication descriptors in the … format.	F2F	9	4	5	4.78	0.441
	VC	9	3	5	4.11	0.601
I found it straightforward to apply the Global Achievement descriptors in the … format.	F2F	11	4	5	4.91	0.302
	VC	8	4	5	4.63	0.518
I feel confident about the accuracy of my ratings in the … format.	F2F	19	4	5	4.68	0.478
	VC	19	2	5	4.47	0.772
1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree						

Open-ended survey responses provided further evidence of a range of examiner perspectives on the different modes and how easy they were to assess in. Several comments and observations made by examiners related to the different interactional environment they experienced online and how it altered their approach.

> *'Difficult to mark interaction as no eye contact or body language (candidates actually turning around to talk to their partners).'*
> (B2 Examiner G)

> *'While examining face to face there is nothing that you miss while in the online environment there may be problems with sound, pronunciation may not always be very intelligible, and interaction is more difficult to establish and assess.'*
> *(A2/B1 Examiner 11)*

While examiners voiced concerns related to the online environment, there were also positive appraisals of the assessor specifically being slightly more removed from the test delivery.

> *'Turning off camera was a positive thing for assessing. I could just focus on listening.'*
> (A2/B1 Examiner 27)

The importance of clear audio and visual channels were a consistent theme underlying examiner comments on rating perceptions. This was mirrored in test-taker responses.

Audio and video quality for VC mode (test takers)

Across all three levels, the majority of test-takers reported little or no impact on their performance caused by poor audio or visual access (Figure 1 and Figure 2). Given that previous studies had highlighted this, it is both encouraging and also a sign of the improvements in technology, particularly since the global Covid-19 pandemic increased the use of VC. Both studies were carried out during the pandemic, where many test-takers had had to adapt to online learning to some extent (see the opening article's section 'Participants' experience with the internet and VC technology').

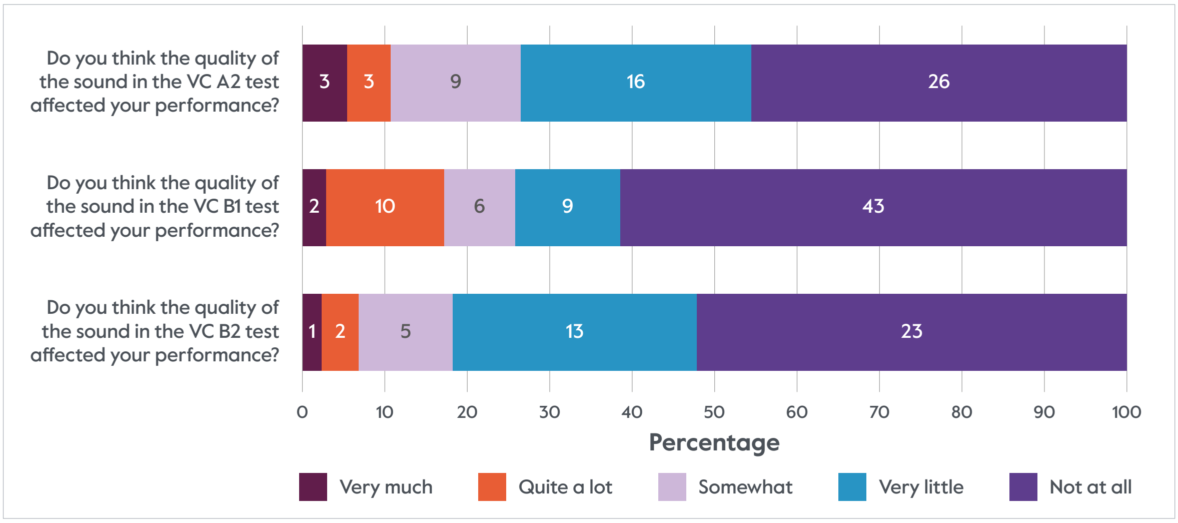


Figure 1: Test-takers ratings of the impact of sound quality on test performance for A2, B1 and B2 exams

In both figures, the relatively high incidence of issues reported for the B1 for Schools was attributable to a specific centre that had recurring issues with connectivity.

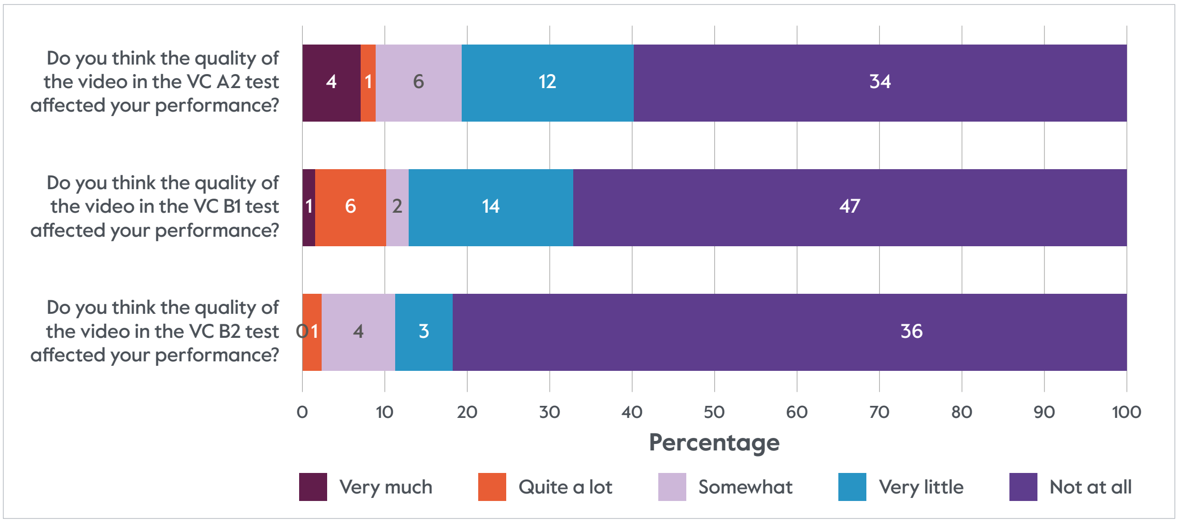


Figure 2: Test-takers ratings of the impact of video quality on test performance for A2, B1 and B2 exams

Test-takers that reported issues in sound or visuals often noted their temporary nature e.g., ‘sometimes the audio was unclear when the internet connection went for a few seconds’ (B2 Candidate 11, Centre B). Primarily, test-taker responses focused on a concern to hear the interlocutor and their partner although a minority of responses also expressed a fear they would not be assessed fairly. One B1 test-taker mentioned that sound quality ‘affected my performance because maybe the examiners didn’t hear my fluency and pronunciation very well’. Although there were reported cases of audio and video issues, for most test-takers (>70%) this was not perceived to have impacted their performance directly. Poor connectivity was, however, often implicated in perceptions of test difficulty, ability to showcase language, and test anxiety, which are explored in the following sections.

Test difficulty (test takers)

Figure 3 highlights that across all three exams a majority of test-takers perceived no difference in test difficulty across modes (A2=39%, B1=57%, B2=52%). A smaller proportion felt the VC mode was more difficult (A2=33%, B1=20%, B2=34%). In all three exams, the F2F mode was perceived as more difficult by a minority of test-takers (A2=28%, B1=22%, B2=14%) with the perception of difficulty with F2F testing decreasing by level.

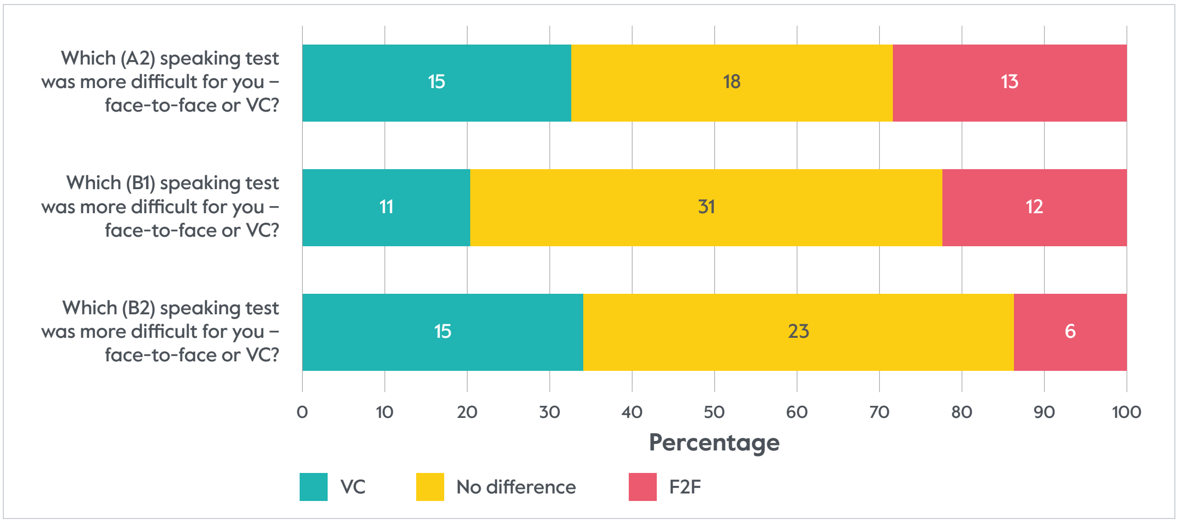


Figure 3: Test-taker perceptions of overall test difficulty for A2, B1 and B2 exams

For the B2 study test-takers were additionally asked to rate the difficulty of each part[3] of the speaking test in the F2F and VC test modes. There were no significant differences in ratings of difficulty across the two test modes (see Table 8).

[3] For the A2 for Schools and B1 for Schools exams, one of which has only two test parts, this approach was avoided due to a concern younger test-takers would be much less aware of test parts.

Table 8: Descriptive statistics for test-takers' perceptions of the difficulty of the B2 test parts in F2F and VC test mode

Q: How difficult was each part of the test?							
B2 test part	**Test mode**	**Min.**	**Max.**	**Mean**	**SD**	**Z score**	**Asymp. Sig. (2-tailed)**
Part 1	F2F	3	5	3.91	0.802	-1.633	0.102
	VC	3	5	4.00	0.807		
Part 2	F2F	1	5	3.36	0.838	-0.577	0.564
	VC	2	5	3.43	0.759		
Part 3	F2F	2	5	3.36	0.780	-1.213	0.225
	VC	1	5	3.25	0.866		
Part 4	F2F	2	5	3.55	0.791	-0.842	0.400
	VC	1	5	3.45	0.901		
1 = Very difficult, 2 = Difficult, 3 = OK, 4 = Easy, 5 = Very easy							

The challenges noted by test-takers in the two modes were often very different. In F2F trials test-takers commented that it can be daunting to be paired with a stranger or someone much older (e.g., A2). With trials being conducted under strict Covid-19 regulations, another challenge of F2F testing was being able to hear your partner clearly through a facemask. The VC mode challenges were often connected to a feeling of distance from the interlocutor and their partner or that task visuals were blurry when presented on screen. Across all three levels, test-takers commented on this issue, and it can be observed in the lower mean values for Part 2 and Part 3 in B2 (Table 8).

Opportunities to demonstrate English ability (test takers)

Test-takers were asked to what extent each test allowed them to show their full English ability. The mean rating in both the B2 and combined A2/B1 studies were significantly lower for the VC test mode than F2F; however, the effect size of the difference was small in both cases. So, while there was a difference in perceptions, the magnitude of this difference was negligible (as seen in the low effect size).

Table 9: Descriptive and inferential statistics for test-taker perception of potential to demonstrate English ability

Q: Did the test you took allow you to show your full English ability?								
Research study	**Test mode**	**Min.**	**Max.**	**Mean**	**SD**	**Z score**	**Asymp. Sig. (2-tailed)**	**Effect size**
B2	F2F	2	5	4.14	0.702	-2.024	0.043	-0.22
	VC	2	5	3.86	0.702			
A2/B1 for Schools	F2F	1	5	4.08	0.872	-2.754	0.006	-0.19
	VC	1	5	3.71	1.149			
1 = Not at all, 2 = Very little, 3 = Somewhat, 4 = Quite a lot, 5 = Very much								

In some cases, test-takers suggested the F2F mode was better because 'it gives the opportunity to have a closer experience with an English speaker' (B2 Candidate 3). Focus group responses included comments along the lines that F2F interaction 'works better' and that 'seeing someone face to face is easier to interact' (B2 Candidates 9 and 10, Centre B). These were reflected in negative appraisals of the VC experience, where test-takers commented 'it is more difficult to understand when to talk and when to stop' (B1 candidate, Italy); 'I didn't know when I had to speak' (A2 candidate, Italy). In focus groups, as in the surveys, inability to show English skills was linked to general test-taker anxiety. As one A2 test-taker remarked: 'my state of mind was pretty nervous, so I don't think I did my best.'

Test anxiety (test takers)

In general, test-taker responses suggest no huge uptick in anxiety in the VC mode (Figure 4). Interestingly, at lower levels, F2F was seen as the more anxiety-inducing mode – perhaps reflecting a more general tendency to feel nervous about exams at a younger age.

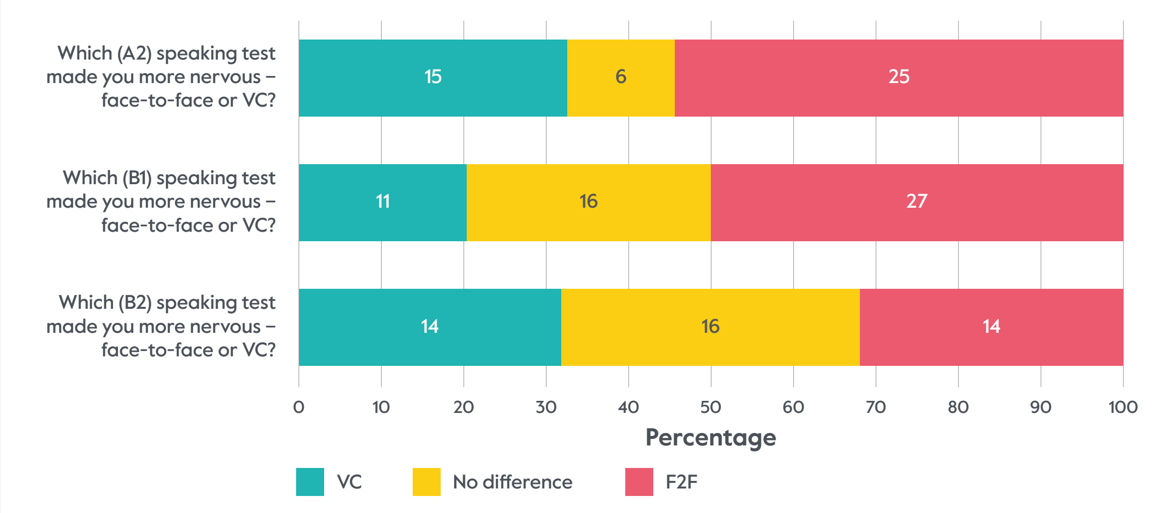


Figure 4: Test-taker perceptions of overall test anxiety for A2, B1 and B2 exams

Of those test-takers that indicated the VC test mode made them more nervous, common reasons were attending to the technical aspects of the VC test and worrying about their internet connection failing and/or the video and audio being affected:

> *'I think in VC you can get more nervous because you have to pay attention on what is happening with the video, the sound, your internet, even your voice. However, in face to face, all is clear and with less nerves than VC.'*
> (B2 Candidate 30)

> *'I was more nervous and when there were problems with the connection I got anxious and I couldn't show all my abilities.'*
> (B1 candidate, Italy)

Additionally, test-takers commented on how doing the test from home meant they were distracted by things happening at home or others disturbing the test. When it came to F2F test anxiety, some candidates reported being nervous in front of the examiner:

> *'When I face to the SE, I feel under pressure, and I can't talk all what I expect.'* (B1 candidate, Vietnam)

There was no consensus on which test mode was more anxiety-provoking, as test-takers' personalities and personal circumstances affected their ease with the two test modes. Importantly, there was no evidence to suggest VC test mode increased anxiety for all candidates. Examiner comments were similarly mixed – one examiner noted that in the VC mode the assessor was not visible, which could also reduce anxiety levels.

Mode preferences (test takers)

As can be seen in Figure 5, across all three exams there was a preference for the F2F mode. This initially appears at odds with the findings reported on anxiety but open responses from test-takers help unpack some of the nuances when it comes to overall preferences.

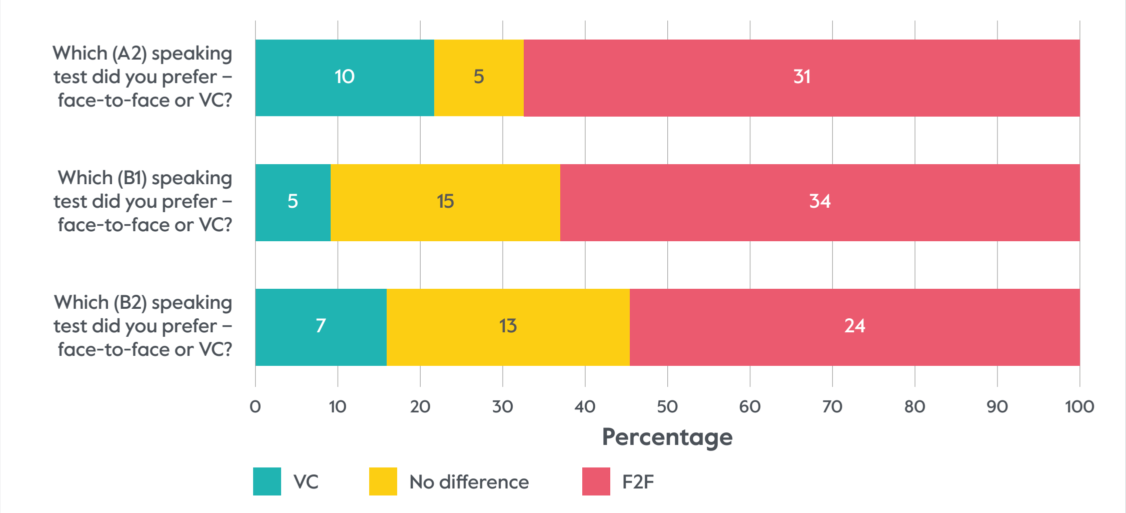


Figure 5: Test-taker preferences for overall test mode for A2, B1 and B2 exams

Responses ranged from those who saw little difference in the mode but just preferred the F2F mode. Others stated a preference for F2F because they had experienced technical problems online (even where the F2F was more anxiety-inducing). Similarly, while test-takers did sometimes find the F2F experience more stressful, they appreciated being able to see their partner better so they could 'co-ordinate better' (B1 candidate) and reported 'it's easier to interact with your partner' (B2 candidate).

Discussion and implications

The results of the B2 and A2/B1 studies chime with those of the existing literature in that while VC speaking can be generally conceptualised as a 'parallel' experience to F2F speaking (Nakatsuhara, Inoue, Berry and Galaczi 2017), it does present different challenges or concerns to stakeholders.

Encouragingly, the underlying issues of connectivity, audio and visual access were not major obstacles for most candidates in this study. However, a note of caution is required here since there was one centre that, much like prior studies (see Ockey et al 2019), was particularly affected by poor internet connectivity. This in turn led to perceptions of anxiety and difficulty in the VC mode. While VC can allow the 'practical advantage of connecting test takers and examiners who could be continents apart' (Galaczi and Taylor 2018:231) the inequality of internet access should be a major concern to test providers.

In terms of assessing VC performance, this study has highlighted that there are potentially factors at play in the online interaction that may make examiners less confident of ratings on specific criteria. This could range from poor audio leading to uncertainty over Pronunciation ratings to more general latency being felt to disrupt the fluidity of the interaction and ratings of Interactive Communication. This is an area for further research since these perceptions arguably relate to examiner and test-taker unfamiliarity with norms of online interaction (something we have all learned very rapidly as a result of the global Covid-19 pandemic). Examiner responses spoke to the need for detailed, practical training in VC testing should remote speaking be operationalised.

Mean values were lower in terms of examiner confidence in rating in the VC mode and test-takers also perceived that they were able to show their language ability to a lesser extent online than F2F (albeit with negligible effect sizes). Yet, the majority of test-takers did not perceive the VC test as being more difficult (Figure 3) nor was it seen to be more anxiety-inducing by a majority (Figure 4). Issues like anxiety were often highly individual, and to some extent, were conflated with test-takers' general test anxiety. It should also be remembered that these trials occurred during the Covid-19 pandemic, so it is unclear how much anxiety regarding F2F exams was due to feelings of being around others in a period of social distancing.

The majority of test-takers expressed an overall preference for the F2F test mode, due to the ease and clarity of interaction with the other speakers. This has obvious implications for test providers in terms of both being seen to ensure connectivity and construct relevance.

References

Craig, D A and Kim, J (2010) Anxiety and performance in videoconferenced and face-to-face oral interviews, *Multimedia-assisted Language Learning* 13 (3), 9–32.

Davis, L, Timpe-Laughlin, V, Gu, L and Ockey, G (2016) Face to face speaking assessment in the digital age: Interactive speaking tasks online, in Davis, J M, Norris, J M, Malone, M M, McKay, T H and Son, Y-A (Eds) *Useful Assessment and Evaluation in Language Education*, Georgetown University Round Table on Languages and Linguistics series, Washington, D C: Georgetown University Press, 115–130.

Galaczi, E and Taylor, L (2018) Interactional competence: Conceptualisations, operationalisations, and outstanding questions, *Language Assessment Quarterly* 15 (3), 219–236.

Jin, Y and Zhang, L (2016) The impact of test mode on the use of communication strategies in paired discussion, in Yu, G and Jin, Y (Eds) *Assessing Chinese Learners of English*, Basingstoke: Palgrave Macmillan, 61–84.

Kim, J and Craig, D A (2012) Performance and anxiety in videoconferencing, in Zhang, F (Ed) *Computer-Enhanced and Mobile-Assisted Language Learning: Emerging Issues and Trends*, Hershey: IGI Global, 137–157.

Lee, H, Patel, M, Lynch, J and Galaczi, E (2021) *Development of the IELTS Video Call Speaking Test: Phase 4 operational research trial and overall summary of a four-phase test development cycle*, IELTS Partnership Research Papers 2021/1, IELTS Partners: British Council/Cambridge Assessment English/IDP: IELTS Australia.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Ockey, G J, Gu, L and Keehner, M (2017) Web-based virtual environments for facilitating assessment of L2 oral communication ability, *Language Assessment Quarterly* 14 (4), 346–359.

Ockey, G J, Timpe-Laughlin, V, Davis, L and Gu, L (2019) Exploring the potential of a video-mediated interactive speaking assessment, *ETS Research Report Series* 2019 (1), 1–29.

Tsunemoto, A, Lindberg, R, Trofimovich, P and McDonough, K (2022) Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency, *Studies in Second Language Acquisition* 44, 659–684.

Video-call mode speaking tests: An examiner verbal protocol study

Susan Gilbert, Cambridge Examinations Basel GmbH, Switzerland
Lyn May, Queensland University of Technology, Australia

Introduction

Speaking tests which have traditionally only been offered in face-to-face (F2F) mode are now being offered in video-call (VC) mode. In research on the IELTS Speaking test, which involves a trained examiner who facilitates the test with a candidate, Nakatsuhara, Inoue, Berry and Galaczi (2017a, 2017b) found that scores received by the same candidate in the two modes were almost identical. However, there were differences in the language elicited from candidates who asked clarification questions more often in the VC mode. Examiners also reported difficulty in interpreting candidates' body language and experienced challenges in turn-taking as they were sometimes unsure of when a candidate had completed a turn and how to signal the initiation of a turn in the VC mode. The question of whether limited eye contact between the candidate and the examiner in VC mode could impact negatively on establishing rapport was also raised by Nakatsuhara et al (2017b).

Given the move towards VC speaking tests and the challenges experienced by raters of F2F paired speaking tests (May 2011), the aim of our paper is to explore the features to which assessors attend in VC mode, and their perceptions on assessing VC speaking tests. Verbal Protocol (VP) research has provided valuable insights into the process of assessing candidate performance in direct speaking tests (May 2011). As part of the larger research project reported in this issue, a VP 'think-aloud' study was carried out to investigate assessor behaviour in VC speaking tests, with the aim of learning more about how examiners respond to candidate performance in this mode. Insights gained into the assessment processes would inform both training programmes for Speaking Examiners (SEs) and test design and delivery.

Research questions

The study set out to collect and analyse data on examiner marking behaviour when assessing A2 and B1 paired speaking tests in VC mode to investigate the following research questions:

- How are the examiners applying the Assessment Scales in the VC mode?
- Which features of L2 speaking do raters notice and how does this relate to the VC mode?
- How do the examiners respond to features of online test design and delivery?

Methodology

Three experienced SEs took part in the study, two working regularly in F2F tests in Spain and one in Italy; they were familiar with VC speaking tests, having participated as interlocutors and assessors in the larger study which is the subject of this issue. They were asked to assess two paired speaking tests delivered in VC mode, one Cambridge A2 Key (henceforth A2) and one Cambridge B1 Preliminary (henceforth B1). The A2 candidates were L1 Spanish-speaking girls aged 11 and 12, while those in the B1 test were L1 Romanian-speaking female teenagers aged 13 and 16. In both cases, the test materials used were the 'for Schools' option, in which topics are appropriate to school-age candidates.

The format of both tests provides opportunities for different interaction patterns using both visual prompts and a script delivered by the interlocutor.

Data collection

The three examiners assessed the same tests; therefore, we collected six verbal reports, two transcribed paired speaking test performances and 12 assessments (three examiners each assessed four candidate performances).

While the examiners were assessing the candidates, they were instructed to stop the video recording at any point as often as they wanted, and to verbalise their thoughts at that moment. They used smartphones to audio-record their comments, leaving the audio recording running through each test; thus it was possible later to listen to the candidate performances alongside the examiners' verbalised thoughts. This enabled us to match their comments to candidate utterances and events in the tests.

The examiners participated in a preliminary training session on the aims and methodology of the study and to familiarize themselves with think-aloud verbal protocols. They practiced until they were comfortable with the procedure, then carried out the assessment tasks. This process resulted in the elicitation, segmenting and coding of a total of 448 think aloud comments: 182 for A2 and 266 for B1.

In subsequent 20-minute individual recall interviews, assessors watched short extracts from the tests they had assessed and expanded on the thoughts they had verbalised at those points in the tests. They were encouraged to make any other comments they felt were relevant. The interviewers asked the assessors to recall their thoughts; they avoided directing the assessors' focus or reacting to the assessors' comments.

Data coding and analysis

Green (1998:68) refers to the problematic nature of developing coding that captures the rich detail provided by the data but at the same time can reliably be used by independent coders, who 'may independently develop different schemes for the analysis of the same body of data'. To minimise this risk, the researchers read through all the data several times and defined a preliminary set of codes, which was based both on the Cambridge Assessment English Speaking Test Assessment Scales and sub-scales, and on online delivery. These codes were used independently by two researchers to code protocol data from one of the six sets collected. This task showed there was agreement between the researchers regarding the coding of comments relating to the Assessment Scales, but further refining of coding related to assessment processes was needed. More detailed coding relating to online delivery was also defined at this stage. Thus a final scheme of codes and subcodes was decided on. The remaining five tests were coded independently, one researcher taking three tests and the other two collaborating only where there were doubts. Finally each researcher read through the other's coding to ensure there was agreement throughout.

The coding scheme was divided into two areas, as below. See the Appendix the for full coding scheme and examples.

Assessment	**Technical aspects**
The assessment criteria (GV, DM, P, IC), and the sub-criteria.	Connectivity, depending at least partly on local conditions such as internet connection, and coded as T (technology, for example screen freeze or delay), or SQ (sound quality).
Marks and/or specific performance bands (M).	
The process of rating (Pro), meaning how the assessor is approaching the task, for example comparing a candidate's performance across different parts of the test.	Online delivery (OLD), referring to test design and delivery to the participants in VC format. Subcodes in this area: • management of the start of the test (Intro); • screen layout (Scrl); • transitions between parts of the test (Tr); • actions of the interlocutor (Itl).
Positive or negative assessment comments (Po/Neg).	
Other assessment comments (OAC), which did not fit the above codes.	

Two other codes were defined for comments which did not fit any of the above codes; ANAC (apparently non-assessment comments about the candidates), for example *'There seems to be a big age gap'* (Assessor A, B1 test), and OTH (Other comments): *'I don't like this question'* (Assessor B, B1 test).

The example below in Table 1 from the A2 test illustrates how each of the assessors' comments were segmented into references to different individual codes and how the coding scheme was applied. In all the extracts provided, the interlocutor is referred to as I, candidates as C1 and C2 and the three assessors as AA, AB and AC. Comments are unedited to maintain authenticity.

Table 1: Illustration of coding, A2 test (1)

Candidate utterance	Assessor comment	Coding explanation
Candidate 1 (C1): I don't like climb to the mountain because I think it's not funny and I don't like it.	**AA:** I don't know if … has Candidate 2 frozen on the screen? **(T/Neg)** We're 4 minutes into the test **(Pro)** and I'm just thinking these candidates are definitely in the 4+ range **(M)** so, I think I'll start focusing on the 5 scales in the analytical scales **(Pro)** and start looking back from there. **(Pro)**	**T/Neg:** Negative comment on technology: *AA makes a reference to screen freeze.* **Pro:** *AA is aware of the progress of the test and perhaps the need to think about marks.* **M:** *AA makes a specific reference to possible marks.* **Pro:** *AA comments on how they plan to refer to the Assessment Scales moving forward.*

A single comment could be coded more than once, as in T/Neg above, and where several sub-scales were mentioned in one phrase, they were coded additionally and separately, as below.

Table 2: Illustration of coding, A2 test (2)

Assessor comment	Coding explanation
Ok, develops her ideas with erm with no **hesitation** so there's **extended** stretches of language, erm, it's **relevant**.	**DM/Po:** the whole phrase is a positive reference to the DM scale. **Hes:** the assessor is thinking of the hesitation sub-scale. **Ext:** the assessor notices an extended utterance. **Rel:** the assessor judges the utterance relevant.

Results

From the total of 448 coded comments, 301 (67.19%) related to assessment and 98 (21.88%) to technical aspects, including connectivity, online design and delivery; OTH and ANAC made up the remaining 49 (10.94%) comments. This indicates that examiners' focus is principally on assessment, but that technical aspects were also noticed. The following findings result from detailed analysis of the 399 comments related to these two areas.

How are examiners applying the Assessment Scales in the VC mode?

Assessors refer to all the scales

Figure 1 demonstrates that the assessors referred to all the criteria; in both tests over 40% of assessment comments referenced GV. The apparent disparity in attention to IC can be explained by the fact that extent of utterances is a feature of DM at B1, while at A2 it is considered as part of IC. In both tests, there were considerably fewer comments on P than the other criteria.

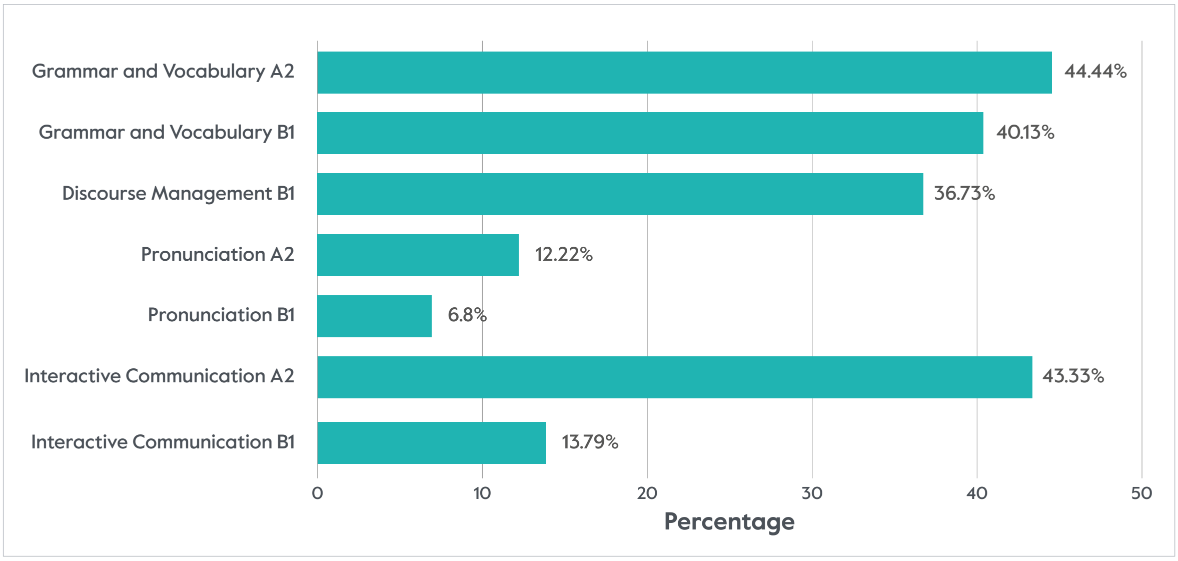


Figure I: Assessors reference to Assessment Scales across both tests

Assessors focus on marks and performance bands later in the tests

Almost all references to marks or performance descriptor bands came later in the tests. In the A2 test, seven of the eight references to marks came in the last part of the test; in the B2 test, 10 of the 12 references were in Parts 3 and 4.

Assessors' approach is positive

The assessors referred more to positive than negative aspects of candidate performance, with a greater difference in the B1 test, as shown in Table 3.

Table 3: Percentage of positive, negative and neutral comments

Assessment comments	% positive	% negative	% neutral
A2: 107 comments	40.19%	30.84%	28.97%
B1: 150 comments	53.9%	14.11%	31.99%

Assessment is cumulative in nature

Comments indicated that assessors were accumulating evidence throughout the tests:

> *'I'm starting to get an idea.'* (AA, A2 test)
>
> *'Would I take her up to a 5? Probably not yet.'* (AB, B1 test)

Assessors compared what they heard with previous parts of the test:

> *'She doesn't have quite as much control as I initially thought.'* (AA, A2 test)
>
> *'And she says "thoroughly interesting", I think, and before, in the previous part she said "really interesting" or "really exciting", so some good adverbs of degree as well from C2.'* (AC, B1 test)

One assessor expressed difficulty awarding marks when performance varies considerably across a test:

> *'She's a bit of a confusing candidate. Discourse Management is fantastic at times.'* (AA, B1 test)

Which features of L2 speaking were noticed by assessors? How does this relate to VC mode?

Assessors refer to all the assessment criteria in almost all test parts

In the A2 test (Figure 2), all the criteria were referred to throughout the test, except for the GV scale in Part 1, Phase 1. However, most of the comments in Part 1, Phase 2 related to GV, for example:

> C1: A sweater and a trainers.
>
> AA: *'A trainers.'* AB: *'A trainers. Vocabulary is good ... but the article.'*
> AC: *'A trainers you can't say but we understand.'*

As the A2 test progressed, assessors commented increasingly on IC, especially in the discussion task Part 2, Phase 2, where half the comments related to interaction, for example:

> C1: I don't like going to the forest because I can't see well the sky.
>
> C2: I like walk in the forest because I think you can find a lot of animals.
>
> AB: *'OK, they're taking it in turns to discuss.'*
>
> AC: *'Some good responses from the two of them, talking about the same picture before they move on to the next one.'*

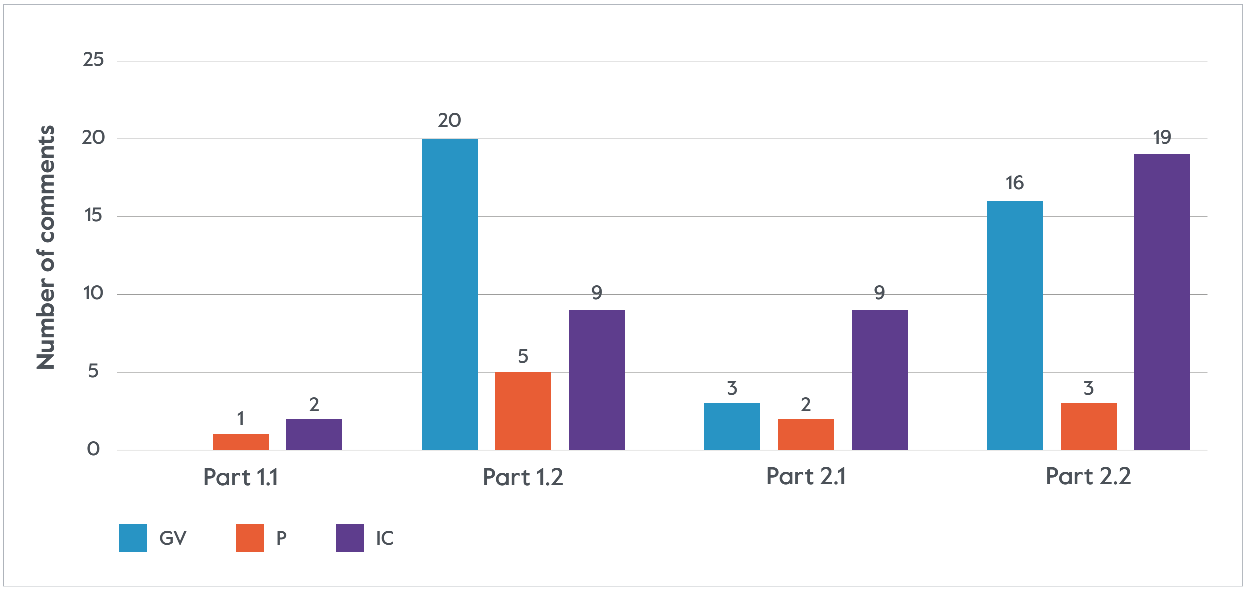


Figure 2: Assessors' focus on individual assessment criteria in A2 test parts

In the B1 test (Figure 3), assessors used all the criteria throughout the test, with the exception of the P criterion in Part 3. The total number of assessment comments was quite consistent in Parts 2, 3 and 4, although GV and DM were commented on much more than P and IC. As in the A2 test, IC was referred to more in the second half of the test (the discussion tasks) and less in Part 2 (the individual long turn).

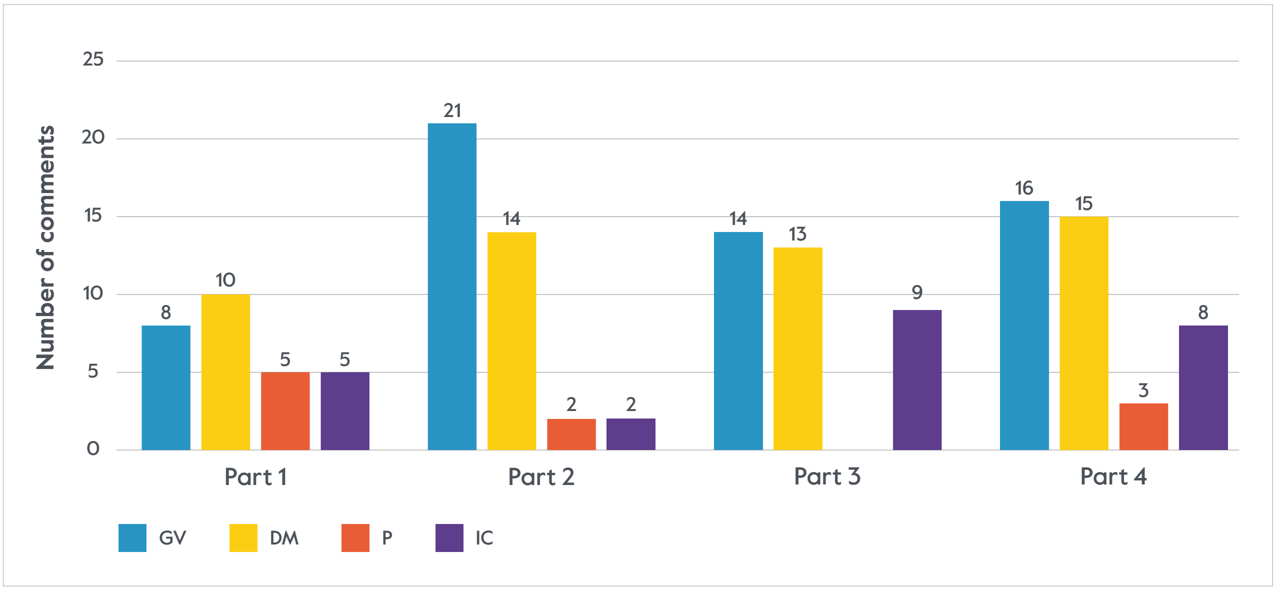


Figure 3: Assessors' focus on individual assessment criteria in B1 test parts

Assessors comment on some assessment sub-scales more than others

In both the A2 and the B1 tests (Figures 4 and 5), assessors focused mainly on control when thinking about grammar:

> *'Attempts to use present perfect but doesn't use "has just".'* (AB, B1 test)

However when assessing vocabulary they focused more on range:

> *'Gives example of different type of movies correctly.'* (AB, B1 test)

When assessing P, as instructed in their guidelines, assessors referred more to intelligibility than analysis of specific pronunciation features, thus across both tests most of the comments on P were general references (Figures 4 and 5):

> *'That's great P initially.'* (AA, A2 test)
> *'Always intelligible.'* (AB, B1 test)

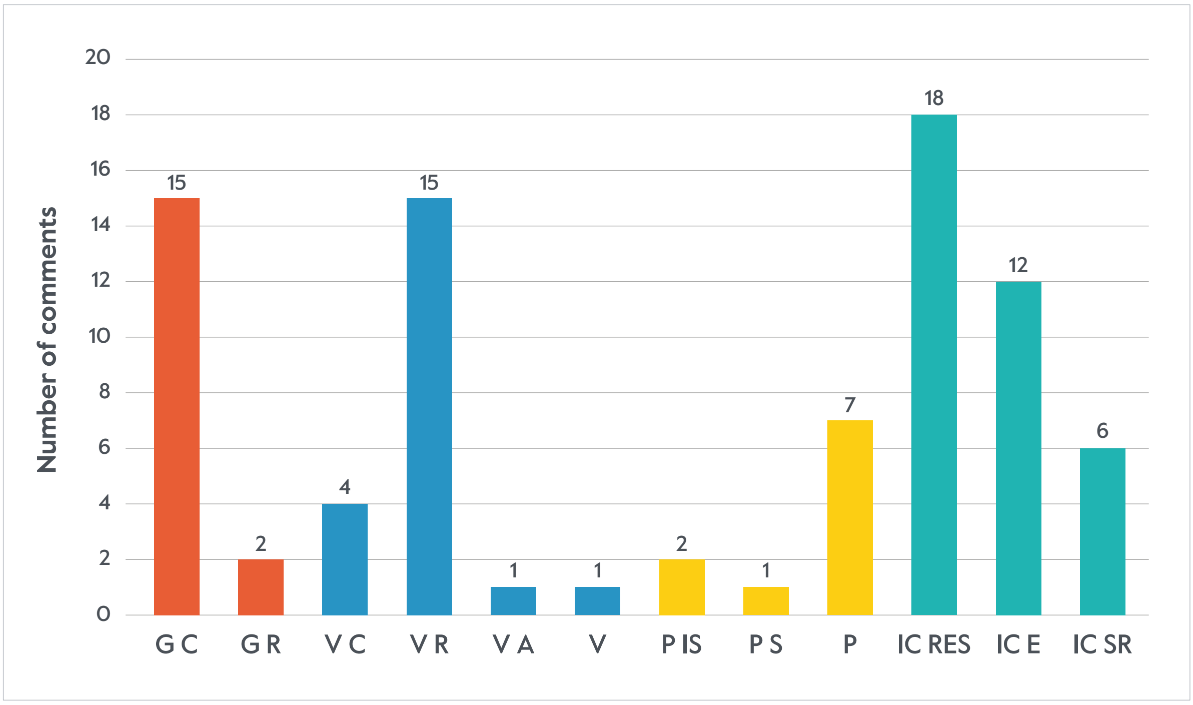


Figure 4: Use of assessment sub-scales in the A2 test

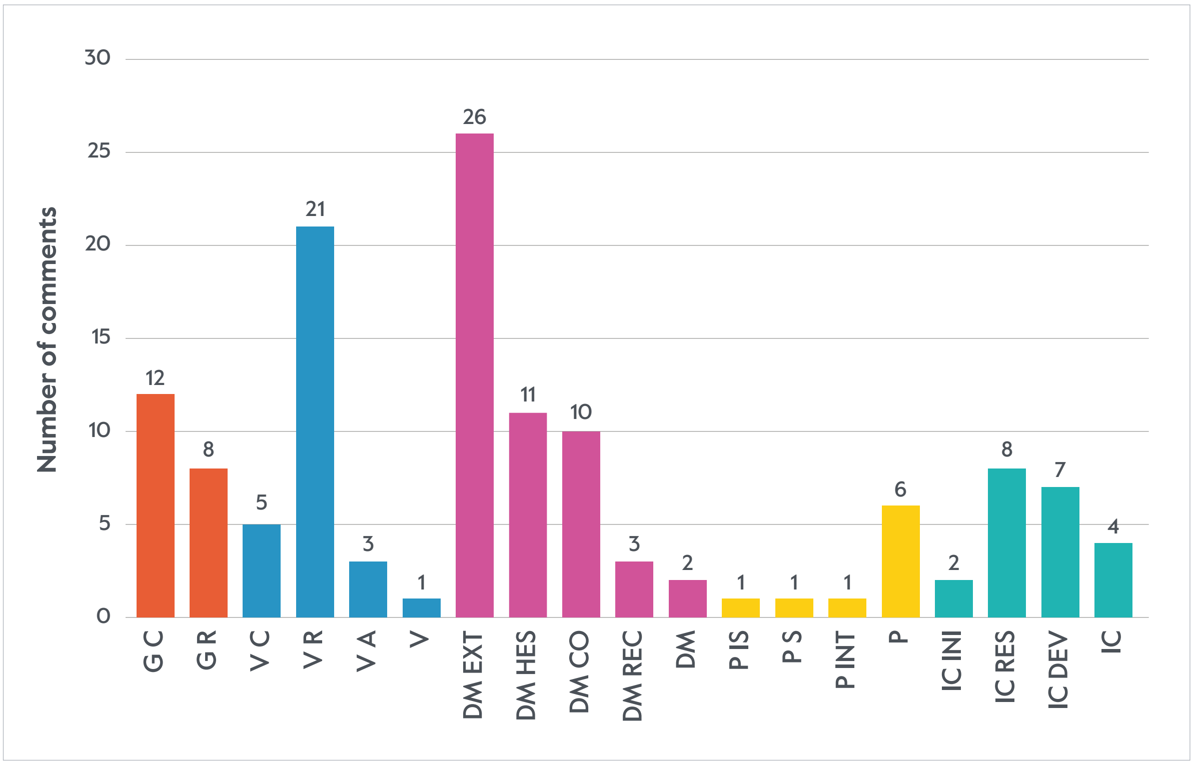


Figure 5: Use of assessment sub-scales in the B1 test

However, all but six of the comments on the other criteria referred to specific sub-scales.

Assessment of IC seems to be impacted in VC mode

Assessment comments indicated occasional doubt about whether the VC mode was impacting on the interaction in the discussion task: *'Is it their level or is it because they are remote? […]. Would it have been better face-to-face? Would they have said "what do you think? Do you like this too?"'* (AB, A2 test)

Comments elicited during the A2 test suggested that the interlocutor was using the 'Why' prompt where in a F2F test they might have used gesture or eye contact to encourage candidates to expand. The extract below shows how use of the 'Why' backup was interpreted differently by the assessors, who when awarding their mark for IC have to consider the amount of support candidates require.

Extract 1: A2 test, candidates-interlocutor discussion

Candidate-interlocutor interaction	Assessor comment	Notes
I: C2, do you think going for a walk in a forest is dangerous? **C2:** Sometimes but not always. **I:** Why? **C2:** Because I think if you are with family and friends, you are safe because they protect you.	**AB:** The interlocutor had to say 'Why?'. A stronger candidate would have said 'because' and carried on and extended. **AA:** Good, a really good extended answer, response, and with very little support as well. **AC:** I think [the interlocutor] is used to saying 'why?' to extend the questions now and the girls are used to waiting to hear the 'why' before they expand, so it's working well.	*AB and AC notice how much the interlocutor is using the 'Why?' backup question: AB views it as a sign of weaker performance, while AC has a positive comment.* *AA feels there was 'very little support', which at A2 refers to the highest performance band descriptor.*

In Extract 1, Assessor B later expanded on whether the interlocutor not being able to use gesture could have an impact on the interaction: *'If they're shyer, ... as interlocutor I think face-to-face I would be able to involve them more because I could hand signal without interrupting. I could try to get them both speaking.'* (AB, interview)

The balance of focus between assessment and online delivery changes across the tests

Figure 6 shows assessors' focus as the tests progressed; at the start of both tests they noticed aspects of online delivery and technology. However in both tests their focus clearly moves on to assessment in the second part, and then continues to be principally on assessment for the remainder of the tests. The exception is in the A2 test, Part 2, Phase 1, where all the assessors commented on an issue in the display of visual prompts.

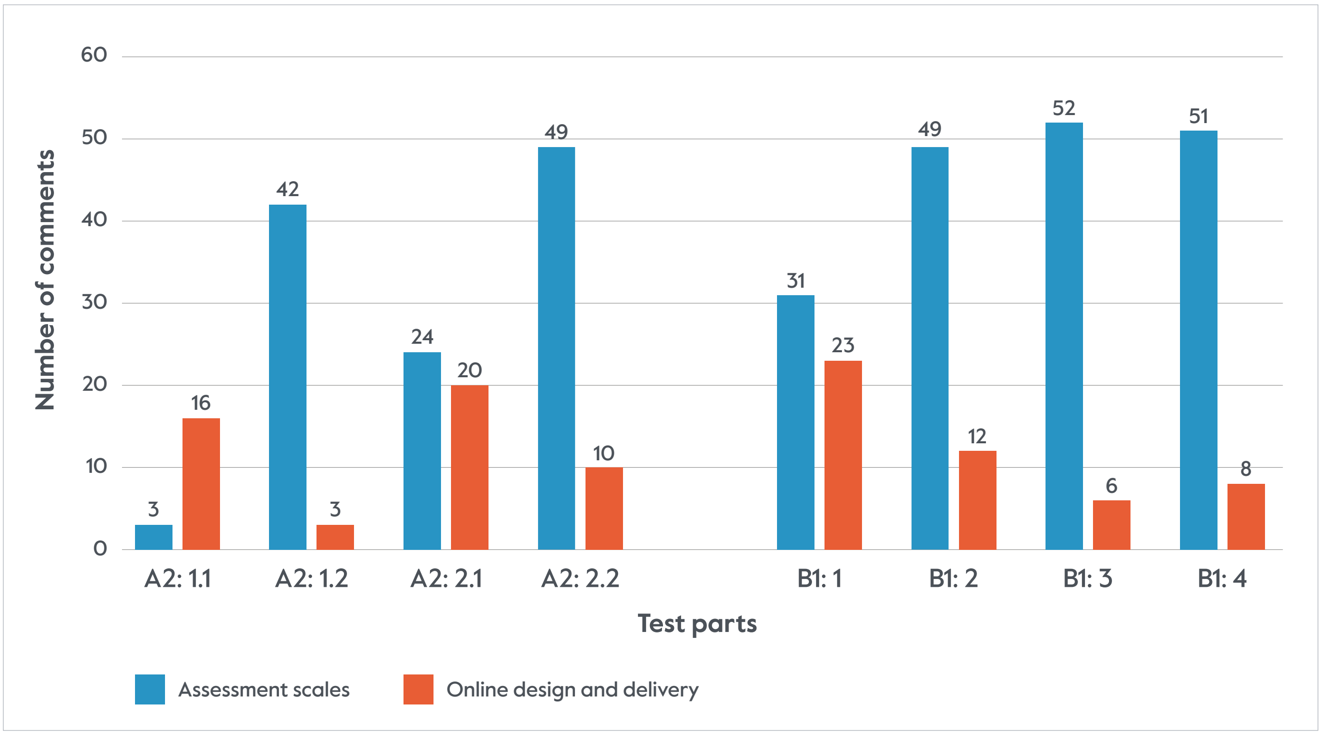


Figure 6: Assessors' focus on rating and online design/delivery in test parts

Assessors refocus quickly onto assessment

The assessors commented on design or delivery aspects, but focused quickly back onto assessment; Extract 2 demonstrates how Assessor A is thinking about grammar when the screen seems to freeze, and the interlocutor intervenes. After this, they quickly return to assessment of grammar, vocabulary and pronunciation.

Extract 2: A2 test candidate-candidate discussion task

Candidate-interlocutor interaction	Assessor comment	Notes
C1: I like walk in the forest because I think you can find a lot of animals.	**AA:** OK, excellently constructed sentence.	*AA comments on grammar/ vocabulary content.*
C2: I don't like climb to the mountain because I think it's not funny and I don't like it.	**AA:** I don't know, has Candidate 1 frozen on the screen? **AA:** We're 4 minutes into the test and I'm just thinking these candidates are definitely in the 4+ range so, I think I'll start focusing on the 5 scales in the analytical scales, and start looking back from there.	*At this point AA is distracted by technical issues.* *AA returns to focus on rating the candidates.*
C1: I don't like climbing the mountain because I think it's difficult.	**AA:** Yeah, she's definitely frozen.	*AA refers again to technical problems.*
I: OK. Candidate 1, do you think going for a walk on a beach is fun?	**AA:** Good idea to ask Candidate 1 the question to see if she's still got a connection.	*AA refers to interlocutor behaviour.*
C1: The picture number four?	**AA:** She's back!	*AA notices that the technical problem is resolved.*
C1: I like walk because I think it's funny and sometimes to see the camp or the sky or sometimes the clouds.	**AA:** Candidate 1's GV, she doesn't have quite as much control as I initially thought. P is definitely a 4.5/5 but maybe GV a bit less.	*AA orients back to assessment focus on GV and P.*

In Extract 3, Assessor B's focus is clearly on the grammatical content of Candidate 2's answer when an unexpected noise occurs. However, the next comment shows Assessor B's focus has returned to the assessment criteria, specifically extent, relevance, and grammatical content.

Extract 3: B1 test candidate-candidate discussion task

Candidate-interlocutor interaction	Assessor comment	Notes
C2: I think that the boy would enjoy watching some sports.	**AB:** Would enjoy.	*AB is focusing on the grammar content, noticing the conditional form.*
C2: Because it may be his favourite sport on the television, and I think he is going to enjoy it. And ...	**AB:** Who's making the noise?	*AB notices an unexpected noise on the audio.*
C2: He can read some comic books, some other books, that he likes. So, I think this is the two things that he can – he could do.	**AB:** OK, develops her ideas with no hesitation so there's extended stretches of language, it's relevant, and she's using a good degree of simple grammatical forms and there is some evidence of more complex.	*AB's focus returns to DM and GV.*

In the follow-up interview, Assessor B pondered why interference from extraneous noises could be more intrusive in an online environment:

> *'Referring to the noise, you get distractions obviously face-to-face, but possibly because we're using headphones, it's obviously a lot more exaggerated, so while you may be getting used to looking at the scales again and concentrating you suddenly look up, which I don't think I really do in a face-to-face test. I can block out any background noise much easier.'* (Assessor B, interview)

How do the examiners respond to features of online test design and delivery?

Uncertainty about candidate names distracts assessors' focus

Over the two tests, there were 83 comments relating to online design and the delivery of the test by the interlocutor, 56.6% concerning screen layout and the transitions between test parts.

The extract below illustrates how assessors were impacted by difficulty establishing candidate names and lack of a clear image of the candidate. They also noticed the lack of gesture available to the interlocutor to resolve the issue, and a potential security risk.

Extract 4: B1 test, start

Candidate-interlocutor interaction	Assessor comment	Notes
I: Good afternoon. I'm (interlocutor name) and this is (assessor name). Which one of you is (C1 name)? **C1:** I am. **I:** OK. And your name is? **C1:** My surname?	**AB:** I'd have preferred to have names on screens, so I know who the candidates are. **AB:** Is the assessor not going to disappear? I'm not sure if this works for introductions, because obviously they can't see who you're looking at.	*The interlocutor's second question (given in bold) is intended for the second candidate, but Candidate 1 thinks the interlocutor is still speaking to her.* *AB notes that the interlocutor cannot use eye contact or gesture to indicate who should answer, as they would in a F2F test.*
I: No (C2 name)? You are (C2 name), right? **C2:** Yes.	**AA:** OK, little bit of a confusing introduction. **AC:** A little confusing when she's asked one name, and then 'Your name is...?' Difficult to elicit who's who.	*The interlocutor is obliged to deviate from the script, to establish the candidates' names. AA and AC are now confused because it isn't immediately clear which candidate is which.*
I: OK, thank you. And how old are you, Candidate 1? **C1:** I'm 13 years old.	**AA:** OK, so C1 is on the top [of the screen]. Is C2 on the phone? Looks like she's on a phone device.	*AA establishes which candidate is which, but becomes distracted by a potential security issue related to a candidate's mobile phone.*
I: OK. And how old are you, C2? **C2:** I'm 16 years old. **I:** OK. Thank you. Where do you live?	**AA:** I can't really see C1, what's her name? It would be nice if we could see a full profile, really, just really see from her chin up. **AC:** We can still see the assessor in the bottom right-hand corner.	*AA returns to doubts about candidate names, and comments on how the candidates are shown on the screen.* *Meanwhile, AC notices that the assessor is also visible on the screen.*

Poor or variable audio quality is a source of tension for assessors

There were 15 comments on technology and connectivity, considerably fewer than on test design and delivery by the interlocutor. However, assessors' tension around possible technical issues such as screen freeze is clear in their comments:

> *'I don't know if it was a frozen frame or whether it was hesitation.'* (AB, A2 test)
>
> *'I'm sure there's a time delay, a lag.'* (AC, B1 test)
>
> *'It was a freeze and it was a bit of a panic, like "Oh dear, this could be a problem."'* (AA, interview)

All the assessors mentioned the importance of audio quality:

> *'It was a struggle. You're really trying hard to understand the words and to see whether she's answering the question and to see if it's fluent [...], but because of the sound it's hard to tune in.'* (AC, interview)

In one case, the rating changed when the audio quality improved:

> *'Once I was able to hear what she was saying it's like 'Oh!'. It was a change around in my opinions in that part.'* (AA, interview)

Difference in sound quality between the candidates impacted on the assessors:

> *'OK, there's quite a big difference in the sound quality between the two. Might have to take this into consideration.'* (AA, B1 test)
>
> *'Because the sound wasn't as good it was difficult to listen to, so it was hard to assess.'* (AC, interview)

The nature of online-delivered sound increases the cognitive load

In the Cambridge Instructions for Speaking Examiners, assessors are instructed to 'refer [to the Assessment Scales] constantly' while listening to the candidates. In the follow-up interviews, the assessors explained the difficulty caused by the fact that unlike assessing F2F speaking tests, listening to online-delivered sound meant that without looking at the candidates they were not always able to know who they were listening to; they needed to look at the screen to see who was speaking, and so could not focus entirely on the Assessment Scales. Assessor C explained:

> *'You'd think the sound would be better with the technology. You can control the volume, but because it's one-dimensional, I think it's a lot harder. The sound waves are completely different with the computer screen as to a room, you don't necessarily know who's speaking.'* (AC, interview)

Assessor B described the impact:

> *'I was looking at my scales [...], and I suddenly thought "who was that?" so I had to go back. [...] For a second I thought "have I got them wrong?" and then you suddenly panic. "Have I been getting them wrong from the start?" and it takes you a second to get your bearings back.'* (AB, interview)

Discussion

Analysis of the think-aloud data showed that the examiners assessed the candidates according to the Instructions guidance; they referred to all the Assessment Scales and sub-scales, there was evidence of the cumulative nature of assessment across the tests, with specific reference to marks coming later in the tests, and the assessment comments were more positive than negative, in line with the 'can do' approach to assessment exemplified by the wording of the Cambridge Assessment Scales. However, the assessment implications of how interlocutors use backup questions and verbal prompts as substitute for eye contact and gesture in VC format should be clarified for SEs.

Minor adjustments could be made to test delivery, including screen layout (so that assessors are always clear which candidate is which and they can easily see the visual prompts), interlocutor script (a clearer introduction, increased use of candidate names and the provision of suggested verbal prompts to use when necessary), and assessment protocols (so that assessors can refer constantly to the Assessment Scales as well as attending to the screen, and know how to assess IC in an online environment). These would reduce the impact of issues of online delivery noted in this study, as would more familiarity with the format. Issues concerning connectivity were infrequent but unpredictable and therefore harder to manage. Assessors' concentration was affected by poor audio quality, screen freeze or time lag, while the unpredictability of connectivity issues is in itself a distractor upfront in assessors' minds. Assessors and interlocutors will need to be provided with strategies and trained to deal with 'the unexpected'; if they know 'what to do if ...', they will feel more relaxed and able to concentrate fully.

The verbal protocol comments indicate that assessors' cognitive load is increased in VC format; Assessor C's comment in the interview, 'you can't relax for a moment', is telling. SEs will need to become used to working with the different nature of online-delivered audio input. Focusing on poor quality sound is exhausting, and a difference in sound quality between the candidates could lead to inequality of assessor's attention, resulting in candidates possibly being disadvantaged. Careful pre-checks of sound quality should avoid this.

We sum up by noting that assessors will need training and adequate practice to become familiar and confident with VC format, and thus manage the associated increase in cognitive load; as Assessor A said in the interview, 'definitely I think we need a lot more practice across the board so that it's going to be familiar, [...], but generally the remote works.'

References

Green, A (1998) *Verbal Protocol Analysis in Language Testing Research*, Studies in Language Testing Volume 5, Cambridge: UCLES/Cambridge University Press.

May, L (2011) *Interaction In a Paired Speaking Test: The Rater's Perspective*, Bern: Peter Lang.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017a) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017b) *Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and video-conferencing delivery (Phase 2)*, IELTS Partnership Research Papers 3, IELTS Partners: British Council/Cambridge English Language Assessment/IDP: IELTS Australia.

Taylor, L and Galaczi, E (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press, 171–233.

Appendix: Codes used in verbal protocol study

General categorisation (where applicable)	Coding and sub-coding	Definition/Gloss with examples
Grammar and Vocabulary	GV	Grammar and Vocabulary (e.g. ' he's just finished, present perfect, good', 'OK, blonde, brunette, good vocab')
	GC	Grammar Control (e.g. 'OK, so "there is /there are" control')
	GR	Grammar Range (e.g. 'OK she's shown a good degree of simple grammatical forms')
	VC	Vocabulary Control (e.g. 'funny, probably wanted to use "fun"')
	VR	Vocabulary Range (e.g. 'with a range of vocabulary')
	VA	Vocabulary Appropriacy (e.g. 'not sure about the use of "comfortable"? If it is appropriate but it helps me understand what she wants to say')
Discourse Management and subcodes	DM	Discourse Management (B1 only) (e.g. 'I definitely think that X's DM is good')
	Ext	Extent of utterance (e.g. 'OK, that was a really good extended response')
	Hes	Hesitation (e.g. 'OK; a bit of hesitation')
	Co	Coherence (e.g. 'extended stretch, using cohesive devices')
	Rel	Relevance (e.g. 'there's extended stretches of language, erm, it's relevant')
Pronunciation	P	Pronunciation (e.g. 'her P is good')
	INT	Intonation, (B1 only) ('e.g. 'so, good intonation')
	S	Stress (e.g. 'there is some word stress')
	IS	Individual Sounds (e.g. 'individual sounds are ok')

General categorisation (where applicable)	Coding and sub-coding	Definition/Gloss with examples
Interactive Communication and subcodes	IC	Interactive Communication (e.g. 'doesn't keep the IC going, really')
	RES	Responding (e.g. 'xx is not really responding to what she said')
	SR	Support Required (A2) (e.g. 'and with very little support as well')
	INI	Initiating (e.g. 'seeing if she can get a response from xx')
	DEV	Developing (B1 only)
	E	Extent of utterance (A2 only) (e.g. 'she only said yes, she hasn't continued')
Subcodes to assessment criteria	Po	'Positive' comments, which used positive adjectives or referred to use of more complex language (e.g. 'that was a really good extended response')
	Neg	'Negative' comments, noticing errors, what was lacking, or using negative adjectives (e.g. 'it's not with correct grammatical forms, she's using strange phrases')
PRO		Process of Assessment, how the assessor seems to be managing the task (e.g. 'so I think more or less I've come to a conclusion on all of them')
Marking	MGV	Marks/bands Grammar and Vocabulary (e.g. 'definitely, I'm looking at a 4/4.5')
	MDM	Marks/bands Discourse Management (B1) (e.g. 'I've gone in DM up to 5')
	MP	Marks/bands Pronunciation (e.g. 'P is definitely intelligible, ... so could be going up to a 4')
	MIC	Marks/bands Interactive Communication (e.g. 'IC's definitely going to be in a 4.5/5')
	M	Marks/bands, but not referring to a scale specifically (e.g. 'these candidates are definitely in the 4+ range')
OAC		Other Assessment Comments: comments on performance using language not in the scales (e.g. "she hasn't completed the task correctly, do I mark her down?")
ANAC		Apparently Non-assessment Comments (e.g. 'she's like, swinging in her chair')
Technical aspects	SQ	Sound Quality (e.g. 'the sound's really not very good at all with C2')
	VQ	Visual (screen) Quality
	T	Technology/connectivity (e.g. 'I'm sure there's a time delay, a lag')
Online delivery	OLD	Online Delivery (e.g. 'did someone just walk past?')
Subcodes to online delivery	OLD Itl	Interlocutor; interlocutor behaviour (e.g. 'nice prompting from interlocutor')
	OLD Intro	Introduction; comments at test start (e.g. 'difficult to elicit who's who')
	OLD Tr	Transition; comments on the transition between test parts (e.g. 'timing of the slides is perfect')
	OLD Scrl	Screen Layout (e.g. 'flowers are distracting')
OTH		Other Comments (e.g. 'Typical that it's his dad's car and not his mum's car. Maybe it is.')

Operational considerations for video-call delivery of paired format speaking tests

Andrew Mullooly, Capabilities, Cambridge University Press & Assessment

Introduction

The circumstances under which a test is delivered are a vital aspect of test validity. As observed in Weir's (2005) socio-cognitive framework, validity does not reside in a test *per se*, but in the scores on a specific administration of a test. The need to ensure that testing conditions remain the same for all test-takers becomes even more pertinent when organisations seek to provide greater access and choice in terms of where and how they can sit their test.

All forms of remote speaking assessment involve meeting certain technical and operational challenges. The nature and complexity of these will vary according to the type of test to be delivered. Aspects such as the number of test participants, whether visual material needs to be shared onscreen, and how high-stakes a test is will all play a part in shaping the challenges faced and most suitable solutions.

As noted elsewhere in this issue, much of the previous research in this area has been limited to single-candidate speaking tests. For example, prior to its operational release as an alternative to in-person (or face-to-face (F2F)) testing, several rounds of studies were conducted exploring validity aspects of a video-call (VC) IELTS Speaking test. One such early study cited examiners' verbal reports as highlighting 'negative effects of delayed video transmission', 'the way the test-taker can impact on the sound quality' and 'the need to control the direction of the interview' as three aspects impacting negatively on their role as facilitator (Nakatsuhara, Inoue, Berry and Galaczi 2017:13). For all three of these, the introduction of a second candidate, as is the case in Cambridge English Qualifications (CEQs) Speaking tests, could be viewed as adding to the complexity.

A more recent IELTS VC Speaking test study reported that examiners felt a brief linguistic turn before the test began was desirable to help build rapport with the test-taker (Lee, Patel, Lynch and Galaczi 2021). It is not just rapport between examiner and test-taker to be considered for paired format tests but also that between the two test-takers. One possible logistical arrangement would be to have test-takers remain co-present in a single location, sitting the test together at a centre, using a shared device to interact with a remote examiner or examiners. As well as alleviating concerns about test-takers not having the chance to interact with each other and the examiners before the start of a VC test, if the goal of remote speaking provision was to improve access to testing where enough examiners could not attend in person, such an approach would help fulfil this, whilst maintaining strict control over the test environment. This would be particularly relevant for high-stakes assessments. However, if visual material needs to be shared onscreen, as happens during both individual long-turns and two-way collaborative tasks in CEQs Speaking tests, such a set-up would require careful positioning of the shared device, so that both test-takers could see clearly, while also ensuring consistent audio and video quality throughout for examiner marking.

For the purposes of this exploratory research, it was decided to focus on a set-up that saw all four participants in different locations. As the scenario furthest removed from the current F2F delivery model, this was viewed as having the strongest requirement for evidence to support comparability of modes. Whilst formal research questions were not defined prior to commencing trialling, the guiding aims could be summarised as follows:

- To evaluate how effectively modifications to existing F2F test materials and examiner frames (see the section 'Procedure') supported the new VC delivery mode.
- To identify the most common issues experienced by test-takers and examiners during VC delivery, and assess severity of impact on the test experience.
- Through observation of examiner behaviours in responding to any issues, to identify what additional guidance or training could be provided to support VC delivery.

Procedure

During the trials, most candidates sat their VC tests at home on their own devices. Where test-takers did come into a centre to sit the test, they did so in separate rooms. Examiners were also either present in a different room at the test centre or at their own homes.

Following consultation with centres participating in the trial, Zoom was selected as an off-the-shelf VC platform with which there was good pre-existing familiarity for centres, examiners and test-takers – albeit typically in a remote teaching rather than assessment context. All participants were given prior written instructions on how to download the software if needed, check their equipment and amend the screen settings to help address the risk of participants' videos partially obscuring onscreen test materials. Headsets were not compulsory but were recommended for use by all examiners to allow for ease of marking.

Although it was important to make sure test-takers had a positive overall experience, it was also desirable not to control conditions to such an extent that none of the potential technical issues needing to be mitigated against and responded to as part of 'real-world' delivery were observable. The guidance provided to volunteer test-takers and examiners therefore did not include minimum hardware specifications or internet performance levels. This consideration is of course different in operational conditions, where hardware and internet specifications are stipulated.

As all test-taker participants were studying towards taking their actual exams (except for two pseudo test-takers from centres where an uneven number of volunteers meant that one participant would otherwise have been left without a test partner), a high level of familiarity with the relevant test format could be assumed.

Test materials were modified to be made more suitable for onscreen sharing. Such adaptations included the insertion of 'holding' pages between tasks so that examiners were less likely to accidentally share any test content before they intended to do so. Examiners were given the option either to read from a digital copy of their script, preferably on a second screen, or print off a hard copy that they could refer to.

Changes to the standard test frame were kept to the minimum deemed necessary to accommodate the VC mode. A brief functionality check was added whenever materials were shared onscreen, with the examiner first asking test-takers to confirm that they could see the visuals before proceeding to begin the task. Accompanying minor changes to wording were made. For example, for B1 Preliminary for Schools (henceforth B1 for Schools) Part 2, the F2F line 'I'm going to give each of you a photograph' was replaced with 'I'm going to show each of you a photograph' for the VC mode. Such changes were in keeping with the approach adopted by IELTS VC speaking studies, which made similar modifications to accommodate the new test delivery medium, for instance at the point where a prompt card appeared onscreen rather than being handed over by the examiner as in the F2F test (Lee et al 2021).

At lower levels, the paired format meant that a revised wording for Part 1 was introduced. The standard F2F Interlocutor Frame for A2 Key for Schools (henceforth A2 for Schools) and B1 for Schools includes the question 'What's your name?' with the examiner asking this of both test-takers in turn. This was changed to more of a confirmatory check for the VC mode, with the question 'And which one of you is (*Candidate A name*)?' followed by 'And your name is (*Candidate B name*)?' This had the advantage of avoiding relying on eye contact to make clear who was being addressed. Examiners were advised to make use of rising intonation to help indicate that Candidate B was intended to respond.

Though examiners received online training in how to deliver these revised script lines, the changes were not considered such that test-takers needed to be informed of them prior to the trial test taking place.

Examiners acting in the role of assessor were instructed to turn off their camera and mute their microphone after being introduced by their colleague towards the start of the test. As assessors do not directly interact with test-takers during the test, it was felt that this would allow them to concentrate on marking, while also reducing the number of visible participants onscreen.

Examiners were not given step-by-step instructions as to how to deal with every technical issue that could arise. They were told that they should keep to the Interlocutor Frame as much as possible but feel free to deviate from it, if, in their judgement, the situation so required. The need to prioritise putting test-takers at ease should any issues arise was stressed.

Findings and recommendations from the B2 First (henceforth B2) trial were incorporated into the planning stage for the later A2/B1 for Schools study. This included a clearer distinction between the role of examiner and centre support staff. For both trials, it was specified that someone at the centre other than the examiners was required to act as an online usher, admitting and welcoming test-takers to the speaking test 'room'. For the latter trial however, this role included carrying out a more standardised pre-test equipment and functionality check. The wording of this was based on suggestions from examiners who had taken part in the B2 trial. The intention behind this change was to minimise the risk of in-test issues occurring and to allow examiners to focus their attention more fully on test delivery.

One in-test change made for the later trial was when examiners acting in the role of interlocutor switched to a document-sharing view. For the B2 trial, examiners were instructed to start the test with a 'holding' page (a Cambridge English logo) displayed onscreen throughout Part 1. Even though no materials are visible during this part of the test, the rationale was that clicking to the next page at the start of Part 2's photograph-based individual long-turn, as opposed to sharing for the first-time mid-test, would place fewer cognitive demands on the examiner and remove the risk of a noticeable in-test pause. Examiners were to stop sharing before the start of Part 4, in which no visual or written prompts are used, as it was felt that devoting more screen space to the participants' videos could help facilitate the three-way interaction between examiner and candidates in this part of the test. However, on reviewing the recordings from the trial, it was observed that this sudden switch in view appeared disconcerting for some test-takers. For the A2/B1 for Schools trial therefore, document-sharing was only used during those parts of the tests where candidates respond to visual prompts. The benefit of consistency in adopting the larger video view whenever materials were not needed onscreen was seen as outweighing the increased risk of a slight mid-test pause.

Analysis

For the B2 study, transcripts and video recordings from all 28 VC tests without pseudo candidates were checked for any technical or test delivery issues arising during or immediately prior to the test. Rather than working to a pre-determined list of categories, these were generated by looking across the full range of issues captured.

Following the A2 /B1 for Schools trial, transcripts of all VC recordings were checked for any issues that might have negatively impacted on test-taker experience or examiners' ability to assess test-taker performance. Where possible issues were identified, the relevant recordings were reviewed to allow for a more detailed analysis.

Given the differences in approach described above between the two trials, a quantitative comparison of issues across the three proficiency levels was not considered appropriate. Instead, this paper will highlight the range and severity of issues encountered, as exemplified in selected excerpts from the test transcripts.

Findings

Broadly speaking, issues could be categorised according to whether they were primarily technical in nature or related more to aspects of the examiner's test delivery. In both instances, low severity issues, resolved within a single conversational turn with minimal, if any, impact on test-taker experience or examiner ability to assess performance, were most frequent, as in the following example from B1 for Schools.

Extract B1.1: Centre MX1 – VC mode

1 EX Thank you. (Candidate 5) how often do you use a mobile phone
2 and why?
3 C5 Can you repeat it please?
4→ EX **How often do you use a mobile phone?**
5 C5 Sorry (.) I have problems with my internet (.)
6 I can't understand you.
7→ EX **Do you often use a mobile phone?**
8 C5 Ah (.) yes.

After first repeating the question once (Line 4), the examiner asks the scripted back-up prompt (Line 7). Similar requests for repetition also occurred during F2F test delivery, with examiners free to repeat lines from the Interlocutor Frame or to utilise the supporting back-up prompts as required.

Extract B1.2: Centre ES2 – F2F mode

1 EX Thank you. And (.)(Candidate 2) (.) is it important for
2 everybody to learn to swim?
3 C2 I don't hear
4→ EX **Is it important for everybody to learn to swim?**
5 C2 I think yes because it's fun to walk in a beach or to
6 a swimming pool ...

The fact that most test-takers were taking their VC tests from home meant that when connectivity issues did occur, there was a lack of on-site support to assist. As test security was not a primary focus of the research, test-takers were not asked to complete a pre-test room sweep or given any instructions to restrict their movements throughout the test. In this following example from A2 for Schools, a candidate sought to resolve a poor internet connection by moving to a different part of the house. As can be seen, the examiner felt it necessary to deviate from the frame (Lines 4, 6, 8, 11 and 12), to give advice to try and improve the quality of the audio, and then to briefly pause the test.

Extract A2.1: Centre MX1 – VC mode

1	EX	(Candidate 4) (.) what clothes do you like to wear when you are
2		at home? (Candidate 4)?
3	C4	Yes (.) can you repeat?
4→	EX	**Open your microphone.**
5	C4	Can you repeat the question?
6→	EX	**Sure (.) open your microphone too.**
7		what clothes do you like to wear when you are at home?
8→		**(Candidate 4), are you having problems with the internet?**
9	C4	Yes (.) I (.) having problems but I move to another part
10		of the house.
11→	EX	**Are you going to change to another part of the house now?**
12→		**Let's wait for (Candidate 4) (.) OK? OK.**
13		(Candidate 4) (.) what clothes do you like to wear when
14		you are at home?
15	C4	I like the shorts and pants and t-shirts.

At the extreme end of the severity scale, during the B2 trial, one test-taker lost connection completely and had to re-join the call. Since this happened during a part of the test (Part 2) where there is no interaction between the test-takers, it was relatively straightforward for the examiner to pick up again from the start of the individual's long-turn.

Extract B2.1: Centre B – VC mode

1	EX	Hi (Candidate 7) (.) are you back?
2	C7	Yes.
3→	EX	**We can start from Part 2 (.) we will go back to your picture.**
4→		**This was your picture at the beginning, right?**
5	C7	What?
6→	EX	**This was your picture? OK (.) I will say it again.**
7		Here are your photographs (.) they show children doing
8		different things on the school trip.
9		Can you see the photographs?
10	C7	Yes.

During an A2 for Schools test, the examiner acting as interlocutor dropped off the call mid-test. In this case, the paired-examiner format allowed for the assessor (AS) to quickly step in and reassure the test-takers until it was possible to continue (Lines 8–10).

Extract A2.2: Centre IT1 – VC mode

1	EX	(Candidate 8) (.) when was the last time you went to the beach?
2	C8	I think last summer (.) I was with four friends (.) if I
3		remember well (.) and were to [unintelligible 00:09:58].
4		If I remember well (.) this is a- these are a beautiful day
5		(.) and after the beach (.) with my friends (.) we went to
6		an house of one of my friends and we had lunch
7		*[Lost connection with interlocutor]*
8→	AS	**OK (.) (interlocutor) had a problem with connection and she**
9→		**is coming back (.) just one moment more (.) Thank you for**
10→		**waiting (.) it won't be much longer (.) I promise.**
11	EX	OK. Hello (.) I'm back (.) Sorry (.) Can you all hear me?
12	C8	Yes.
13	C7	Yes.
14	EX	OK (.) We will continue. (Candidate 7) (.) when was the
15		last time you went to the beach?

Though advised to use a laptop or similar device, it could be observed from the recordings that some test-takers chose to access the test via a mobile phone. This was not seen to have impacted significantly on their test experience.

As assessors are not directly involved in the interaction, they generally would not be a visible presence during the test past the point at which they had been introduced by the interlocutor. On a couple of occasions, an assessor forgot to switch off their video until mid-way through Part 1, but this had no notable impact on candidate performance.

As anticipated, the point where the interlocutor shared test materials onscreen was found to be a juncture at which a high proportion of both technical and delivery-related issues occurred. Technical issues observed included material not displaying as expected or the video display obstructing task content. When test-takers reported problems in seeing the visuals, the issue would often swiftly resolve itself, without need for any corrective action by either the examiner or test-takers, as in the below example from B1 for Schools (Lines 5–7).

Extract B1.3: Centre MX1 – VC mode

1	EX	Thank you (.) now in this part of the test you're going
2		to talk about something together for a minute (.)
3		for two minutes >sorry< can you see the pictures?
4	C1	Yes.
5→	EX	**(Candidate 2) (.) no you can't?**
6→	C2	*[Candidate 2 first shakes head then nods]* **Yes now I can.**
7→	EX	**Now you can (.) OK.**

In the case of the video display obscuring test content, either the examiner was able to deviate from the Interlocutor Frame to provide support to resolve the issue, or the test-taker ultimately resolved the problem themselves, as in the following example from B2.

Extract B2.2: Centre ES1 – VC mode

1 EX OK (.) thank you. Now I'd like you to talk about something
2 together for about two minutes. I'd like you to imagine that
3 a country wants to encourage its people to do more sports.
4 Here are some things that they are thinking about and a
5 question for you to discuss. Can you see the task?
6 C14 Yes
7→ C13 **I am going to move me a little bit, yeah** *[laughs]*
8 EX Can you see the task?
9 C14 Yes.
10 C13 Yes.
11 EX OK (.) First you have some time to look at the task.
12→ C13 **OK (.) Oh (.) I don't know where put me.**
13→ EX **OK?**
14→ C13 **Yeah.**

This was not found to be a problem in the A2 for Schools and B1 for Schools trial, something that was probably at least partly attributable to the more standardised pre-test checks but could also have been a result of increased test-taker familiarity with the VC software over the course of the pandemic.

Test material issues related to aspects of the examiners' test delivery included the wrong content being shown. It should be noted that when this was the case, it was generally only very briefly, whilst the examiner was navigating between tasks.

In terms of the changes made to the Interlocutor Frame, when examiners followed the revised A2 for Schools and B1 for Schools Part 1 wording as intended, it generally worked well. However, there were several occasions of interlocutors deviating from it, occasionally leading to avoidable confusion for test-takers.

In the below example, the fact that the examiner failed to use Candidate 6's first name in Line 5 meant that Candidate 1 believed this to be a follow-up question and still directed at them (Line 6).

Extract B1.4: Centre RO – VC mode

1 EX Good afternoon (.) I'm (interlocutor full name) (.)
2 and this is (assessor full name).
3 Which one of you is (Candidate 1 first name)?
4 C1 I am (Candidate 1 first name).
5→ EX **OK. And your name is?**
6→ C1 **My surname?**
7 EX No (Candidate 6 first name)?
8 You are (Candidate 6 first name), right?
9 C6 Yes.

Conclusions

This research strand had three guiding aims: firstly, to evaluate modifications made to test materials for VC delivery; secondly, to identify the different types and severity of in-test issues encountered; and finally, when issues did occur, through reflection on how examiners responded to them, to help identify what additional guidance or training could support VC delivery.

The minor modifications made to test materials were found to be generally both effective and sufficient. This helps support the assertion that equivalent F2F and VC modes of paired speaking tests can be offered, without the need for major revisions to either test content or examiner frames. The most common issue observed with the revised materials, whereby examiners did not always keep to the wording for the name check at the start of A2 for Schools and B1 for Schools tests, should be largely avoidable through clearer written instructions and formatting.

Whilst there were relatively few issues with Interlocutor Frame delivery, the omissions and slips that did occur could perhaps be taken as indicative of increased cognitive demands being placed on interlocutors in terms of managing both the technology and the interaction at the same time. This was particularly evident in some of the first tests that examiners delivered online. This should not be taken as suggesting that remote delivery is intrinsically more challenging than F2F, rather that it brings with it a different set of challenges that examiners must adapt to.

Though in-test technical issues were often encountered at the point at which materials were shared onscreen, they could occur at any point in the test. Realistically, however extensive the examiner training, test-taker guidance and pre-test preparations, any test provider considering a VC speaking solution should do so with the recognition that technical issues will still happen. Alongside minimising their frequency, priority should be placed on mapping out measures to be taken in each different scenario, so that examiners are well-prepared and do not need to improvise a response. This is particularly important in view of the need for uniformity of administration as highlighted by Weir (2005).

Depending on the test stakes, to avoid compromising test security in an operational context, a complete loss of connection by either test-taker or examiner may well be sufficient grounds for the test to be stopped and retaken at a later date. If tests were allowed to continue, examiners would need clear guidance as to how far into a task it would be permissible to re-start. Were the test-taker more than halfway through an individual long-turn, for instance, a reduced speaking window or opportunity to complete a replacement contingency task upon restart may be more appropriate.

As well as potential breaches to test security, another threat to scoring validity could be posed through poor or variable conditions for rating (Taylor and Galaczi 2011). Issues should be evaluated on the extent to which they interfere with the examiners' ability to accurately assess performance. For example, with hesitation a relevant performance feature under the Discourse Management assessment scale, if connectivity was found to be a recurring issue throughout a VC test, it would be important for examiners to be able to confidently and correctly distinguish this from test-takers repeatedly needing prolonged time to formulate their responses.

The paired format of CEQs means that a technical problem for one test-taker has the potential to impact on the performance of their partner. Guidance should be made available for test-takers, so that they are informed about what will happen in different scenarios and can be confident that their result will not be affected.

References

Lee, H, Patel, M, Lynch, J and Galaczi, E (2021) *Development of the IELTS Video Call Speaking Test: Phase 4 operational research trial and overall summary of a four-phase test development cycle*, IELTS Partnership Research Papers 2021/1, IELTS Partners: British Council/Cambridge Assessment English/IDP: IELTS Australia.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Taylor, L and Galaczi, E (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press, 171–233.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

Where your world grows

Find out more at
cambridge.org/english

Shaftesbury Road, Cambridge, CB2 8EA