ISSN 2515-1703
2016

IELTS Partnership Research Papers

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

Fumiyo Nakatsuhara, Chihiro Inoue, Vivien Berry and Evelina Galaczi

IELTS™
BRITISH COUNCIL
idp IELTS AUSTRALIA
CAMBRIDGE ENGLISH Language Assessment
Part of the University of Cambridge

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

This paper presents the results of a preliminary exploration and comparison of test-taker and examiner behaviour across two different delivery modes for an IELTS Speaking test: the standard face-to-face test administration, and test administration using Internet-based video-conferencing technology.

Funding

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

Acknowledgements

The authors gratefully gratefully acknowledge the participation of Dr Lynda Taylor for the design of both Examiner and Test-taker Questionnaires, and Jamie Dunlea for the FACETS analysis of the score data; their input was very valuable in carrying out this research. Special thanks go to Jermaine Prince for his technical support, careful observations and professional feedback; this study would not have been possible without his expertise.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2016.

This publication is copyright. No commercial re-use. The research and opinions expressed are of individual researchers and do not represent the views of IELTS. The publishers do not accept responsibility for any of the claims made in the research.

How to cite this paper

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, Cambridge English Language Assessment and IDP: IELTS Australia. Available at https://www.ielts.org/teaching-and-research/research-reports

Introduction

The IELTS partners – British Council, Cambridge English Language Assessment, and IDP: IELTS Australia – are pleased to introduce a new series called the *IELTS Partnership Research Papers*.

The IELTS test is supported by a comprehensive program of research, with different groups of people carrying out the studies depending on the type of research involved. Some of that research relates to the operational running of the test and is conducted by the in-house research team at Cambridge English Language Assessment, the IELTS partner responsible for the ongoing development, production and validation of the test. Other research is best carried out by those in the field, for example, those who are best able to relate the use of IELTS in particular contexts.

With this in mind, the IELTS partners sponsor the IELTS Joint Funded Research Program, where research on topics of interest are independently conducted by researchers unaffiliated with IELTS. Outputs from this program are externally peer reviewed and published in the *IELTS Research Reports*, which first came out in 1998. It has reported on more than 100 research studies to date — with the number growing every few months.

In addition to 'internal' and 'external' research, there is a wide spectrum of other IELTS research: internally conducted research for external consumption; external research that is internally commissioned; and, indeed, research involving collaboration between internal and external researchers.

Some of this research will now be published periodically in the *IELTS Partnership Research Papers*, so that relevant work on emergent and practical issues in language testing might be shared with a broader audience.

We hope you find the studies in this series interesting and useful.

About this report

The first report in the *IELTS Partnership Research Papers* series provides a good example of the collaborative research in which the IELTS partners engage and which is overseen by the IELTS Joint Research Committee. The research committee asked Fumiyo Nakatsuhara, Chihiro Inoue (University of Bedfordshire), Vivien Berry (British Council) and Evelina Galaczi (Cambridge English Language Assessment) to investigate how candidate and examiner behaviour in an oral interview test event might be affected by its mode of delivery – face-to-face and internet video-conferencing. The resulting study makes an important contribution to the broader language testing world for two main reasons.

First, the study helps illuminate the underlying construct being addressed. It is important that test tasks are built on clearly described specifications. This specification represents the developer's interpretation of the underlying ability model – in other words, of the construct to be tested. We would therefore expect that a candidate would respond to a test task in a very similar way in terms of language produced, irrespective of examiner or mode of delivery.

If different delivery modes result in significant differences in the language a candidate produces, it can be deduced that the delivery mode is affecting behaviour. That is, mode of delivery is introducing construct-irrelevant variance into the test. Similarly, it is important to know whether examiners behave in the same way in the two modes of delivery or whether there are systematic differences in their behaviour in each. Such differences might relate, for example, to their language use (e.g. how and what type of questions they ask) or to their non-verbal communication (use of gestures, body language, eye contact, etc.).

Second, this study is important because it also looks at the ultimate outcome of task performance, namely, the scores awarded. From the candidates' perspective, the bottom line is their score or grade, and so it is vitally important to reassure them, and other key stakeholders, that the scoring system works in the same way, irrespective of mode of delivery.

The current study is significant as it addresses in an original way the effect of delivery mode (face-to-face and tablet computer) on the underlying construct, as reflected in test-taker and examiner performance on a well-established task type.

The fact that this is a research 'first' is itself of importance as it opens up a whole new avenue of research for those interested in language testing and assessment by addressing a subject of growing importance. The use of technology in language testing has been rightly criticised for holding back true innovation – the focus has too often been on the technology, while using out-dated test tasks and question types with no understanding of how these, in fact, severely limit the constructs we are testing.

This study's findings suggest that it may now be appropriate to move forward in using tablet computers to deliver speaking tests as an alternative to the traditional face-to-face mode with a candidate and an examiner in the same room. Current limitations due to circumstances such as geographical remoteness, conflict, or a lack of locally available accredited examiners can be overcome to offer candidates worldwide access to opportunities previously unavailable to them.

In conclusion, this first study in the IELTS Partnership Research Papers series offers a potentially radical departure from traditional face-to-face speaking tests and suggests that we could be on the verge of a truly forward-looking approach to the assessment of speaking in a high-stakes testing environment.

On behalf of the Joint Research Committee of the IELTS partners

Barry O'Sullivan, British Council
Gad Lim, Cambridge English Language Assessment
Jenny Osborne, IDP: IELTS Australia

October 2015

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

Abstract

This report presents the results of a preliminary exploration and comparison of test-taker and examiner behaviour across two different delivery modes for an IELTS Speaking test: the standard face-to-face test administration, and test administration using Internet-based video-conferencing technology. The study sought to compare performance features across these two delivery modes with regard to two key areas:

- an analysis of *test-takers' scores* and *linguistic output* on the two modes and *their perceptions* of the two modes
- an analysis of *examiners' test management and rating behaviours* across the two modes, including *their perceptions* of the two conditions for delivering the speaking test.

Data were collected from 32 test-takers who took two standardised IELTS Speaking tests under face-to-face and internet-based video-conferencing conditions. Four trained examiners also participated in this study. The convergent parallel mixed methods research design included an analysis of interviews with test-takers, as well as their linguistic output (especially types of language functions) and rating scores awarded under the two conditions. Examiners provided written comments justifying the scores they awarded, completed a questionnaire and participated in verbal report sessions to elaborate on their test administration and rating behaviour. Three researchers also observed all test sessions and took field notes.

While the two modes generated similar test score outcomes, there were some differences in functional output and examiner interviewing and rating behaviours. This report concludes with a list of recommendations for further research, including examiner and test-taker training and resolution of technical issues, before any decisions about deploying (or not) a video-conferencing mode of the IELTS Speaking test delivery are made.

Authors

Fumiyo Nakatsuhara, Chihiro Inoue, CRELLA, University of Bedfordshire

Vivien Berry, British Council

Evelina Galaczi, Cambridge English Language Assessment

Table of contents

1 Introduction **7**

2 Literature review **7**

2.1. Underlying constructs 8

2.2. Cognitive validity 10

2.3. Test-taker perceptions 11

2.4. Test practicality 11

2.5. Video-conferencing and speaking assessment 12

2.6. Summary 13

3 Research questions **14**

4 Methodology **15**

4.1. Research design 15

4.2. Participants 15

4.3. Data collection 16

4.4. Data analysis 19

5 Results **21**

5.1. Score analysis 22

5.2. Language function analysis 28

5.3. Analysis of test-taker interviews 33

5.4. Analysis of observers' field notes, verbal report sessions with examiners, examiners' written comments, and examiner feedback questionnaires 35

6 Conclusions **45**

References **49**

Appendices **52**

Appendix 1: Exam rooms 52

Appendix 2: Test-taker questionnaire 53

Appendix 3: Examiner questionnaire 55

Appendix 4: Observation checklist 58

Appendix 5: Transcription notation 61

Appendix 6: Shifts in use of language functions from Parts 1 to 3 under face-to-face/ video-conferencing conditions 62

Appendix 7: Comparisons of use of language functions between face-to-face (f2f)/ video-conferencing (VC) conditions 63

Appendix 8: A brief report on technical issues encountered during data collection (20–23 January 2014) by Jermaine Prince 66

1 Introduction

This paper reports on a preliminary exploration and comparison of test-taker and examiner behaviours across two different delivery modes for the same L2 speaking test – the standard test administration, and internet-based video-conferencing test administration using Zoom[1] technology. The study sought to compare performance features across these two delivery modes with regard to two key areas:

- an analysis of *test-takers' scores and linguistic output* on the two modes and *their perceptions* of the two modes
- an analysis of *examiners' test management and rating behaviours* across the two modes, including *their perceptions* of the two conditions for delivering the speaking test.

This research study was motivated by the need for test providers to keep under constant review the extent to which their tests are accessible and fair to as wide a constituency of test users as possible. Face-to-face tests for assessing spoken language ability offer many benefits, particularly the opportunity for reciprocal spoken interaction. However, face-to-face speaking test administration is usually logistically complex and resource-intensive, and the face-to-face mode can be difficult or impossible to conduct in geographically remote or politically sensitive areas. An alternative would be to use a semi-direct speaking test, in which the test-taker speaks in response to recorded input delivered via a CD-player or computer/tablet. A disadvantage of the semi-direct approach is that this delivery mode does not permit reciprocal interaction between speakers, i.e. test-taker and interlocutor(s), in the same way as a face-to-face format. As a result, the extent to which the speaking ability construct can be maximally represented and assessed within the speaking test format is significantly constrained.

Recent technical advances in online video-conferencing technology make it possible to engage much more successfully in face-to-face interaction via computer than was previously the case (i.e., face-to-face interaction no longer depends upon physical proximity within the same room). It is appropriate, therefore, to explore how new technologies can be harnessed to deliver and conduct the face-to-face version of an existing speaking test, and what similarities and differences between the two formats can be discerned. The fact that relatively little research has been conducted to date into face-to-face delivery via video-conferencing provides further motivation for this study.

2 Literature review

A useful basis for discussing test formats in speaking assessment is through a categorisation based on the delivery and scoring of the test, i.e. by a human examiner or by machine. The resulting categories (presented visually as quadrants 1, 2 and 3 in Figure 1) are:

- 'direct' human-to-human speaking tests, which involve interaction with another person (an examiner, another test-taker, or both) and are typically carried out in a face-to-face setting, but can also be delivered via phone or video-conferencing; they are scored by human raters
- 'semi-direct' tests (also referred to as 'indirect' tests in Fulcher (2003)), which involve the elicitation of test-taker speech with machine-delivered prompts and are scored by human raters; they can be either online or CD-based
- automated speaking tests which are both delivered and scored by computer.

1 Zoom is an online video-conferencing program (http://www.zoom.us), which offers high definition video-conferencing and desktop sharing. See Appendix 8 for more information.

(The fourth quadrant in Figure 1 presents a theoretical possibility only, since the complexity of interaction cannot be evaluated with current automated assessment systems.)

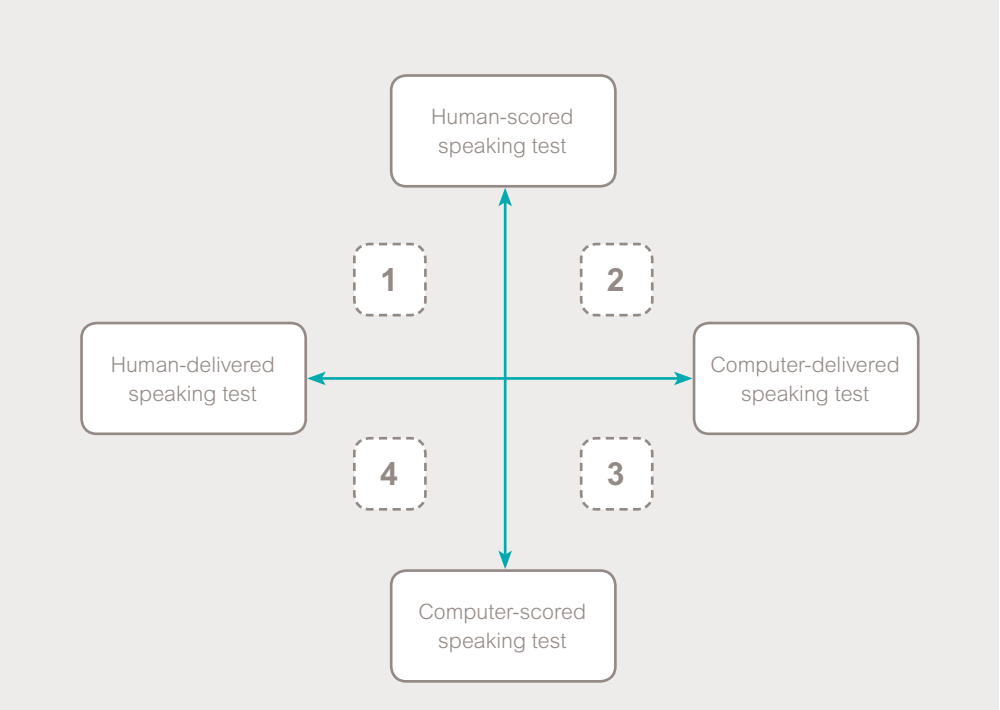


Figure 1: *Delivery and scoring formats in speaking assessment*

Empirical investigations and theoretical discussions of issues relevant to these three general test formats have given rise to a solid body of academic literature in the last two decades, which has focused on a comparison of test formats and, in the process, has revealed important insights about their strengths and limitations. This academic literature forms the basis for the present discussion, since the new speaking test format under investigation in this study is an attempt to overcome some of the limitations associated with existing speaking test formats which the academic literature has alerted us to, while preserving existing strengths.

In the overview to follow, we will focus on key differences between certain test formats. For conciseness, the overview of relevant literature will be mostly limited to the face-to-face direct format and computer-delivered semi-direct format, since they have the greatest relevance for the present study. Issues of scoring will be touched on marginally and only when theoretically relevant. We will, in addition, leave out discussions of test reliability in the context of different test formats, since they are not of direct relevance to the topic of interest here. (Broader discussions of different speaking test modes can be found in Fulcher (2003), Luoma (2004), Galaczi (2010), and Galaczi and ffrench (2010)).

2.1 Underlying constructs

Construct validity is an overriding concern in testing and refers to the underlying trait which a test claims to assess. Since the 1980s, speaking language tests have aimed to tap into the construct of Communicative Competence (Canale and Swain 1980) and Communicative Language Ability (Bachman 1990). These theoretical frameworks place an emphasis on the use of language to perform communicative functions rather than on formal language knowledge. More recently, the notion of Interactional Competence – first introduced by Kramsch (1986) – has taken a central role in the construct definition of speaking tests. Interactional competence goes beyond a view of language competence as residing within an individual to a more social view where communicative

language ability and the resulting performance reside within a social and jointly-constructed context (McNamara and Roever 2006). Direct tests of speaking are, as such, seen as the most suitable when communicative language ability is the construct of interest, since they have the potential to tap into interaction. However, they do have practical limitations, as will be discussed later, which impact on their use.

A fundamental issue to consider is whether and how the delivery medium – i.e. the face-to-face vs. computer-delivered test format in this case – changes the nature of the trait being measured (Chapelle and Douglas 2006; Xi 2010). The key insight to emerge from investigations and discussions of the speaking test formats is that the constructs underlying different speaking test formats are overlapping, but nevertheless different. The construct underlying direct face-to-face speaking tests (and especially paired and group tests) is viewed in *socio-cognitive* terms, where speaking is viewed both as a cognitive trait and a social interactional one. In other words, the emphasis is not just on the knowledge and processing dimension of language use, but also on the social, interactional nature of speaking. The face-to-face speaking test format is interactional, multi-directional and co-constructed. Responsibility for successful communication is shared by the interlocutor and (any) clarifications, speaker reactions to previous turns and other modifications can be accommodated within the overall interaction.

In contrast, computer-delivered speaking assessment is uni-directional and lacks the element of co-construction. Performance is elicited through technology-mediated prompts and the conversation has a pre-determined course which the test-taker has no influence upon (Field 2011, p. 98). As such, computer-based speaking tests draw on a *psycho-linguistic* definition of the speaking construct which places emphasis on the cognitive dimension of speaking. A further narrowing down of the construct is seen in automated speaking tests which are both delivered and scored by computer. These tests represent a narrow psycho-linguistic construct (van Moere 2012) and aim to tap into 'facility in L2' (Bernstein, van Moere and Cheng 2010, p. 356) and 'mechanical' language skills (van Moere 2010, p. 93), i.e. core linguistic knowledge which every speaker of a language has mastery of, and which is independent of the domain of use. These core language skills have been contrasted with 'social' language skills (van Moere 2010, p93), which are part of the human-to-human speaking test construct.

Further insights about similarities and differences between different speaking test formats come from a body of literature focusing on comparisons between the scores and language generated in comparison studies. Some studies have indicated considerable overlap between direct and semi-direct tests in the statistical correlational sense, i.e. people who score high in one format also score high in the other. Score equivalence has, by extension, been seen as construct equivalence. Stansfield and Kenyon, for example, in their comparison between the face-to-face Oral Proficiency Interview and the tape-based Simulated Oral Proficiency Interview concluded that 'both tests are highly comparable as measures of the same construct – oral language proficiency' (1992, p. 363). Wigglesworth and O'Loughlin (1993) also conducted a direct/semi-direct test comparability study and found that the candidates' ability measures strongly correlated, although 12% of candidates received different overall classifications for the two tests, indicating some influence of test method. More recently, Bernstein et al. (2010) investigated the concurrent validity of automated scored speaking tests; they also reported high correlations between human administered/human scored tests and automated scoring tests.

A common distinguishing feature of the score-comparison studies is the sole reliance on statistical evidence in the investigation of the relationship and score equivalence of the two test formats. A different set of studies attempted to address not just the statistical equivalence between computer-based and face-to-face tests, but also the comparability of the linguistic features generated, and extended the focus to qualitative analyses of the language elicited through the two formats. In this respect, Shohamy (1994) reported

discourse-level differences between the two formats and found that when the test-takers talked to a tape recorder, their language was more literate and less oral-like; many test-takers felt more anxious about the test because everything they said was recorded and the only way they had for communicating was speaking, since no requests for clarification and repetition could be made. She concluded that the two test formats do not appear to measure the same construct. Other studies have since then supported this finding (Hoejke and Linnell 1994, Luoma 1997, O'Loughlin 2001), suggesting that 'these two kinds of tests may tap fundamentally different language abilities' (O'Loughlin 2001, p169).

Further insights about the differences in constructs between the formats come from investigations of the functional language elicited in the different formats. The available research shows that the tasks in face-to-face speaking tests allow for a broader range of response formats and interaction patterns, which represent both speech production and interaction, e.g., interviewer–test-taker, test-taker–test-taker, and interviewer–test-taker–test-taker tasks. The different task types and patterns of interaction allow, in turn, for the elicitation and assessment of a wider range of language functions in both monologic and dialogic contexts. They include a range of functions, such as informational functions, e.g., providing personal information, describing or elaborating; interactional functions, e.g., persuading, agreeing/ disagreeing, hypothesising; and interaction management functions, e.g., initiating an interaction, changing the topic, terminating the interaction, showing listener support (O'Sullivan, Weir and Saville 2002).

In contrast, the tasks in computer-delivered speaking tests are production tasks entirely, where a speaker produces a turn as a response to a prompt. As such, computer-delivered speaking tests are limited to the elicitation and assessment of predominantly informational functions. Crucially, therefore, while there is overlap in the linguistic knowledge which face-to-face and computer-delivered speaking tests can elicit, (e.g. lexico-grammatical accuracy/range, fluency, coherence/cohesion and pronunciation), in computer-delivered tests that knowledge is sampled in monologic responses to machine-delivered prompts, as opposed to being sampled in co-constructed interaction in face-to-face tests.

To sum up, the available research so far indicates that the choice of test format has fundamental implications for many aspects of a test's validity, including the underlying construct. It further indicates that when technology plays a role in existing speaking test formats, it leads to a narrower construct. In the words of Fulcher (2003, p. 193): 'given our current state of knowledge, we can only conclude that, while scores on an indirect [i.e. semi-direct] test can be used to predict scores on a direct test, the indirect test is testing something *different* from the direct test'. His contention stills holds true more than a decade later, largely because the direct and semi-direct speaking test formats have not gone through any significant changes. More recently, Qian (2009, p. 116) similarly notes that 'the two testing methods do not necessarily tap into the same type of skill'.

2.2 Cognitive validity

Further insights about differences between speaking test formats come from investigations of the cognitive processes triggered by tasks in the different formats. The choice of speaking test format has key implications for the task types used in a test. This in turn impacts on the cognitive processes which a test can activate and the cognitive validity of a test (Weir 2005; also termed 'interactional authenticity' by Bachman and Palmer 1996).

Different test formats and corresponding task types pose their own specific cognitive processing demands. In this respect, Field (2011) notes that tasks in an interaction-based paired test entail processing input from several interlocutors (including a peer), keeping track of different points of view and topics, as well as the need for test-takers'

familiarity with each other's' L2 variety and the forming of judgements in real-time about the extent of accommodation to the partner's language. These kinds of cognitive decisions during a face-to-face speaking test impose processing demands on test-takers that are absent in computer-delivered tests. In addition, arguments have been put forward that even similar task types – e.g., a long-turn task involving the description of a visual, which is used in all speaking test formats – may become cognitively different when presented through the different channels of communication, due to the absence of body language and facial gestures, which provide signals of listener understanding (Chun 2006; Field 2011). In a computer-delivered context, retrospective self-monitoring and repair, which are part of the cognitive processing of speaking, are also likely to play a smaller role (Field 2011).

The difference in constructs can also be seen not just between different test formats but within the same format. For example, a direct speaking test can be delivered not just face-to-face, but also via a phone. Such a test involves co-construction between two (or more) interlocutors. It lacks, however, the visual and paralinguistic aspect of interaction, and, as such, imposes its own set of cognitive demands. It could also could lead to reduced understanding of certain phonemes due to the lower sound frequencies used, and often leads to intonation assuming a much more primary role than in face-to-face talk (Field 2011).

2.3 Test-taker perceptions

Test-taker perceptions of computer-based tests have received some attention in the literature as well, mostly in the 1980s and 1990s, which was the era of earlier generations of computer-based speaking tests. In those investigations, test-takers reported a sense of lack of control and nervousness (Clark 1988; Stansfield 1990). Such test-taker concerns have been addressed in some newer-generation computer-based oral tests, which give test-takers more control over the course of the test. For example, Kenyon and Malabonga's (2001) investigation of candidate perceptions of several test formats (a tape-delivered semi-direct test, a computer-delivered semi-direct test and a face-to-face test) found that the different tests were seen by test-takers as similar in most respects. The face-to-face test, however, was perceived by the study participants to be a better measure of real-life speaking skills. Interestingly, the authors found that at lower proficiency levels, candidates perceived the computer-based test to be less difficult, possibly due to the adaptive nature of that test which allowed the difficulty level of the assessment task to be matched more appropriately to the proficiency level of the examinees.

In a more recent investigation focusing on test-taker perceptions of different test formats, Qian (2009) reported that although a large proportion of his study participants (58%) had no particular preference in terms of direct or semi-direct tests, the number of participants who strongly favoured direct testing exceeded the number strongly favouring semi-direct testing. However, it should be noted that the two tests used in Qian's study were not comparable in terms of the task formats included. The computer-based test was administered in a computer-laboratory setting, and topics were workplace-oriented. In contrast, the face-to-face test used was an Academic English test. As a result, the differences in task formats and test constructs might have also affected the participating students' perceptions towards the two test formats, in addition to the difference in test delivery formats.

2.4 Test practicality

Discussions focusing on different speaking test formats have also addressed the practicality aspects associated with the different formats. One of the undoubted strengths of computer-delivered speaking tests is their high practical advantage over their face-to-face counterparts. After the initial resource-intensive set-up,

computer-based speaking tests are cost-effective, as they allow for large numbers of test-takers to be tested at the same time (Qian, 2009). They also offer greater flexibility in terms of time, since computer-delivered speaking tests can in principle be offered at any time, unlike face-to-face tests which are constrained by a 'right-here-right-now' requirement. In addition, computer-delivered tests take away the need for trained examiners to be on site. In contrast, face-to-face speaking tests require the development and maintenance of a network of trained and reliable speaking examiners who need regular training, standardisation and monitoring, as well as extensive scheduling during exam sessions (Taylor and Galaczi 2011).

2.5 Video-conferencing and speaking assessment

The body of literature reviewed so far indicates that the different formats that can be used to assess speaking ability offer their unique advantages, but inevitably come with certain limitations. Qian (2009, p. 124) reminded us of this:

> There are always two sides of a matter. This technological development has come at a cost of real-life human interaction, which is of paramount importance for accurately tapping oral language proficiency in the real world. At present, it will be difficult to identify a perfect solution to the problem but it can certainly be a target for future research and development in language testing.

Such a development in language testing can be seen in recent technological advances which involve the use of video-conferencing in speaking assessment. Such a new speaking test mode preserves the co-constructed nature of face-to-face speaking tests, while at the same time, offering the practical advantage of remotely connecting test-takers and examiners who could be continents apart. As such, it reduces some of the practical difficulties of face-to-face tests while preserving the interactional advantage of face-to-face tests.

The use of a video-conferencing system in English language testing can be traced back to the late 1990s. One of the pioneers was ALC, an educational company in Japan, which developed a test of spoken English in conjunction with its online English lessons in collaboration with Waseda University (a private university in Japan), Panasonic and KDDI (IT companies in Japan) in 1999. As part of their innovative collaborative project, they offered group online lessons of spoken English and an online version of the Standard Speaking Test (ALC 2015) using the same technology, where a face-to-face interview test was carried out via computer. The test was used to measure the participating students' oral proficiency before and after a series of lessons. The computer-delivery format was used until 2001, and a total of 1638 students took the test during the three years (Hirano 2015, personal communication). Despite the success of the test delivery format, the test format was not continued after three years, as the university favoured face-to-face English lessons and tests. Since online face-to-face communication was not very common at the time of developing this test, practicality and the costs involved in the use of the technology might have contributed to that decision.

A more recent example of using video-conferencing comes from the University of Nottingham and China and a speaking test based on video-conferencing. In this test, Skype is used to run a speaking assessment for a corporation with staff spread throughout the country (Dawson 2015, personal communication).

2.6 Summary

As a summary, let us consider two key questions: What can machines do better? What can humans do better? As the overview and discussion have indicated so far, the use of technology in speaking assessment has often come at the cost of a narrowing of the construct underlying the test. The main advantage of computer-delivered and computer-scored speaking tests is their convenience and standardisation of delivery, which enhances their reliability and practicality (Chapelle and Douglas 2006; Douglas and Hegelheimer 2007; Jamieson 2005; Xi 2010). The trade-offs, however, relate to the inevitable narrowing of the test construct, since computer-based speaking tests are limited by the available technology and include constrained tasks which lack an interactional component. In computer-based speaking tests, the construct of communicative language ability is not reflected in its breadth and depth, which creates potential problems for the construct validity of the test. In contrast, face-to-face speaking tests and the involvement of human examiners introduces a broader test construct, since interaction becomes an integral part of the test, and so learners' interactional competence can be tapped into. The broader construct, in turn, enhances the validity and authenticity of the test. The caveat with face-to-face speaking tests is the low practicality of the test and the need for a rigorous and ongoing system of examiner recruitment, training, standardisation and monitoring on site.

The remote face-to-face format is making an entry into the speaking assessment field and holds potential to optimise strengths and minimise shortcomings by blending technology and face-to-face assessment. Its advantages and limitations, however, are still an open empirical question. As can be seen in the literature reviewed above, much effort has been put into exploring potential differences between interactive face-to-face oral interviews and simulated or computer oral interviews (SOPI and COPI respectively). The primary differences between the two are that, in the former, a 'live' examiner interacts in real time with the test-taker or test-takers, whereas in the latter, these individuals respond to pre-recorded tasks; while the former is build on interaction, there is no interactivity in the latter. In the two cases where attempts have been made to deliver an oral interview in real time, with a 'live' examiner interacting with test-takers, no empirical evidence has been gathered or reported to support or question the approach.

The present study aims to provide a preliminary exploration of the features of this new and promising speaking test format, while at the same time, opening up a similarly new and exciting area of research.

3 Research questions

This study considered the following six research questions. The first three questions relate to test-takers and the rest relate to examiners.

RQ1a: Are there any differences in **test-takers' scores** between face-to-face and video-conferencing delivery conditions?

RQ1b: Are there any differences in **linguistic output, specifically types of language function**, elicited from test-takers under face-to-face and video-conferencing delivery conditions?

RQ1c: What are **test-takers' perceptions** of taking the test under face-to-face and video-conferencing delivery conditions?

RQ2a: Are there any differences in **examiners' test administration behaviour (i.e. as interlocutor)** under face-to-face and video-conferencing delivery conditions?

RQ2b: Are there any differences in **examiners' rating behaviour** when they assess test-takers under face-to-face and video-conferencing delivery conditions?

RQ2c: What are **examiners' perceptions** of examining under face-to-face and video-conferencing delivery conditions?

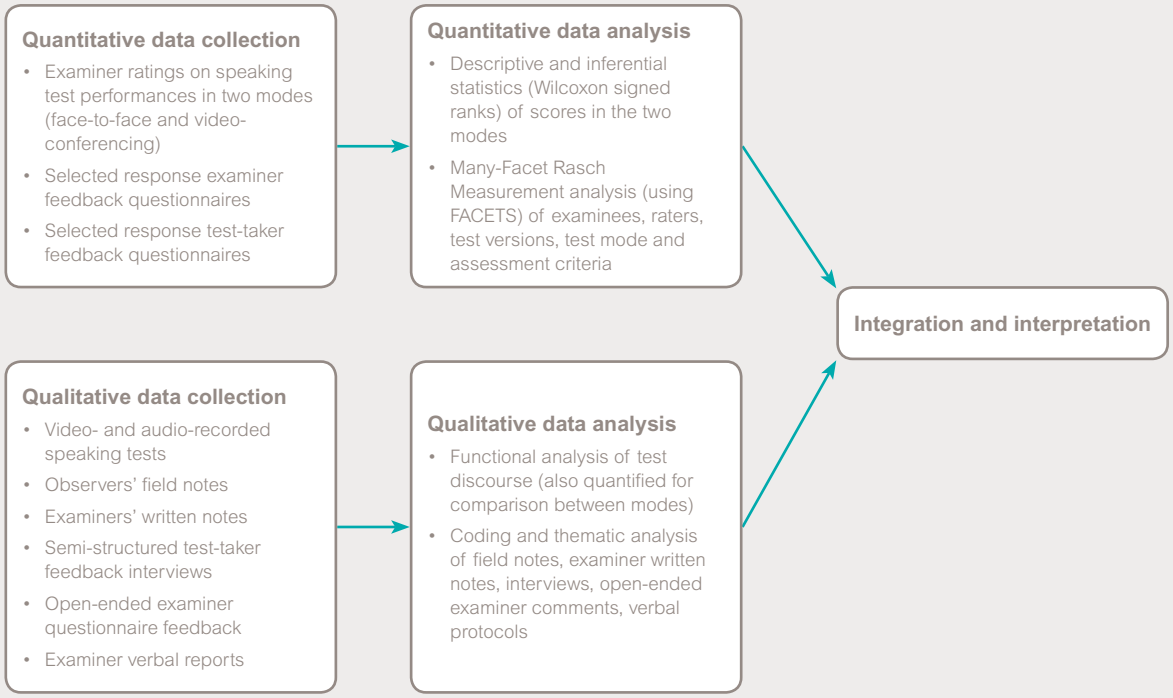
4 Methodology

4.1 Research design

The study used a convergent parallel mixed methods design (Creswell and Plano Clark 2011), where quantitative and qualitative data were collected in two parallel strands, analysed separately and findings were integrated. The two data strands provide different types of information and allow for a more in-depth and comprehensive set of findings.

Figure 2 presents information on the data collection and analysis strands in the research design.

Figure 2: *Research design*



4.2 Participants

Thirty-two test-takers, who were attending IELTS preparation courses at Ealing, Hammersmith & West London College, signed up in advance to participate in the study. As an incentive, they were offered the choice of either 1) having their fee paid for a live IELTS test, or 2) a small honorarium. Class tutors were asked to estimate their IELTS Speaking Test scores; these ranged from Band 4.0 to Band 7.0. Due to practical constraints, not all of the original students were able to participate and some substitutions had to be made; the range of the face-to-face IELTS Speaking scores of the 32 students who ultimately participated was Band 5.0 to Band 8.5. This score range was higher than originally planned by the research team (see Figure 4 in Section 5), but nevertheless was considered adequate for the purposes of the study.

Four trained, certificated and experienced IELTS examiners (i.e., Examiners A–D) also took part in the research. Examiners were paid the normal IELTS rate for each test plus an estimated amount for time spent on completion of the questionnaire and participation in the verbal report protocols. All travel expenses were also reimbursed.

Each examiner examined eight test-takers in both modes of delivery across two days.

Signed consent forms were obtained from all test-takers and examiners.

4.3 Data collection

4.3.1 Modes of delivery for the speaking tests

- Face-to-face is the traditional delivery mode for IELTS speaking tests and consists of spoken interaction between a test-taker and an examiner sitting opposite each other in an examination room.
- The video-conferencing mode was operationalised using iPad hardware and Zoom software (see Appendix 8 for information relating to this software). In this mode, the spoken interaction between the test-taker and the examiner also took place in real time but the test-taker and the examiner were in different rooms and interacted with each other via computer screens.

Data on both delivery modes were collected from all three parts of the test: Part 1 – Question and Answer exchange; Part 2 – Test-taker long turn, and Part 3 – Examiner and test-taker discussion.[2]

4.3.2. Speaking test performances and questionnaire completion

All 32 test-takers took both face-to-face and video-conferencing speaking tests in a counter-balanced order.

Two versions of the IELTS Speaking test (i.e. Versions 1 and 2[3]; retired test versions obtained from Cambridge English Language Assessment) were used, and the order of the two versions was also counter-balanced.

The data collection was carried out over four days. On each day, two parallel test sessions were administered (one face-to-face and one via video-conferencing). Two examiners carried out test sessions on each day, and they were paired with different examiners on these two days (i.e. Day 1: Examiners A and B; Day 2: Examiners B and C; Day 3: Examiners C and D; Day 4: Examiners D and A). Table 1 shows the data collection matrix used for the data collection on Day 1.

All test sessions were audio- and video-recorded using digital audio recorders and external video cameras. The video-conferencing test sessions were also video-recorded using Zoom's on-screen recording technology (see Appendix 1 for the test room settings).

After two test sessions (i.e. one face-to-face and one video-conferencing test), test-takers were interviewed by one of the researchers. The interview followed eight questions specified in a test-taker questionnaire (see Appendix 2), and they were also asked to elaborate on their responses wherever appropriate. The researchers noted test-takers' responses on the questionnaire, and each interview took less than five minutes.

A week before the test sessions, two mini-trials were organised to check: 1) how well the Zoom video-conferencing software worked in the exam rooms; and 2) how well on-screen recording of the video-conferencing sessions, as well as video-recording by external cameras in the examination rooms, could be carried out. The four examiners were also briefed as to the data collection procedures and how to administer tests using Zoom.

2 For more information on each task type, see www.ielts.org

3 These two versions of the test were carefully selected to ensure comparability of tasks (e.g. topic familiarity, expected output).

Table 1: *Data collection matrix*

Time	Face-to-face	Video-conferencing
0–20 (inc. 5-min admin time)	Examiner A – Test-taker 1	Examiner B – Test-taker 2
20–40	Examiner B – Test-taker 2	Examiner A – Test-taker 1
5 mins for test-taker interview	Researcher 1 – Test-taker 2	Researcher 2 – Test-taker 1
45–65	Examiner B – Test-taker 4	Examiner A – Test-taker 3
65–85	Examiner A – Test-taker 3	Examiner B – Test-taker 4
5 mins for test-taker interview	Researcher 2 – Test-taker 3	Researcher 1 – Test-taker 4
15 mins + 5 mins above	– Examiner break –	
105–125	Examiner A – Test-taker 5	Examiner B – Test-taker 6
125–145	Examiner B – Test-taker 6	Examiner A – Test-taker 5
5 mins for test-taker interview	Researcher 1 – Test-taker 6	Researcher 2 – Test-taker 5
150–170	Examiner B – Test-taker 7	Examiner A – Test-taker 8
170–190	Examiner A – Test-taker 8	Examiner B – Test-taker 7
5 mins for test-taker interview	Researcher 2 – Test-taker 8	Researcher 1 – Test-taker 7

4.3.3 Observers' field notes

Three researchers stayed in three different test rooms and took field notes. One of them (Researcher 3) stayed in the test-takers' video-conferencing room so that she could see all students performing under the video-conferencing test conditions.

The other two researchers (Researcher 1 and Researcher 2) observed test sessions in both face-to-face and examiners' video-conferencing rooms. Each of them followed one particular examiner on each day, to enable them to observe the same examiner's behaviour under the two test delivery conditions. The research design ensured that Researchers 1 and 2 could observe all four examiners on different days (e.g. Examiner B's sessions were observed by Researcher 1 on Day 1 and by Researcher 2 on Day 2).

4.3.4 Examiners' ratings

Examiners in the live tests awarded scores on each analytic rating category (i.e. *Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, Pronunciation*), according to the standard assessment criteria and rating scales used in operational IELTS tests. (In the interest of space, the rating categories are hereafter referred to as Fluency, Lexis, Grammar and Pronunciation.)

Score analysis was planned to be carried out using only the single-rated scores awarded in the live tests at this stage of the study. Various options to carry out multiple ratings using video-recorded performances were considered during the research planning stage. However, the research team felt that this could introduce a significant confounding variable at the current stage of the exploration, namely rating video-recorded performance on face-to-face and video-conferencing delivery modes, whose effect we were not able to predict at this stage, due to the lack of research in this area. Given the limited time available for this study, together with its preliminary and exploratory nature, it was felt that it would be best to limit this study to the use of live rating scores obtained following a rigorous counter-balanced data collection matrix (see Table 1). Nevertheless, this does not exclude the possibility of carrying out a multiple ratings design which could form a separate follow-up study in the future.

4.3.5 Examiners' written comments

After each live test session, the examiners were asked to make brief notes on why they awarded the scores that they did on each of the four analytic categories. Compared with the verbal report methodology (as described below), a written description is likely to be less informative. However, given its ease in collecting larger datasets in this manner, it was thought to be worth obtaining brief notes from examiners to supplement a small quantity of verbal report data (e.g., Isaacs 2010).

4.3.6 Examiner feedback questionnaires

After completing all speaking tests of the day, examiners were asked to complete an examiner feedback questionnaire about: 1) *their own behaviour* as interlocutor under video-conferencing delivery and face-to-face test conditions; and 2) *their perceptions* of the two test delivery modes. The questionnaire consisted of 28 questions and free comments boxes, and took about 20 minutes for examiners to complete (see Appendix 3).

4.3.7 Verbal reports by examiners on the rating of test-takers' performances

After completing all speaking tests of the day, together with a feedback questionnaire, examiners took part in verbal report sessions facilitated by one of the researchers. Each verbal report session took approximately 50 minutes.

Seven video-conferencing and seven face-to-face video-recorded test sessions by the same seven test-takers were selected for collecting examiners' verbal report data. The intention was to select at least one test-taker from each of IELTS Bands 4.0, 5.0, 6.0 and 7.0 respectively, to cover a range of performance levels. However, due to the lack of test-takers at IELTS Band 4.0, the IELTS overall band scores of the seven test-takers included Bands 4.5, 5.0, 5.5, 6.0, 6.5, 7.0 and 7.5 in one or both of the delivery modes (see Section 5.4).

The same four trained IELTS examiners participated in the verbal report sessions, and one of the two researchers who observed the live interviews acted as a facilitator. A single verbal report per test session was collected from the examiner who actually interviewed the test-taker. All examiners carried out verbal report sessions with both of the researchers across the four days. In total, 14 verbal reports were collected (one examiner was only available to participate in two verbal report sessions).

The examiners were first given a short tutorial that introduced the procedures for verbal report protocols. Then, following the procedure used in May (2011), verbal report data were collected in two phases, using stimulated recall methodology (Gass and Mackey 2000):

- **Phase 1:** Examiners watched a video without pausing while looking at the written comments they made during the live sessions, and made general oral comments about a test-taker's overall task performance.
- **Phase 2:** Examiners watched the same video clip once again, and were asked to pause the video whenever necessary to make comments about any features that they found interesting or salient related to the four analytic rating categories, and any similarities and differences between the two test delivery modes. The participating researcher also paused the video and asked questions to the examiners, whenever they wished to do so.

The order of verbal reporting sessions on video-conferencing and face-to-face videos for the four examiners was counter-balanced. The researchers took notes during the verbal report sessions, and all sessions were also audio-recorded.

4.4 Data analysis

Scores awarded under face-to-face and video-conferencing conditions were compared using both Classical Testing Theory (CTT) analysis with the Wilcoxon signed-rank tests[4], and Many-Facet Rasch analysis (MFRM) using the FACETS 3.71 analysis software (Linacre 2013a). The two analyses are complementary and add insights from different perspectives in line with the mixed-methods design outlined earlier. The CTT analysis was, however, from the outset seen as the primary quantitative analysis procedure to address RQ1a because of the constraints imposed by the data collection plan (i.e. each examiner rated the same test-takers in both modes).

The Wilcoxon signed-rank tests (CTT) were used to examine whether there are any statistically significant differences between the two test-delivery conditions (RQ1a, see Section 5.1). The counter-balanced research design was implemented to minimise scoring errors related to different rater severity levels, given that the single rating design would not allow for the identification of variable rater harshness within the CTT analysis. The CTT design is thus based on the assumption that any such rater differences have been controlled, and that scoring differences will be related to test-taker performance and or delivery mode.

The MFRM analysis was carried out to add insight into the results of the main CTT analysis of scoring differences across the two modes. The MFRM analysis adds insight into the impact of delivery mode on the scores, but also helps us to investigate rater consistency, as well as potential differences in difficulty across the analytic rating scales used in the two modes. The method used for ensuring sufficient connectivity in the MFRM analysis, and the important assumptions and limitations associated with this methodology are discussed further in Results and Conclusion (Sections 5 and 6).

All 32 recordings were analysed for language functions elicited from test-takers, using a modified version of O'Sullivan et al.'s (2002) observation checklist (see Appendix 4 for the modified checklist). Although the checklist was originally developed for analysing language functions elicited from test-takers in paired speaking tasks of the Cambridge Main Suite examinations, the potential to apply the list to other speaking tests (including the IELTS Speaking Test) has been demonstrated (e.g., Brooks 2003, Inoue 2013). Three researchers who are very familiar with O'Sullivan et al.'s checklist watched all videos and coded elicited language functions specified in the list.

The codings were carried out to determine whether each function was elicited in each part of the test, rather than how many instances of each function were observed; it did not seem to be feasible to count the number of instances when the observation checklist was applied to video-recorded performances without transcribing them (following the approach of O'Sullivan et al. 2002). The researchers also took notes of any salient and/or typical ways in which each language function was elicited under the two test conditions. This was to enable transcription of relevant parts of the speech samples and detailed analysis of them. The results obtained from the face-to-face and video-conferencing delivered tests were then compared (RQ1b, see Section 5.2).

Closed questions in test-takers' feedback interview data were analysed statistically to identify any trends in their perceptions of taking the test under face-to-face and video-conferencing delivery conditions (RQ1c, see Section 5.3). Their open-ended comments were used to interpret the statistical results and to illuminate the results obtained by other data sources.

4 The Wilcoxon signed-rank test is the non-parametric equivalent of the paired samples t-test. The non-parametric tests were used, as the score data were not normally distributed.

When Researchers 1 and 2 observed live test sessions, they noted any similarities and differences identified in examiners' behaviours as interlocutors. These observations were analysed in conjunction with the first part of the examiners' questionnaire results related to their test administration behaviour (RQ2a, see Section 5.4).

All written comments provided by the examiners on their rating score sheets were typed out so these could be compared across the face-to-face and video-conferencing conditions. The two researchers who facilitated the 14 verbal report sessions took detailed observational notes during the verbal report sessions, and recorded examiners' comments. Resource limitations made it impossible to transcribe all the audio/video data from the 14 verbal report sessions with the examiners. Instead, the audio/video recordings were reviewed by the researchers to identify key topics and perceptions referred to by the examiners during the verbal report sessions.

These topics and comments were then captured in spreadsheet format so they could be coded and categorised according to different themes, such as 'turn taking', 'nodding and back-channelling' and 'speed and articulation of speech'. A limited number of relevant parts of the recordings were later transcribed, using a slightly simplified version of Conversation Analysis notation (modified from Atkinson and Heritage 1984; Appendix 5). The quantity and thematic content of written commentaries and verbal reports were then compared between the face-to-face and video-conferencing modes.

Examinations were also carried out as to whether either mode of test delivery led to examiners' attention being oriented to more positive or negative aspects of test-takers' output related to each analytic category (RQ2b, see Section 5.4).

The second part of the examiner questionnaire regarding examiners' perceptions towards the two different delivery modes were analysed, in conjunction with the results of other analyses as described above (RQ2c, see Section 5.4).

The results obtained in the above analyses of test-takers' linguistic output, test scores, test-taker questionnaire responses, examiners' questionnaire responses, written comments and verbal reports were triangulated to explore and give detailed insight into how the video-conferencing delivery mode compares with the more traditional face-to-face mode from multiple perspectives.

5 RESULTS

This section presents the findings of the research, while answering each of the six research questions raised in Section 3. Before moving on to the findings, it is necessary to briefly summarise the participating students' demographic information and the way in which the planned research design was implemented.

The 32 participating students consisted of 14 males (43.8%) and 18 females (56.3%). Their ages ranged from 19 to 51 years with a mean of 30.19 (SD=7.78) (see Figure 3 for a histogram on age distribution). The cohort comprised 21 different first language (L1) speakers as shown in Table 2. As mentioned earlier, their speaking proficiency levels were higher than expected, and their face-to-face IELTS Speaking scores ranged from Bands 5.0 to 8.5 (see Figure 4). In retrospect, this is perhaps not totally unsurprising as it may be unrealistic to expect students who are considered to be at Band 4 to willingly participate in IELTS tests.

Figure 3: *Age distribution of test-takers*
(N=31 due to one missing value)

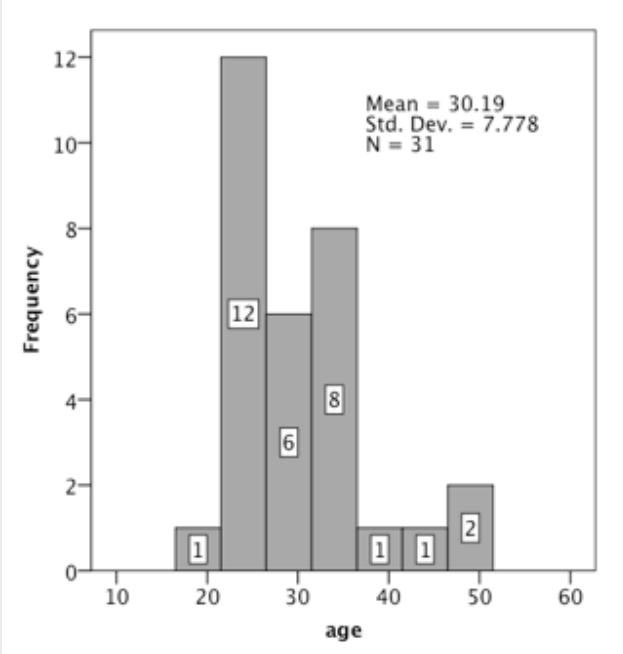


Figure 4: *Participating students' IELTS Speaking test scores (face-to-face)*

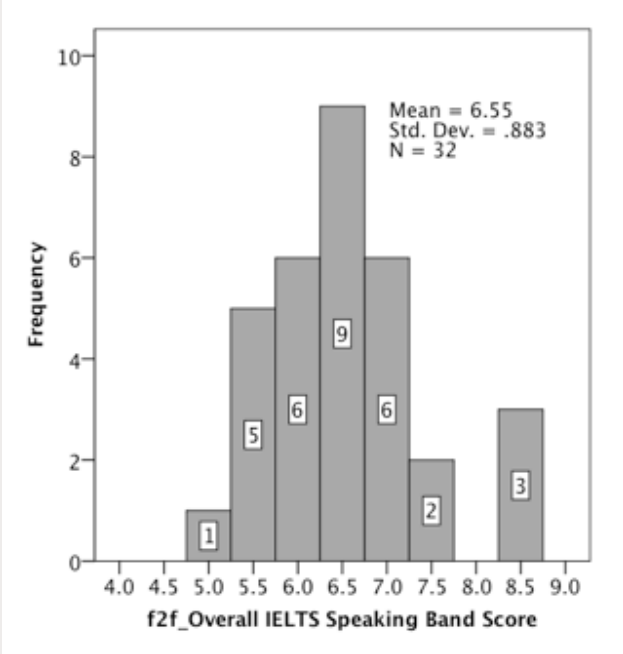


Table 2: *Participants' first languages*

L1	No. of participants	L1	No. of participants
Arabic	1	Kosovan	1
Armenian	2	Nepalese	1
Bengali	1	Polish	2
Bulgarian	1	Russian	1
Burmese	2	Somali	2
English	1	Spanish	4
Estonian	1	Sudanese	1
French	2	Tagalog	1
Italian	3	Turkish	1
Japanese	3	Urdu	1

This sample can be considered representative of the overall IELTS population, since all participants' L1s, with the exception of Kosovan and Sudanese, are in the typical IELTS top 50 test-taker L1s (www.ielts.org).

Table 3 shows that the 64 tests carried out with the 32 students were perfectly counter-balanced in terms of the order of the two test modes and the order of the two test versions. These tests were equally distributed to the four examiners. With this data collection design, we can assume that any order effects or examiner effects can be minimised, if not cancelled out.

Table 3: *Test administration*

Order of test modes	Order of test versions	Examiner	Frequency	Percent (%)
Face-to-face → Video-conferencing	ver 1 → ver 2	Examiner A	2	25.0
		Examiner B	2	25.0
		Examiner C	2	25.0
		Examiner D	2	25.0
		Total	*8*	*100.0*
	ver 2 → ver 1	Examiner A	2	25.0
		Examiner B	2	25.0
		Examiner C	2	25.0
		Examiner D	2	25.0
		Total	*8*	*100.0*
Video-conferencing → Face-to-face	ver 1 → ver 2	Examiner A	2	25.0
		Examiner B	2	25.0
		Examiner C	2	25.0
		Examiner D	2	25.0
		Total	*8*	*100.0*
	ver 2 → ver 1	Examiner A	2	25.0
		Examiner B	2	25.0
		Examiner C	2	25.0
		Examiner D	2	25.0
		Total	*8*	*100.0*

5.1 Score analysis

We now move on to score analysis to answer RQ1a: Are there any differences in test-takers' scores between face-to-face and video-conferencing delivery conditions?

5.1.1. Classical Test Theory Analysis

Table 4 shows that there were no significant differences in test scores awarded to the four rating categories and two overall scores (mean and rounded).

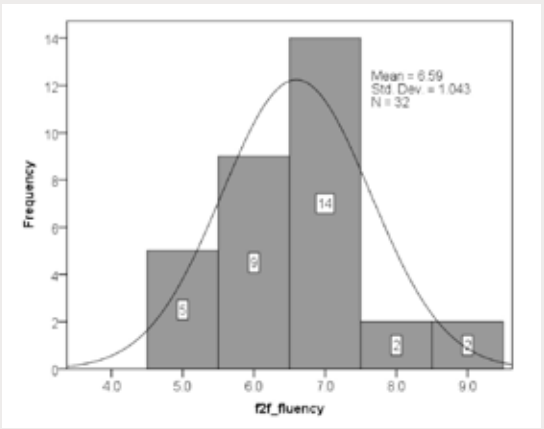
Table 4: *Wilcoxon signed rank test on test scores*

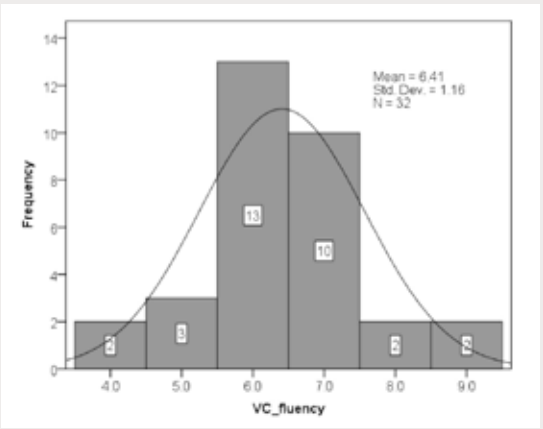
Rating category	Test mode	Median	Mean	Standard deviation	Z	Sig. (2-tailed)
Fluency	Face-to-face	7.000	6.594	1.043	-1.732	.083
	Video-conferencing	6.000	6.406	1.160		
Lexis	Face-to-face	7.000	6.750	1.047	-.302	.763
	Video-conferencing	7.000	6.719	1.143		
Grammar	Face-to-face	6.500	6.625	1.008	.000	1.000
	Video-conferencing	7.000	6.625	1.100		
Pronunciation	Face-to-face	7.000	6.688	.780	-1.667	.096
	Video-conferencing	7.000	6.531	.879		
Overall[5] (mean)	Face-to-face	6.750	6.664	.829	-1.503	.133
	Video-conferencing	6.500	6.570	.982		
Overall (rounded)	Face-to-face	6.500	6.547	.883	-1.031	.302
	Video-conferencing	6.500	6.469	.991		

5 The first overall category shows mean overall scores, and the second overall category shows overall scores that are rounded down as in the operational IELTS test (i.e. 6.75 becomes 6.5, 6.25 becomes 6.0).

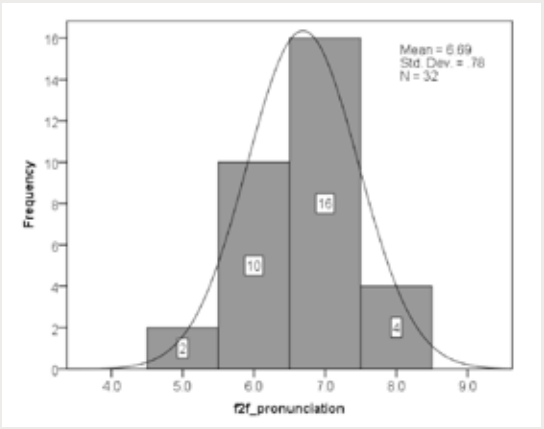
Further descriptive analyses were performed for Fluency and Pronunciation, since the mean differences (although not statistically significant) were slightly larger than the two analytical categories, although the differences are still negligibly small. Figures 5.1 to 6.2 present histograms for Fluency and Pronunciation scores under the face-to-face and video-conferencing conditions, respectively.

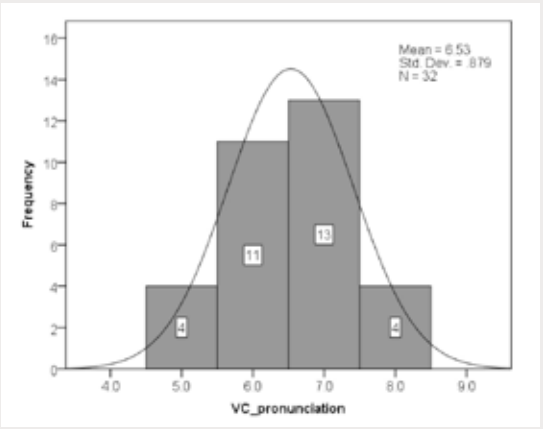
Figures 5.1 and 5.2: *Histograms of Fluency scores*





Figures 6.1 and 6.2: *Histograms of Pronunciation scores*





For Fluency, the most frequent score awarded was 7.0 in the face-to-face mode, and 6.0 in the video-conferencing mode. For Pronunciation, the most frequent score was 7.0 in both modes, but the video-conferencing mode showed higher frequencies of lower scores (i.e. Score 7.0 (n=13), 6.0 (n=11), 5.0 (n=4)) than the face-to-face mode did (i.e. Score 7.0 (n=16), 6.0 (n=10), 5.0 (n=2)). Although neither of these differences led to statistical significance at p=0.5, it is worth investigating possible reasons why these differences were obtained. However, it must be remembered that non-significant results are likely to mean that there is nothing systematic happening and therefore our consideration of them is simply speculative.

Additionally, following examiners' comments that those who get affected most by the delivery mode might be older test-takers and/or lower achieving test-takers (see Section 5.4), further comparisons were made for different age groups and different proficiency groups. For two-group (above/below mean) and three-group comparisons (divided by the points at ±1 SD away from mean)[6], no clear difference seemed to emerge. However, descriptive statistics indicated that the lowest achieving group who scored less than 1 SD below the mean (i.e., Band 5.5 or below, N=6) showed consistently lower mean scores under the video-conferencing condition across all rating categories. The younger group of test-takers who were below the mean age (i.e., 30 years old or younger) scored statistically significantly lower in Pronunciation under the video-conferencing condition (Median: 7.00 in face-to-face, 6.50 in video-conferencing, Mean: 6.83 in face-to-face, 6.44 in video-conferencing, Z=-2.646, p=0.008).

It seems worth investigating possible age and proficiency effects in the future with a larger dataset, as the small sample size of this study did not allow meaningful inferential statistics. Therefore, no general conclusions can be drawn here.

6 For the proficiency level comparisons, two groups were with 1) Band 7.0 and above (N=11) and 2) Band 6.5 and below (N=21). Three groups were with 1) Band 7.5 and above (N=5), 2) Bands 6.0 to 7.0 (N=21), and 3) Bands 5.5 and below (N=6). For the age comparisons, two groups were with 1) 31 years old and older (N=13) and 2) 30 years old and younger (N=18). Three groups were with 1) 38 years old and older (N=4), 2) 23 to 37 years old (N=25) and 3) 22 years old and younger (N=2).

5.1.2 Many-Facet Rasch Measurement (MRFM) analysis

Two MFRM analyses using FACETS 3.71.2 (Linacre 2013a) were carried out: a 5-facet analysis with *examinee*, *rater, test version, mode* and *rating criteria* as facets, and a 4-facet analysis with *examinees, raters, test version* and *rating criteria* as facets.
The reason for conducting the two analyses was to allow for investigation of the effect of delivery mode on scores in the 5-facet analysis, and to investigate the performance of each analytic rating scale in each mode as a separate "item" in the 4-facet analysis. The difference lies in the conceptualisation of the rating scales as items. In the 5-facet analysis, only four rating scales were designated as items, and examinees' scores on the four analytic rating criteria were differentiated according to delivery mode (i.e. an examinee received a score on Fluency in the face-to-face mode, and a separate score on the same scale in the video-conferencing mode). In this analysis, there were four rating scale items, *Fluency, Lexis, Grammar* and *Pronunciation*. In the 4-facet analysis, delivery mode was not designated as a facet, and each of the analytic rating scales was treated as a separate item in each mode resulting in eight item scores, one for *Face-to-Face Fluency*, one for *Video-Conferencing Fluency*, one for *Face-to-Face Lexis*, one for *Video-Conferencing Lexis*, etc.

Before discussing the results of the two analyses, it is first necessary to specify how sufficient connectivity was achieved, and the caveats this entails for interpreting the results.

As noted above, there was no overlap in the design between raters and examinees, resulting in disjoint subsets and insufficient connectivity for a standard MFRM analysis. One way to overcome disjoint subsets is to use group anchoring to constrain the data to be interpretable within a common frame of reference (Bonk and Ockey 2003; Linacre 2013b). Group anchoring involves anchoring the mean of the groups appearing as disjoint subsets, in this case examinees grouped according to the examiner by whom they were rated. The group mean was anchored at 0 for these examinee groups, which still allows the individual elements (i.e. each examinee) to float in relation to the

designated mean. Group anchoring allows sufficient connectivity for the other facets to be placed onto the common scale within the same measurement framework, and quantitative differences in terms of rater severity, difficulty of delivery mode, and difficulty of individual rating scale items to be compared on the same Rasch logit scale. Nevertheless, this anchoring method also entails some limitations, which will be described in the conclusion.

The common frame of reference was further constrained by anchoring the difficulty of the test versions. The assumption of test versions being equivalent is borne out by the straightforward means, with both Version 1 and 2 having identical observed score means of 6.66. However, given that the administration of versions was completely counter-balanced, and the data indicate that any test-version effect is very small, the estimates of the other elements would not be likely to change whether a Version is anchored or not (Linacre, personal communication).

The measurement report for raters in both the 5- and 4-facet analyses, showing the severity in terms of the Rasch logit scale and the Infit Mean Square index (commonly used as a measure of fit in terms of meeting the assumptions of the Rasch model) are shown in Table 5. Although the FACETS program provides two measures of fit, Infit and Outfit, only Infit is addressed here, as it is less susceptible to outliers in terms of a few random unexpected responses. Unacceptable Infit results are thus more indicative of some underlying inconsistency in an element.

Table 5: *Rater measurement report (5-facet analysis & 4-facet analysis)*

Rater	5-facet analysis			4-facet analysis		
	Logit measure[7]	Standard error	Infit mean square	Logit measure	Standard error	Infit mean square
A	0.8	0.24	0.83	0.81	0.24	0.81
B	-2.87	0.23	0.95	-2.89	0.23	0.97
C	0.04	0.26	0.91	0.04	0.26	0.92
D	0.44	0.24	1.16	0.45	0.24	1.16

The same output for rating scales in the 5- and 4-facet analyses is shown in Tables 6 and 7, respectively.

Table 6: *Delivery mode measurement report (5-facet analysis)*

Test mode	Logit measure	Standard error	Infit mean square
Face-to-face	-0.16	0.17	1.08
Video-conferencing	0.16	0.17	0.85

(Population): Separation .00; Strata .33; Reliability (of separation) .00
(Sample): Separation .87; Strata 1.49; Reliability (of separation) .43
Model, Fixed (all same) chi-square: 1.8; d.f.: 1; significance (probability): .19

7 Rater severity in this table is not discussed intentionally due to possible inaccuracy caused by the group anchoring method. Because we have based our connectivity on group anchoring of the examinees, the MFRM analysis thus prioritises the interpretation that any group differences are due to differential rater severity. Table 5 shows that Examiner B was potentially more lenient than the other raters. However, we cannot judge from this analysis whether Examiner B was actually more lenient than the other raters or that the group Examiner B assessed had higher proficiency than the other groups. This limitation of the MFRM analysis in this study will be revisited in the conclusion section.

Table 7: *Rating scale measurement report (4-facet analysis)*

Test mode	Logit measure	Standard error	Infit mean square
Face-to-face Fluency	0.13	0.34	1.01
Face-to-face Lexis	-0.57	0.35	1.33
Face-to-face Grammar	0.02	0.34	0.93
Face-to-face Pronunciation	-0.22	0.34	1.06
Video-conferencing Fluency	0.71	0.34	0.82
Video-conferencing Lexis	-0.45	0.35	0.79
Video-conferencing Grammar	0.02	0.34	0.96
Video-conferencing Pronunciation	0.36	0.34	0.83

(Population): Separation .57; Strata 1.09; Reliability (of separation) .24
(Sample): Separation .71; Strata 1.28; Reliability (of separation) .34
Fixed (all same) chi-square: 10.5; d.f.: 7; significance (probability): .16

Infit values in the range of 0.5 to 1.5 are 'productive for measurement' (Wright and Linacre 1994), and the commonly acceptable range of Infit is between 0.7 and 1.3 (Bond and Fox 2007). Infit values for all the raters and the rating scales except face-to-face Lexis fall within the acceptable range, and the Infit value for face-to-face Lexis is only marginally over the upper limit (i.e. 1.33). The lack of misfit gives us confidence in the results of the analyses and the Rasch measures derived on the common scale. It also has important implications for the construct measured by the two modes being uni-dimensional.

The results of placing each element within each facet on a common Rasch scale of measurement are shown visually in variable maps produced by FACETS. The variable map for the 5-facet analysis is shown in Figure 7, and that for the 4-facet analysis in Figure 8.

Figure 7: *Variable map (5-facet)*

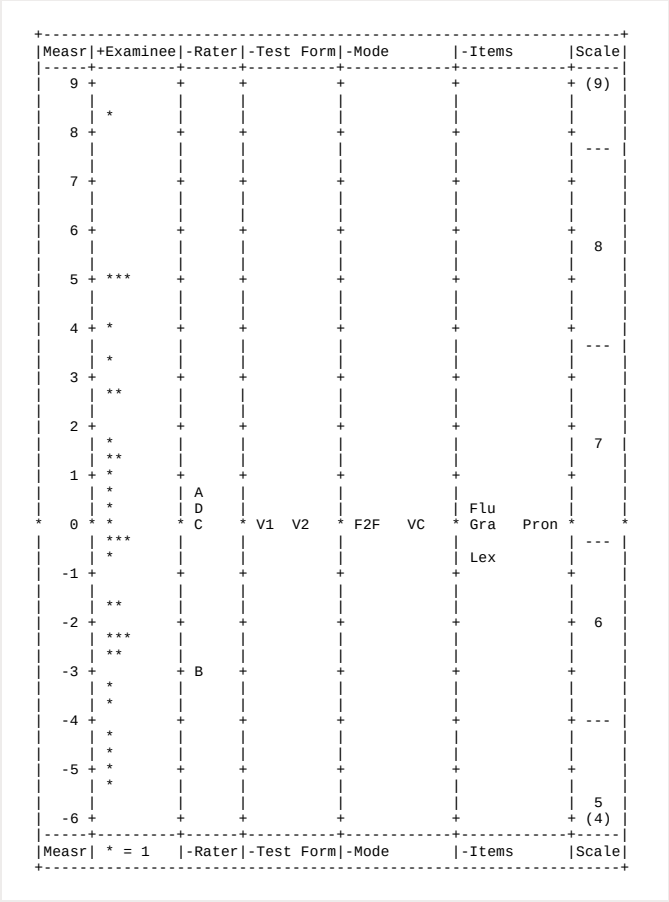
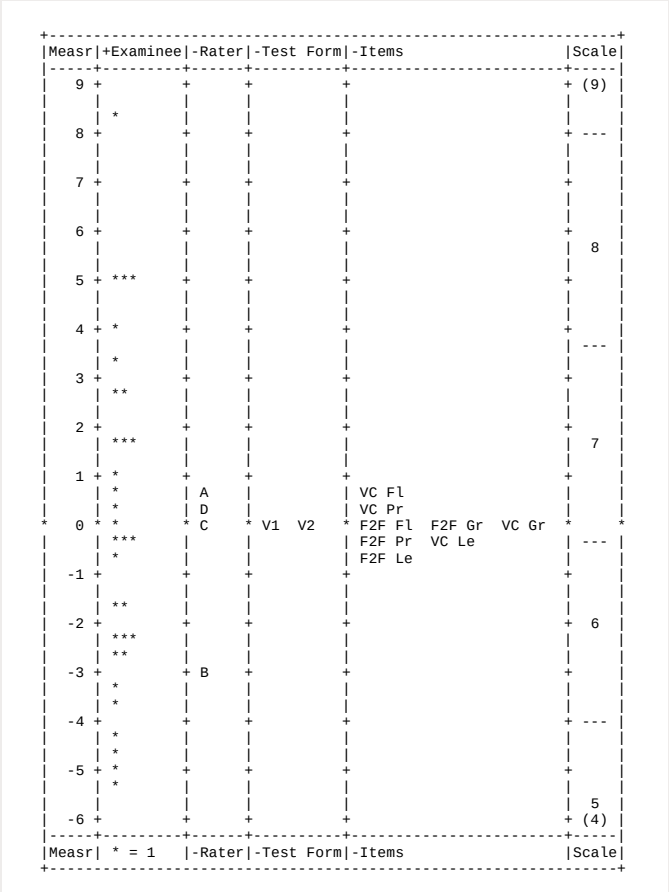


Figure 8: *Variable map (4-facet)*



Note: VC=Video-Conferencing, F2F=Face-to-Face

Of most importance for answering RQ1a are the results for the delivery mode facet in the 5-facet analysis. Figure 7 for the 5-facet analysis shows the placement of the two modes on the common Rasch scale. While video-conferencing is marginally more difficult than the face-to-face mode, fixed chi-square statistics, which test the null hypothesis that all elements of the facets are equal, indicate that the two modes are not statistically different in terms of difficulty (X^2=1.8, p=0.19; see the measurement report for delivery mode in Table 7 above). This reinforces the results of the CTT analysis, and strengthens the suggestion that no significant differences impacting on scores were demonstrated for the effect of delivery mode on actual scores.

The 4-facet analysis further supports the results of the CTT analysis in that *Video-Conferencing Fluency* and *Video-Conferencing Pronunciation* are the most difficult scales, while the other scales cluster together with no pattern of difference related to whether the scale is for the face-to-face mode or video-conferencing mode (see the 4-facet variable map in Figure 8). Although eight rating scales (i.e. four rating scales in face-to-face and four rating scales in video-conferencing) did not show statistically significant differences (see fixed chi-square statistics in Table 8; X2=10.5, p=0.16), the scales for *Fluency* and *Pronunciation* do seem to demonstrate some interaction with delivery mode. As will be later discussed in Section 5.4, the fact that *Pronunciation* was slightly more difficult in the video-conferencing mode seems to relate to the issues with sound quality noted by examiners. For *Fluency*, there seems to be a tendency (at least in some examiners) to constrain back-channelling in the video-conferencing mode (although other examiners emphasised it). The interaction between the mode and back-channelling might have resulted in slightly lower *Fluency* scores under the video-conferencing condition.

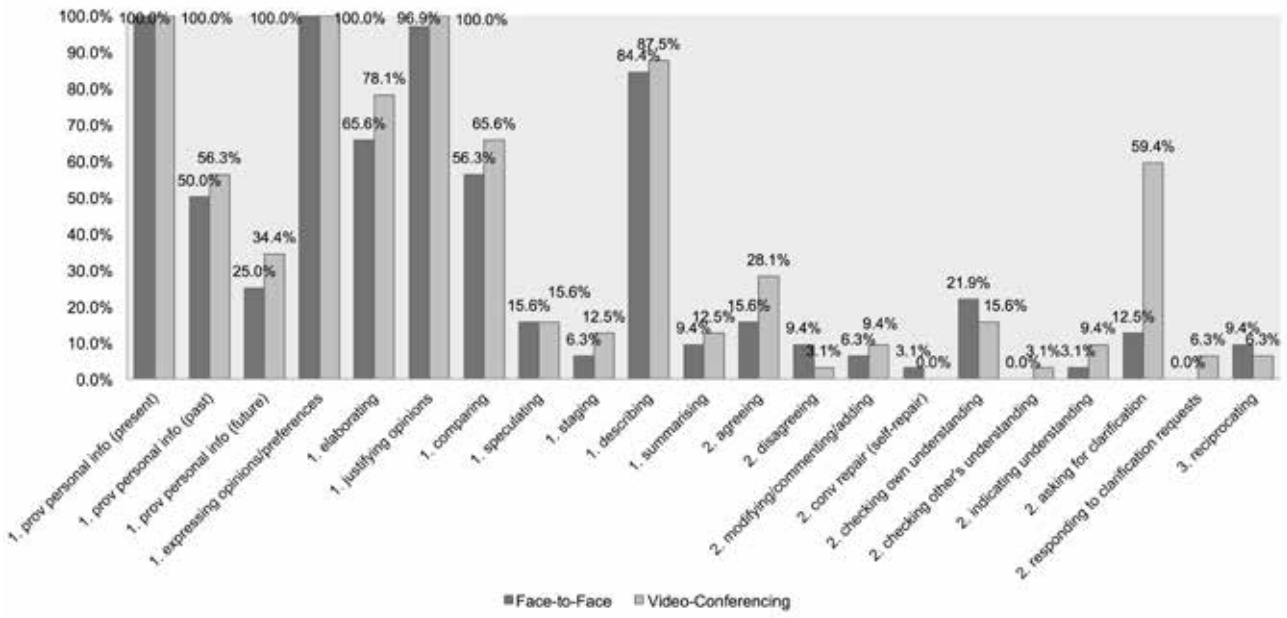
To sum up, the MFRM analysis using group anchoring of examinees provided information which complements and reinforces the results from the CTT analysis. The results of both the 5- and 4-facet analyses indicate little difference in difficulty between the two modes. Lack of misfit is associated with uni-dimensionality (Bonk and Ockey, 2003) and by extension can be interpreted as both delivery modes in fact measuring the same construct.

5.2 Language function analysis

This section reports on the analysis of language functions elicited in the two delivery modes, in order to answer RQ1b: *Are there any differences in linguistic output, specifically types of language function, elicited from test-takers under face-to-face and video-conferencing delivery conditions?*

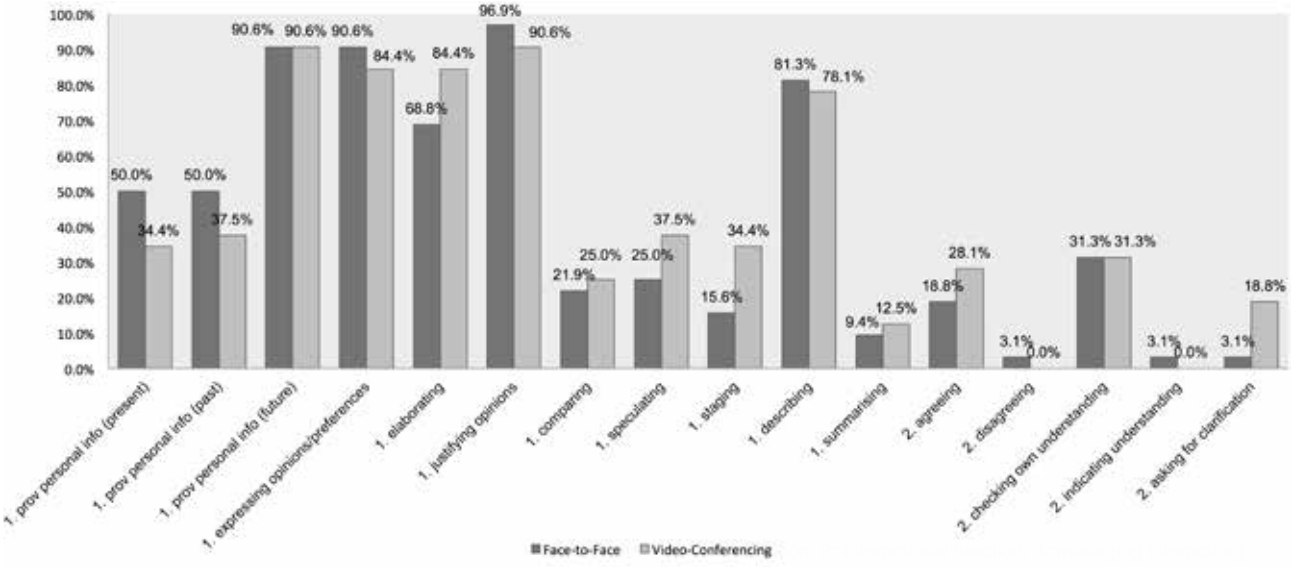
Figures 9, 10 and 11 illustrate the percentage of test-takers who employed each language function under the face-to-face and video-conferencing delivered conditions across the three parts of the IELTS test. For most of the functions, the percentages were very similar across the two modes. It is also worth noting that more advanced language functions (e.g. speculating, elaborating, justifying opinions) were elicited as the interviews proceeded in both modes, just as the IELTS Speaking test was designed to do, which is encouraging evidence for the comparability of the two modes (Appendix 6 visualises the similar shifts in function use between the two modes).

Figure 9: *Language functions elicited in Part 1*



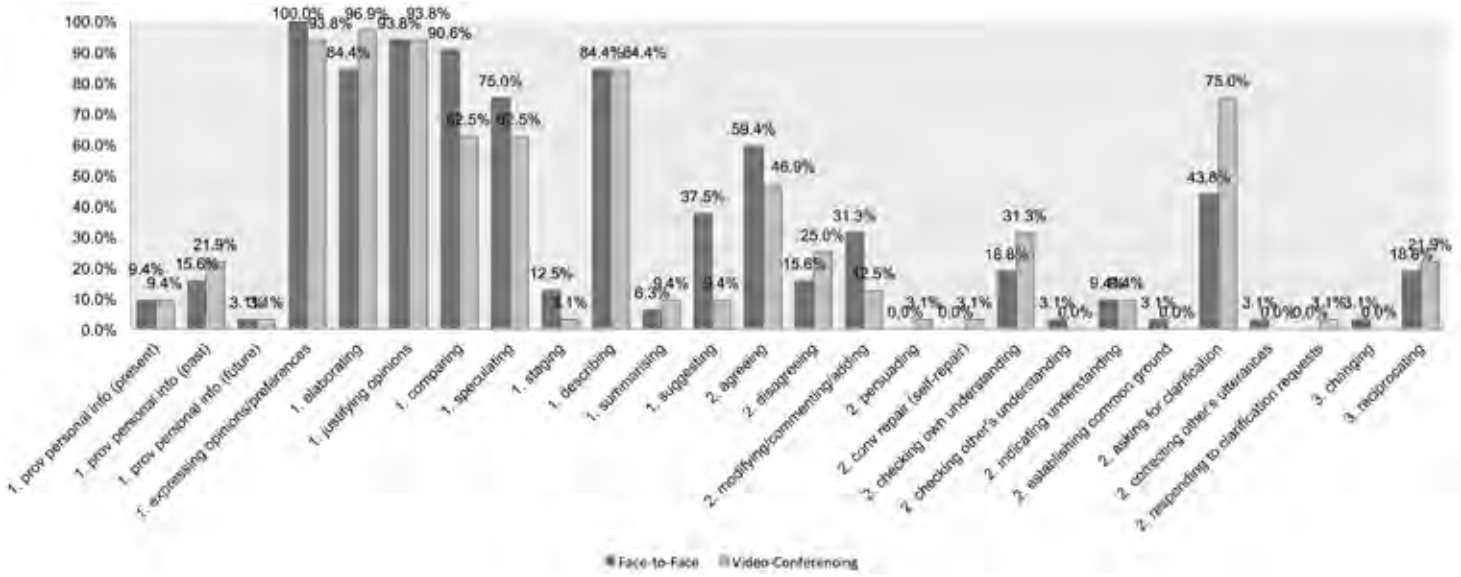
(1=informational function, 2=interactional function, 3=managing interaction)

Figure 10*: Language functions elicited in Part 2*



(1=informational function, 2=interactional function, 3=managing interaction)

Figure 11: *Language functions elicited in Part 3*



(1=informational function, 2=interactional function, 3=managing interaction)

As shown in Table 8, there were five language functions that test-takers used significantly differently under the two test modes. The effect sizes were small or medium, according to Cohen's (1988) criteria (i.e., small: r=.1, medium: r=.3, large: r=.5) (see Appendix 7 for all statistical comparisons). It is worth noting that these differences emerged only in Parts 1 and 3. There was no significant difference in Part 2, indicating that the two delivery modes did not make a difference for individual long turns.

Table 8: *Language functions and number of questions asked in Part 3 (N=32)*

[Part] Function	Test mode	Median	Mean	SD	Z (df=31)	Sig. (2-tailed)	Effect size (r)
[Part 1] asking for clarification	Face-to-face	.00	.13	.34	-3.638	.000	-.455
	Video-conferencing	1.00	.59	.50			
[Part 3] elaborating	Face-to-face	1.00	.84	.37	-2.000	.046	-.250
	Video-conferencing	1.00	.97	.18			
[Part 3] comparing	Face-to-face	1.00	.91	.30	-2.714	.007	-.339
	Video-conferencing	1.00	.63	.49			
[Part 3] suggesting	Face-to-face	.00	.38	.49	-2.714	.007	-.339
	Video-conferencing	.00	.09	.30			
[Part 3] modifying/ commenting/adding	Face-to-face	.00	.31	.47	-2.121	.034	-.265
	Video-conferencing	.00	.13	.34			
[Part 3] asking for clarification	Face-to-face	.00	.44	.50	-2.496	.013	-.312
	Video-conferencing	1.00	.72	.46			
N of Part 3 Qs asked	Face-to-face	5.00	4.56	1.19	-1.827	.068	-
	Video-conferencing	4.00	4.09	1.49			

More test-takers asked *clarification questions* in Parts 1 and 3 of the test under the video-conferencing condition. This is congruent with the examiners' and test-takers' questionnaire feedback (see Sections 5.3 and 5.4) in which they indicated that they did not find it always easy to understand each other due to the sound quality of the video-conferencing tests. Due to poor sound quality, test-takers sometimes needed to make a clarification request even for a very simple, short question as in Excerpt (1) below.

Excerpt (1) E: Examiner C, C: S14, Video-conferencing

1 E: do you like ( )?
2→ C: sorry?
3 E: <do you like (.) travelling?>
4 C: <do I like travelling.> I wouldn't say it's uh my favourite thing but er (.) it's it's good
5 to go out er out of where you are living once in a while yeah.

Under the video-conferencing condition, more test-takers *elaborated* on their opinions. This is in line with the examiners' reports that it was more difficult for them to intervene appropriately in the video-conferencing mode (see Section 5.4). As a consequence, test-takers might have provided longer turns while elaborating on their opinions. Excerpt (2) illustrates how *S30* produced a relatively long turn by elaborating on her idea in Part 3. During the elaboration, *Examiner A* refrained from intervening but instead nodded silently quite a few times. The non-verbal back-channelling seemed to have encouraged S30 to continue with her long turn. This is consistent with previous research which suggested that the types and amount of back-channelling could affect the production of speech (e.g. Wolf 2008). However, while such increased production of long turns might look positive, it could potentially be problematic as Part 3 of the test is supposed to elicit interactional language functions, as well as informational language functions.

Excerpt (2) E: Examiner A, C: S30, Video-conferencing

1 E: what kinds of difficulties do travellers experience today?
2 C: erm I think there is not many difficulties for travelling, but
3→ for example er there is lots of people that they are afraid of erm taking a plane,
4 E: ((nodding))
5 C: but I really think that it's not as erm (.) bad as they think >because it's really safe<
6 to go by plane [everywhere
7 E: [((nodding))
8 C: lo- the distance are much shorter than going by car or [bus (.) or by train
9 E: [((nodding))
10 C: erm (.) I think by plane is a safe option for everyone.

The three functions *comparing*, *suggesting* and *modifying/commenting/adding* were more often used under the face-to-face condition. As expressed in test-takers' interviews, some of them thought that relating to the examiner during in the video-conferencing test was not as easy as it was in the face-to-face test. This might explain why test-takers were able to use more *suggesting* and *modifying/commenting/adding*, both of which are interactional functions, under the face-to-face condition. Excerpt (3) shows very interactive discourse between *Examiner C* and *S32* under the face-to-face condition. In the frequent, quick turn exchanges, *S32* demonstrated many language functions including the three functions, comparing (line 21), suggesting (lines 11, 19–20) and commenting (line 3).

Excerpt (3) E: Examiner C, C: S32, Face-to-Face

1 E: let's talk now about classroom (.) erm how important is it for teachers to make feel
2 their students that they're doing well or they've done well?
3→ C: (1.0) uh huh ok .hhh yeah I think that is a (.) good point er:=
4 E: =should they give a reward?
5 C: no=
6 E: =no rewards=
7 C: =uh sorry?
8 E: should teachers give rewards to [students?
9 C: [yes of course.
10 E: what sorts of rewards?
11→ C: they can say good, [good, excellent, excellent, yeah (.) er [if if something is good
12 E: [uh hu [what about certificates or prize?
13 C: n(h)o=
14 E: =why not?=
15 C: =why not (.) because erm that is er not polite
16 E: uh huh
17 C: not formal you know? because college is er not market, college is er huh
18 E: er OK so it's not appropriate= =uh huh [let's talk about
19→ C: =not appropriate= [just just just they can talk and show
20→ they they something they are happy for you and they are happy uh for your progress or
21→ something that is will be more better than (.) this ways.

The last row of Table 8 shows the total number of questions asked in Part 3 of the test. It was decided to count the number, as all examiners mentioned in their verbal report sessions that they had to slow down their speech and articulate their utterances more clearly under the video-conferencing delivered condition. One examiner also added that, as a consequence, she might have been able to use fewer questions in Part 3 in the video-conferencing mode (see Section 5.4). However, although the descriptive statistics showed that more questions were asked under the face-to-face condition (4.56 for face-to-face and 4.09 for video-conferencing), there was no significant difference in the total number of questions used by examiners between the two modes (Z (31) =-1.827, p=0.068).

5.3 Analysis of test-taker interviews

This section describes results from the test-taker feedback interviews to respond to *RQ1c: What are test-takers' perceptions of taking the test under face-to-face and video-conferencing delivery conditions?*

Table 9: *Results of test-taker questionnaires (N=32)*

About each test mode (f2f=face-to-face, vc=video-conferencing)							
	Test mode	**Median**	**Mean**	**SD**	**Wilcoxon test**		**Effect size** (r)
					Z (df=31)	**Sig.**	
Q1 + Q3: Did you understand the examiner? (1. Never – 3. Sometimes – 5. Always)	f2f	5.00	4.72	.46	-4.096	.000	.512
	vc	4.00	3.72	.77			
Q2 + Q4: Did you feel taking the test was… (1. V difficult – 3. OK – 5. V easy)	f2f	4.00	3.84	.85	-3.048	.002	.381
	vc	3.00	3.13	.83			

Comparison of the two test modes: frequency (%)			
	Face-to-face	**Video-conferencing**	**No difference**
Q5: Which speaking test made you more nervous – the face-to-face one, or the one using the computer?	9 (28.1%)	**15 (46.9%)**	8 (25.0%)
Q6: Which speaking test was more difficult for you – the face-to-face one, or the one using the computer?	4 (12.5%)	**21 (65.6%)**	7 (21.9%)
Q7: Which speaking test gave you more opportunity to speak English – the face-to-face one, or the one using the computer?	**16 (50.0%)**	6 (18.8%)	10 (31.3%)
Q8: Which speaking test did you prefer – the face-to-face one, or the one using the computer?	**27 (84.4%)**	3 (9.4%)	2 (6.3%)

As summarised in Table 9, test-takers reported that they understood the examiner better under the face-to-face condition (mean: 4.72) than the video-conferencing condition (mean: 3.72), and the mean difference was statistically significant (Q1 and Q3). They also felt that taking the test face-to-face was easier (mean: 3.84) than taking the test using a computer (mean: 3.13), and again the difference was statistically significant (Q2 and Q4). The effect sizes for these significant results were large (r=.512) and medium (r=.381), respectively, according to Cohen's (1988) criteria (i.e., small: r=.1, medium: r=.3, large: r=.5).

Test-takers' comments on these judgements included the following.

Understanding the examiner (Q1 and Q3)

- *[Face-to-face – always] I asked the examiner to repeat one question, but apart from that, I was able to understand him perfectly. Having good eye contact and seeing his facial expressions and postures helped me to understand him better (S06).*
- *[Video-conferencing – sometimes] The sound quality was not as good as the face-to-face mode, and I needed to focus more on what the examiner was saying (S25).*
- *[Video-conferencing – sometimes] The quality of sound wasn't always good (S24).*

Test difficulty (Q2 and Q4)

- *[Face-to-face – easy] The face-to-face test was easy. I felt it was more real (S25).*
- *[Video-conferencing – OK] I was able to understand the examiner most of the time, but I felt rather tense as I felt somewhat reluctant to ask the examiner to repeat questions via computer (S23).*
- *[Video-conferencing – OK] Taking the test using a computer was not as easy as face-to-face. I saw myself on screen and it made me more nervous (S29).*

The last four questions (Q5 – Q8) related to comparisons between the two test delivery modes.

Of the test-takers, 46.9% reported that they felt more nervous when they were taking the test using a computer, while 28.1% of them felt that the face-to-face test made them more nervous and 25.0% did not find any difference (Q5). Test-takers' comments on these judgements included:

- *[Face-to-face] I'm shy and I always become nervous when I speak to a person face-to-face. Instead, the computer test made me more relaxed (S07).*
- *[Video-conferencing] I can develop a better relationship with the examiner face-to-face. I was able to feel the examiner's rapport directly, which made me more comfortable and encouraged. But I couldn't feel the same level of friendliness from the computer screen (S22).*
- *[No difference] I was nervous in both modes, but for different reasons. The face-to-face mode made me nervous as the examiner was in front of me. The computer mode made me nervous as the sound quality was not good and the image didn't synchronise with the sound (S25).*

Regarding the test difficulty, 65.6% of the test-takers thought that taking the test using the computer was more difficult, while 12.5% felt the face-to-face test was more difficult and 21.9% did not find any difference (Q6). Test-takers' comments included:

- *[Face-to-face] I felt that the face-to-face test was more difficult, as I became more nervous (S07).*
- *[Video-conferencing] I'm used to chatting with my friends on screen, but not to having formal conversation with the examiner. So, I felt the computer test was more difficult (S04).*
- *[Video-conferencing] I like to talk to people directly and I found it easier to talk to the examiner face-to-face (S06).*
- *[No difference] Both tests were OK. I didn't find any difference between the two modes (S27).*

Half (50%) of the test-takers felt that the face-to-face test gave them more opportunity to speak English, while 18.8% felt that they were able to speak more in the computer mode, and 31.3% did not find any difference (Q7). Test-takers' comments included:

- *[Face-to-face] I felt that I was able to speak more face-to-face, because I was able to understand when to speak and chip in (S17).*
- *[Video-conferencing] On the computer, it doesn't matter what I say as the*

interviewer is not in front of me. I felt freer and was able to speak more. Perhaps I made more mistakes, though (S05).

- *[Video-conferencing] I felt the computer-delivery mode gave me more opportunities to speak, because the use of body language was limited, so I couldn't rely on gestures to complement my language (S27).*
- *[No difference] I didn't feel any difference. I think I spoke about the same amount in both modes (S28).*

Finally, when asked about their preferred test mode, 84% of the test-takers reported that they preferred the face-to-face test, while 9.4% preferred the computer-delivered test and 6.3% did not have any preference (Q8). Test-takers' comments included:

- *[Face-to-face] The computer mode was more like a 'test' and the face-to-face mode was more like real non-test communication (S16).*
- *[Face-to-face] I like to see the examiner's body language while speaking to her. In the computer mode, I was able to see only part of her top body (S29).*
- *[Face-to-face] I preferred the face-to-face mode. The examiner gave me lively encouragement. In the computer mode, the communication didn't feel real (S10).*
- *[Face-to-face] I liked the face-to-face test better. It was clear whether the examiner understood me or not. But in the computer test, I was not sure if she understood me, because I couldn't see her facial expressions very well (S05).*
- *[Video-conferencing] I first thought I wouldn't like the computer-delivery mode very much, but after I've taken both modes, I actually preferred the computer mode. Because, seeing the examiner face-to face felt very direct and I felt more stressed (S13).*
- *[No difference] The both modes were the same. It's still speaking to someone (S20).*

5.4 Analysis of observers' field notes, verbal report sessions with examiners, examiners' written comments, and examiner feedback questionnaires

We have so far reported on test-takers' data in relation to their output language, test scores and feedback interviews. We now move on to presenting results on examiners, including: examiners' test administration behaviour (RQ2a), their rating behaviour (RQ2b), and their perceptions of examining (RQ2c) under the two delivery conditions. These points will be discussed one by one following analyses of data from four different sources: retrospective verbal report sessions with examiners; observers' field notes; examiners' written comments; and examiner feedback questionnaires.

RQ2a: Are there any differences in examiners' test administration behaviour (i.e. as interlocutor) under face-to-face and video-conferencing delivery conditions?

Analysis of the verbal report audio data highlighted several aspects of test administration that appeared to differ across the face-to-face and video-conferencing delivery conditions. By listening to the audio recordings of the examiners' verbal reports and making extensive notes on the topics and reactions mentioned by the examiners, the research team was able to analyse examiner comments according to two main categories:

- differences reported by examiners in their interaction with test-takers under the face-to-face and video-conferencing condition
- issues that are perceived as specific to administering a speaking test via a computer-delivered mode.

5.4.1 Differences reported by examiners in their interaction with test-takers under the face-to-face and video-conferencing condition

Examiners commented on differences in the following aspects of their own (i.e. the examiner's) interactional behaviour:

- role and frequency of examiner response tokens, e.g. nodding, back-channelling
- rate and articulation of examiner speech
- effect of examiner intonation
- use of gestures by examiners (and awareness of gestures used by test-takers)
- issues related to turn-taking management
- requests for clarification.

These interactional aspects on the part of the examiner are discussed more fully below, supported by examples from the researchers' notes on the verbal report data.

Some examiners were observed to use *nodding and back-channelling* less in the video-conferencing condition, while others used it more. A number of possible reasons for this were hypothesised.

- *Examiner D tended to nod more on face-to-face mode to facilitate the test-taker, but on computer mode, he didn't do it very much due to the delays in video transmission and for fear that doing so may delay it further; he felt that under the face-to-face condition, he was more himself and more natural; he could interject naturally 'Oh really?' and interpreted more naturally.*
- *Examiner C tried to be more human-like on computer with lots of nodding and smiling.*
- *Examiner A nodded much more on computer; she said this might have been because the test-taker asked for repeated instructions and questions; she was afraid that he might think she's not getting what he said if she doesn't show understanding (and verbally back-channelling isn't recommended).*

It seems possible that nodding may be deliberately constrained by an examiner for technical reasons to avoid video transmission delay; alternatively, it may be deliberately used by an examiner as an interactional strategy to compensate for the lack of physical proximity that results in the video-conferencing condition.

Three examiners commented on *slower and more articulate speech* as a noticeable feature of their own interaction in the video-conferencing condition.

- *Examiner D needed to articulate each word more slowly and he needed to slow down his speech.*
- *Examiner C needed to articulate each syllable very clearly and to speak slowly on computer; S32 doesn't need that level of graded language, but she had to do so to make sure so he understands her; this happened for other test-takers as well, and this would prevent her from giving Band 9 – as she cannot really examine test-takers who can keep up with her when she speaks fast to see their limit.*
- *Examiner A spoke more slowly and clearly on computer because the test-taker was leaning towards the computer with one ear closer to the screen, and also because there were delays in transmission; with video-conferencing Examiner A needed to speak slowly and her speech sounds unnatural, she feels.*

The video-conferencing condition appears to provoke in examiners a sense that they need to speak more slowly and articulate more clearly, in order to ensure that test-takers understand them, and possibly to mitigate any perceived technical challenges, e.g. transmission delay or poor sound quality.

One examiner comment in the verbal report data referred to *intonation*.

Examiner D uses 'Why?' with an intonation which doesn't make it sound like interrogation – which is easy to convey in face-to-face, but with video-conferencing, due to sound quality and transmission speed, he's not sure if that 'subtlety' is conveyed; he's done double-marking for test-takers with a jagged profile and has noticed that the intimidating intonation of 'why?' by the examiner can affect the whole interview.

This single comment suggests at least one examiner's awareness of the potential for subtleties of tone and implicature to be distorted when the interaction takes place via computer rather than face-to-face.

A large number of examiner comments concerned *gestures and body language*.

- *Examiner D uses gestures with the topic card (i.e. patting the topic card when saying "Please don't write anything on the topic card"; finger-pointing the instruction as he explains; showing his hand when saying "please start talking"), but cannot do so with computer; he thinks that such gestures won't affect the score, but helps test-takers (especially weaker ones) a lot; S20 wasn't looking at the screen as much as she did face-to-face; her eyes often looked up and sideways a lot (and it wasn't because of the location of the camera lens on the iPad); there's no real eye contact on computer; there was one occasion that Examiner D said something but she wasn't looking at the screen, so she kept talking without noticing that she was addressed; she is also swinging back and forth a lot – which made Examiner D wonder if having an examiner on the other end is meaningful.*
- *Examiner D doesn't think he can use gestures on computer as effectively as he could under the face-to-face condition; when S16 was struggling to answer a question in face-to-face, Examiner D turned around to take the pressure off from her; S16 moved forward to show that she didn't understand Examiner D face-to-face; but Examiner D is not sure if he could pick this up under the video-conferencing condition; Examiner D's question on 'how important...?' in the face-to-face mode, Examiner D sensed that S16 didn't understand the question based on her facial expressions and repeated the question; Examiner D's introduction of the new topic 'Being the best' – S16 changed facial expressions signalling that she didn't understand the topic; he feels that he wouldn't be able to get such signs through computer; in his second watching of the video-conferencing video, it was very obvious from her face that she didn't understand some questions, but Examiner D didn't pick it up when he was examining; this is perhaps because he was focusing more on sound than her facial expressions.*
- *Examiner C comments she cannot see the test-taker's hands on computer; therefore there's limited information available – of course examiners are supposed to rate what the test-taker says and not the body language or gestures, but it affects the examiners' impression… cannot get the same rapport on computer as face-to-face interviews; the test-taker used her hands much more face-to-face than on computer; she was also smiling more and seemed more relaxed.*
- *Examiner C felt that her gestures in the face-to-face mode might have been too much and distracting; this wouldn't happen on computer; interlocutor frame would take a back seat in the computer mode; test-takers see no mess on the examiner's table; the use of non-verbal response tokens (nodding, smiling) should be standardised, especially for the computer mode; best rapport can be obtained by eyes, and giving smiles motivates test-takers; she feels that this could be lost under the computer condition; she was able to make good eye contact face-to-face; but with video-conferencing, the amount of the test-taker's body language was reduced, and it was not easy to make the exact eye contact – if they look at each other, they never have perfect eye contacts (as cameras are not in the middle of the screen); there was a lot more ambiguity under the computer mode;*

in the face-to-face test, she could tell subtle differences in the test-taker's facial expressions to signal his request for clarification; but she is not sure if she could do the same on computer.

- *Under the face-to-face condition, Examiner A pointed out sub-questions on the prompt card using her finger; examiners are not allowed to say anything to facilitate test-takers' Part 2 speech (apart from 'Can you tell me something more about this?') but they are allowed to point to sub-questions; in the video-conferencing session, S19 again stopped talking but Examiner A couldn't do the finger pointing on computer.*
- *Examiner A does lots of Skype lessons, so she is used to making eye contact (although it is not possible to make a perfect eye contact on computer); but S19 was looking away in the computer mode.*
- *Some test-takers moved a lot under the computer mode, and their movements were rather distracting to Examiner B.*

It is clear that all four examiners were sensitive to the impact of gesture, movement and body language in the interaction and how this appeared to differ across the two conditions. Specific mention was made of the following gestures used by both test-takers and examiners in the interaction:

- use of hands and fingers (to point, to emphasise, to initiate)
- facial expressions, e.g. smiling, frowning (to encourage, to query)
- nodding
- upper body movement, e.g. swinging/leaning forwards, shifting in one's chair
- eye contact to establish/maintain rapport.

The consistent message from the examiner comments seems to be that, in the video-conferencing condition, examiners can find it harder to use natural gestures as part of their own interaction, and to perceive/read similar gestures when these are used by the test-taker. Limited eye contact in the video-conferencing condition may limit rapport with the test-taker. Furthermore, some simple gestures which routinely support the smooth administration of the face-to-face test may simply not be possible in the video-conferencing condition, e.g. finger-pointing; alternatively, such gestures may be distracting due to transmission delay. There may also be a tendency for some examiners to exaggerate nodding or smiling in an attempt to compensate for the latter.

Examiner C's comment that the examiner's interlocutor frame 'would take a back seat in the computer mode' is an interesting one and may be worthy of further investigation. We can only speculate on exactly what Examiner C had in mind when making this comment but it may reflect a concern that, in the computer mode, the examiner would be forced to deviate more from the interlocutor frame (i.e. examiner script) to compensate for some aspects of the computer mode. This seems to be an important consideration which may warrant attention in any further studies, since adhering to the examiner script is an essential aspect of the validity of the IELTS Speaking test. It should, however, be noted that findings from the literature as to the way examiners' language affects the language test-takers produce are not entirely consistent (cf. Brown and Hill 1998; O'Sullivan and Yang 1996).

A number of examiner comments related to the challenges posed by the video-conferencing condition for the management of turn-taking.

- *Examiner D nods more and smiles more in face-to-face; he also sits back after giving instruction to show that it's now the test-taker's turn; he also mentioned that it is difficult to figure out why the test-taker leaned into the computer (whether she couldn't hear well or didn't understand, etc.); a tiny sound delay on computer made it difficult to turn-take, as he cannot judge whether test-takers intend to continue or not; he couldn't hear her sometimes, which generated an awkward*

sequence: S16 – 'Can you repeat..?' Examiner D – 'What did you say?' S16 – 'What did you say?'

- *Examiner C commented that in the computer mode it takes longer to do the examiner frame (usually the topic frame takes 9 secs face-to-face, but on computer it took 12–13 secs); turn-taking is different – due to the sound quality and slow communication; because turn-taking is slower, a smaller amount of language (evidence) is elicited during the time allowed.*
- *Examiner A commented that, unlike face-to-face, it was difficult to take turns appropriately on the computer; it was hard to judge when S19 would stop talking.*

These comments from three of the four examiners highlight the increased challenge for turn management posed by the video-conferencing condition, e.g. it may be more difficult to signal the interlocutor's turn, or to determine when the interlocutor's turn is complete. This potentially slows down the turn-taking rate and it may result in a reduced, as well as a somewhat stilted, language sample being gathered within the time available.

One examiner reported receiving more *clarification questions* in the video-conferencing condition. This examiner comment is confirmed by the results of the functional analysis discussed earlier (Figures 9 to 11).

5.4.2 Issues that are apparently specific to administering a speaking test via a video-conferencing mode

In their verbal reports, the examiners raised a number of issues that were linked to the nature of a computer-based speaking test which they felt impacted negatively on their own role as the facilitator in the test, as well as on the smooth running of the speaking test. Specific comments were made regarding:

- the negative effects of delayed video transmission
- the way the test-taker can impact on the sound quality
- the need to control the direction of the interview.

Three examiners commented on how the *delayed transmission* in the video-conferencing condition made it difficult to stop a test-taker from continuing to talk, or to intervene and help a test-taker if needed.

- *It is easier to stop or interrupt the test-taker face-to-face; due to the delay in video transmission, it is difficult to do it on computer; it is also difficult to catch if the test-taker wants to talk more (from their subtle facial expression); Examiner D mentioned that such small clues help the smooth running of the interviews.*
- *Due to the delay in transmission, stopping the test-taker can be problematic; sometimes Examiner C had to stop them by using her hand (like a policeman), which she didn't feel was very nice; also the use of hands by the examiner was very limited; Examiner C had to hold the speaker close to her at the same time trying not to make much noise to hear the test-taker better.*
- *Interrupting and turn-taking were difficult with video-conferencing; 'sorry to interrupt you'; also, finding the right timing to ask a rounding-off question was difficult; to end the conversation Examiner C wouldn't normally use a hand to stop test-takers while saying 'thank you', but she thought it might be necessary on computer?*
- *Examiner A commented it was easier to stop or interrupt the test-taker face-to-face; also, Examiner A helped the test-taker more face-to-face because he was looking more nervous (e.g. wrapping up what he wanted to say like 'so you mean the minority is…')*
- *To stop the test-taker's speech, Examiner B often signals it both verbally and non-verbally (putting a hand forward); with video-conferencing she needed to make the gesture closer to her face to be captured on the screen.*

- *Examiner B rephrased a question because of the test-taker's pauses and her facial expressions, without an explicit clarification request – but is not sure if she could get similar information under the video-conferencing condition; due to a slight sound delay Examiner B couldn't intervene to help the test-taker at the exact point she wanted to, which she would have done face-to-face.*

Two examiners commented on how in the video-conferencing condition, the *test-taker sometimes impacted negatively, albeit unintentionally, on the sound quality* of their own speech.

- *The test-taker was often putting her hand near her mouth, which obscured the sounds; doesn't happen face-to-face.*
- *The test-taker sometimes covered her face, face-to-face; if she had done this in the video-conferencing condition, it'd be very difficult to understand her.*

One examiner commented on the need to *control the direction of the interview* in certain circumstances.

- *Examiner C felt that she needed to provide more language support under the video-conferencing condition to make sure that students don't go off topic.*

Analysis of observer field notes provided corroborating evidence for all the issues and interactional features discussed above that relate to the examiners' test administration behaviour across the two conditions.

RQ2b: Are there any differences in examiners' rating behaviour when they assess test-takers under face-to-face and video-conferencing delivery conditions?

The verbal report data contained insightful comments from examiners regarding their experience of rating test-taker spoken performance in the two conditions. These data were gathered as the four examiners watched videos of themselves rating a small number of test-takers under both conditions. It was clear from the analysis that *sound quality* had impacted significantly on examiners' ability to rate speech output in a consistent manner. Poor sound quality in the video-conferencing condition seems to have forced the examiner to allocate extra resources to certain aspects of the interaction (careful listening, coping with delayed transmission), possibly at the expense of their attentional capacity for the actual activity of rating. Evidence of this can be seen in the following excerpts from the data where examiners reflected on their experience of rating in the face-to-face and video-conferencing conditions.

- *Examiner D mentioned he noticed S20's lexical and grammatical errors more face-to-face; he assumes it was because he was more relaxed face-to-face and was able to concentrate on assessing, rather than on listening very carefully on computer due to poor sound quality; due to the delay in transmission, probing the test-taker by speeding up the delivery of questions in Part 3 is difficult; of course, listening is not difficult, but the number of Qs that can be asked is affected.*
- *With video-conferencing, Examiner D couldn't really hear her word endings, or other micro-phonologic features…he didn't want to give the benefit of the doubt, so scored 5.0 for Pronunciation and it could have been even 4.0 (6.0 in face-to-face mode); he gave 5.0 for Grammar (6.0 in face-to-face mode).*
- *Because the sound quality wasn't great, Examiner C was unable to judge whether the test-taker said 'paper' instead of 'pepper' due to L1 influence, or it was just the poor sound; good sound quality is crucial for reliable assessment; if technical issues prevented L1-specific influences (such as schwa), or repetition, ums and ers, the ratings can be affected.*
- *Due to the delay in transmission, probing the test-taker by speeding up the delivery of questions in Part 3 is difficult; might affect the rating because there's less evidence; big problems with the delay in video transmission and occasional*

skipping of utterances; delay was noticeable; synchronising is crucial for accurate assessment – lip-reading and processing input in real time; gave 5.0 on computer for test-taker's use of simple structures, but gave 6.0 f2f – but maybe the test-taker was using simple sentences because the transmission was delayed and choppy and she wanted to convey what she meant quickly.

- *Examiner C felt that score 'inflation' happened in face-to-face; the test-taker's pronunciation is clear, but it should have been 6.0 instead of 7.0 (which she originally gave during the face-to-face test).*
- *Pronunciation was not clear on computer; 'campus' sounded like 'compost'; I 'planet' (planned) to go there; she couldn't hear phonologic/phonetic features the face-to-face test clearly.*
- *Sometimes she couldn't hear him – she gave 5.0 for Lexis on video-conferencing (but 6.0 in the face-to-face mode), as she didn't want to give him the benefit of the doubt.*
- *The sound quality wasn't great – the test-taker's voice sounded muffled.*
- *Pronunciation is definitely clearer in face-to-face; she wondered if this test-taker was at Band 7.0 on pronunciation towards the end of face-to-face session – it was easier to hear stress, timing and rhythm.*
- *Examiner A feels she could or perhaps should have given a higher score on Grammar in the video-conferencing mode (4.0 on video-conferencing, 5.0 face-to-face) but she couldn't hear S19 during the exam.*
- *The face-to-face mode provided more accurate language samples to rate for Grammar and Lexis – it was clear that she said 'I like' rather than 'I like it' and 'I don't like' rather than 'I don't like it' – but with video-conferencing, it was not clear whether she really missed 'it' or Examiner B couldn't hear due to sound quality; Examiner B feels that the face-to-face rating is more accurate and that she perhaps overrated S07's Grammar on computer (video-conferencing at 7.0, face-to-face at 6.0).*
- *Examiner B concentrated more on Pronunciation under the video-conferencing condition, due to the slight delay of sounds and more patchy speech; Examiner B wasn't sure whether she needed extra efforts due to the test-taker's pronunciation problem or due to the technology; S07 had some L1 Spanish influences such as 'estudy' for 'study; also, visual information is very important to rate Pronunciation, but due to an unclear view of the test-taker and her picture and sound not exactly synchronised, Examiner B felt that she perhaps overrated the test-taker's pronunciation under the video-conferencing condition (video-conferencing at 6.0 and face-to-face at 5.0).*
- *Examiner B, however, felt that she was able to rate test-takers more objectively under the video-conferencing condition; as an examiner trainer, she is used to rating audio-recorded performances and she felt that rating video-conferencing performances was somewhat easier.*

A persistent theme in these comments is the perceived negative impact of poor sound quality in the video-conferencing condition on examiners' judgements of *Pronunciation* and *Grammar*; the impact seems less pronounced where judgements of *Lexis* are concerned. There were no explicit comments regarding the *Fluency* criterion, although, judging from some comments, delayed transmission did affect the 'confluence' (McCarthy 2006) of the co-constructed interaction if not the individual test-taker's fluency and coherence, especially in Part 3 where fewer questions could be asked by the examiner resulting in less topic coverage.

Additional data linked to rating activity was available from examiners in the form of the notes they entered onto the mark sheet during rating, under the four assessment criteria for the speaking test: *Fluency, Lexis, Grammar and Pronunciation*. The notes show that

examiners had no difficulty recording evidence for each of the four criteria across the two conditions. However, the notes in the video-conferencing condition make regular reference to technical difficulties and the problem of hearing clearly enough to make the necessary judgement, typically where Pronunciation quality was concerned, for example:

- *I had trouble hearing a lot of words [Examiner C]*
- *very difficult to rate – sound quality and delay [Examiner C]*
- *a little difficult at times to make out what some words were [Examiner D].*

It should be noted, however, that despite the examiners' comments on the sound quality and quality of pronunciation, the analysis of their scoring behaviour does not show any significant differences between the two modes of delivery.

RQ2c: What are examiners' perceptions of examining under face-to-face and video-conferencing delivery conditions?

This section presents the results from the examiner questionnaire which was completed by the four examiners involved in the study. The short questionnaire was designed to gather examiner perceptions of administering and rating the speaking test under the two conditions – face-to-face and video-conferencing delivery. It included both Likert scale responses (1=Strongly disagree to 5=Strongly agree) and free text boxes to capture additional comments. Questions and responses focused upon examiner perceptions of the ease of *administering* the speaking test in the face-to-face and video-conferencing conditions, the ease of *rating* the speaking test in the two conditions, and any *comparisons* across the test modes.

Table 10: *Examiner perceptions concerning ease of administration (N=4)*

	Test mode	Min	Max	Mean (SD)
Comfortable in overall administration	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	2	5	3.75 (1.26)
Ease of administering Part 1	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	2	5	4.00 (1.41)
Ease of administering Part 2	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	3	5	4.00 (0.82)
Ease of administering Part 3	Face-to-face	2	5	4.00 (1.41)
	Video-conferencing	2	5	4.00 (1.41)
Ease of administering interlocutor frame	Face-to-face	3	5	4.50 (1.00)
	Video-conferencing	2	5	3.75 (1.89)

In terms of ease of *administration*, mean values for the face-to-face mode were in almost all cases higher (except for Part 3), suggesting that examiners felt more comfortable with this mode for the speaking test and perceived it as marginally easier to deliver than the video-conferencing condition. Interestingly, however, the difference was marginal and it could be seen as encouraging that examiners who were not necessarily as familiar with the computer-mediated approach to speaking assessment reported as positively on the experience as they did.

Table 11: *Examiner perceptions concerning ease of rating (N=4)*

	Test mode	Min	Max	Mean (SD)
Comfortable overall in rating performance	Face-to-face	3	5	4.50 (1.00)
	Video-conferencing	1	5	3.50 (1.73)
Ease of applying Fluency and Coherence scale	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	1	5	3.75 (1.89)
Ease of applying Lexical Resource scale	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	3	5	4.25 (0.96)
Ease of applying Grammatical Range and Accuracy scale	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	2	5	4.00 (1.41)
Ease of applying Pronunciation scale	Face-to-face	4	5	4.75 (0.50)
	Video-conferencing	1	5	4.00 (1.41)
Confidence in accuracy of rating	Face-to-face	3	5	4.00 (0.82)
	Video-conferencing	1	5	3.50 (1.73)

In terms of *ease of rating*, mean values for the face-to-face mode were in all cases higher, again suggesting that examiners felt more comfortable rating in this mode of the speaking test and perceived it as marginally easier to rate than in the video-conferencing condition. Once again, the difference was marginal; despite not being as familiar with the computer-mediated approach, examiners nonetheless reported positively on the experience.

Table 12: *Comparison of the two modes*

	Face-to-face	Video-conferencing	No difference
Which mode of speaking test did you feel more comfortable with?	1	2	1
Which mode of speaking test did you feel was easier for you to *administer*?	2	0	2
Which mode of speaking test did you feel was easier for you to *rate*?	3	1	0
Which mode of speaking test do you think gave a better chance for the test-taker to demonstrate their level of English language proficiency?	1	1	2
Which speaking test did you prefer?	2	1	1

The questionnaire responses from the four examiners suggest that, with regard to ease of administration, the computer-based video-conferencing speaking test may be the

slightly dis-preferred mode; with regard to ease of rating, the face-to-face mode seemed to be the preferred mode. Overall, the face-to-face mode attracted a more positive response, although a noticeable percentage of examiner responses indicated no difference across the two modes. It may, however, be worth remembering that three of the four examiners had indicated that they had not used video-conferencing previously in either teaching or examining and this may have impacted on their preferences.

The free text box responses from examiners provided explanation or justification for some of the questionnaire responses and were also consistent with themes concerning test administration raised by examiners in their verbal reports (see above). Analysis of the verbal report data also highlighted some insights which may have useful implications in two respects:

Examiner exposure to, and training in, computer-based tests

Examiner comments suggest that exposure to computer-based testing of speaking encourages familiarity and confidence with that mode.

- *As this was Examiner D's second day, he said he got more used to examining through computer and looked more on screen than on the first day (he was paying more attention to listening rather than looking at the test-taker then).*
- *Examiner A has done online teaching a lot, so she is used to understanding facial expressions and signals from students for the need to repeat questions even on computer.*

It is possible that the examiners' lack of experience with video-conferencing delivery of speaking tests impacted both on their comments and on how they responded to the questionnaire as reported in Tables 10 and 11. Familiarity and confidence in delivering speaking tests though a video-conferencing mode can presumably be enhanced through appropriate training and more experience.

Future development of the computer-based administration

Other comments highlighted practical issues that may be worth considering in any future development of the computer-based format for the speaking test. Comments relate to matters of time management and examiner note-taking activity.

- *If the topic card is shown on screen in the future, it will be good to display the preparation time (counting down) too for both the examiner and test-taker.*
- *Examiner C feels that assessing Band 9 on computer may be very difficult, if faster delivery of questions and checking comprehension is not possible.*
- *In the face-to-face mode, Examiner C can guarantee that test-takers get a correct topic card, but she cannot do this with computer (unless the topic is displayed on screen).*
- *During the preparation time in Part 2, Examiner A did not take notes and she wouldn't be tempted to do so; it's a habit not to take notes, unless she needs to remember something specific; she thinks that would make test-takers nervous.*[8]
- *Examiner B made lots of notes during the planning time in Part 2; making notes is very easy under the video-conferencing condition as test-takers cannot see what examiners are writing; they are not allowed to do that in face-to-face; sometimes Examiner B even continued to take notes after the test-taker started talking.*

These comments suggest the value of having the timing displayed on screen for both examiner and test-taker, as well as the opportunity for the examiner to make notes during their assessment.

8 In the operational face-to-face IELTS Speaking test, the examiner is trained and instructed NOT to make notes during the test to avoid creating anxiety for the test-taker who may be close enough to see and read them. The 'remote' video-conferencing condition might make it possible for the examiner to make relevant notes on performance during the test without causing anxiety to the test-taker as the examiner is not fully visible to the test-taker. However, this is an administrative detail for the test provider to determine and standardise.

6 Conclusions

This study, using a convergent parallel mixed methods design, has carried out a preliminary exploration and comparison of test-taker and examiner behaviour across two different delivery modes for the same L2 speaking test, i.e., the standard face-to-face and video-conferencing modes. The various types of data and the different methods of analysis have provided different and supplementary pictures of the two delivery modes of the test, which allowed for a more in-depth and comprehensive set of findings.

The findings for each of the six research questions in the beginning of this report are summarised in Table 13.

Table 13: *Summary of findings*

Research questions	Findings
1a: Are there any differences in **test-takers' scores** between face-to-face and video-conferencing delivery conditions?	CTT analysis shows that there were no significant differences in scores between the two modes; this finding is confirmed by the MFRM analysis.
1b: Are there any differences in **linguistic output, specifically types of language function**, elicited from test-takers under face-to-face and video-conferencing delivery conditions?	Comparing, suggesting, and modifying/commenting/adding in Part 3 were used by more test-takers in face-to-face. Asking for clarifications in Parts 1 & 3, elaborating in Part 3 were used by more test-takers using video-conferencing. No significant differences were observed in any of the other language functions.
1c: What are **test-takers' perceptions** of taking the test under face-to-face and video-conferencing delivery conditions?	Test-takers felt they understood the examiner better and that the test was easier in face-to-face mode (statistically significant). Comparing the two modes, the video-conferencing mode made test-takers more nervous, gave less opportunity to speak English, and thus was less preferred. Test-taker interview data supported this view, but there are some interesting/encouraging comments towards video-conferencing too.
2a: Are there any differences in **examiners' test administration behaviour (i.e. as interlocutor)** under face-to-face and video-conferencing delivery conditions?	Examiners reported differences in using response tokens (nodding etc.), articulation and speed of their speech, intonation, gestures, turn-taking and requests for clarification. They also reported that delayed video transmission and sound quality affected their behaviour as interlocutors (e.g. difficulty in intervening).
2b: Are there any differences in **examiners' rating behaviour** when they assess test-takers under face-to-face and video-conferencing delivery conditions?	Examiners reported that sound quality and delayed video affected the ease of rating as they had to allocate attention to some aspects using video-conferencing that they would not have had to do under the face-to-face condition, although there is no evidence that this affected the scores they awarded.
2c: What are **examiners' perceptions** of examining under face-to-face and video-conferencing delivery conditions?	Examiners reported that the face-to-face condition was easier to administer and rate.

The results of this exploratory study comparing face-to-face and video-conferencing delivery modes of the IELTS Speaking Test suggest that while the two modes are comparable in some respects, they also differ in some aspects. While the two modes generated essentially the same test score outcomes, there were some differences in

functional output and examiner interviewing and rating behaviours, although it is worth reiterating that these may have been due to the degree of examiner familiarity with video-conferencing delivery.

As such, it is recommended that before any decisions about deploying (or not) an online video-conferencing system for the IELTS Speaking test delivery are made, further analysis is carried out which (a) focuses on a range of important issues which have remained beyond the scope of this small-scale investigation (see recommendations for further research, below) and (b) seeks to confirm the findings in a larger-scale investigation.

In addition, the effect of the technical issues which were encountered (even in this tightly-managed and carefully-planned study) should not be underestimated. Zoom is a much better, more stable computer-mediated communication software than other programs such as Skype, but the technical issues of sound quality and delayed video transmission were persistent, which were repeatedly reported by both examiners and test-takers (as well as the researchers on-site), and impacted significantly on various aspects of the tests. This is a significant issue which needs to be carefully considered and addressed in any future discussions and decisions about the use of any video-conferencing system. Appendix 8 provides a brief report by a technical expert from the British Council on the technical issues during data collection, and it states strongly that more stable internet connections are required for better sound quality and that meticulous preparations at the local site are an absolute necessity for smoother administration of the video-conferencing delivered mode. It may be useful to examine further some video-conferencing sessions which had much better internet connections than others, since this would allow us to have baseline data of Zoom working at its best at this point (and so the negative influences of sound quality and delayed videos would be minimised as much as possible).

In order to explore the potential of computer-delivered IELTS Speaking tests, we recommend that further analyses of existing data are carried out to further investigate important questions/issues such as those outlined below.

Larger-scale replication and a multiple-marking design

- Replicating the study with a larger data set is essential. As mentioned in Section 5.1, this will reveal any possible differential effects of the delivery mode related to test-taker characteristics, such as age and proficiency level. It will also enable more sophisticated, accurate statistical analysis, leading to more generalisable conclusions.
- A multiple rating design which allows more rigorous MFRM analysis should be implemented in future research. The group anchoring method used in this study assumes that the groups are in effect equivalent. However, the groups in this study contained small numbers of examinees (N=8 each), which limits the generalisability of the results. Furthermore, during data collection, some test-takers who had originally agreed to take part were absent on the day, requiring replacements to be found at short notice. Completely random allocation was thus not possible, with some participants allocated on the basis of convenience and availability. The assumption of equivalence is nonetheless largely borne out by the very close mean raw scores for the four groups, but one of the groups exhibited a slightly higher mean raw score than the other groups.[9] Therefore, it is important to carry out a more rigorous MFRM study with a multiple rating design in order to confirm the results of this study.
- Since every oral test was recorded, the possibility exists of re-marking all audio and/or video recordings by presenting them randomly to two or more experienced raters. This would allow for a more thorough statistical analysis of differences in scores awarded on the two different modes of oral tests.

9 Mean scores of the four groups were as follows: Group 1 (n=8): 6.42; Group 2 (n=9): 7.28; Group 3 (n=7): 6.52; Group 4 (n=8): 6.33.

Examiner and test-taker training

- All comments from both examiners and test-takers pointed to the need for explicit examiner and test-taker training if the introduction of computer-based oral testing is to be considered in the future. The possibility that the interaction between the test mode and discourse features might have resulted in slightly lower Fluency scores highlights the importance of counteracting the possible disadvantages under the video-conferencing mode through examiner training and awareness raising. It would be extremely useful to design a further study to train examiners in the use of the technology and also develop materials for test-takers to prepare themselves for video-conferencing delivery. This study could then be replicated and similar analyses performed without the confounding variable of sound quality and computer familiarity.

Analysis of interactional features

- Some examiners felt that they might not have been able to ask as many questions with video-conferencing in Part 3. Although the number of questions did not show any statistically significant difference (see Table 8 under RQ1b), it would be useful to examine discourse features such as examiners' and test-takers' rate of speech and test-takers' length of responses. Any significant difference in such features may impact on the amount of speech which the test allows test-takers to produce, and may have implications for the comparability between the two test modes and the validity of the online mode. Transcribing the recorded interviews will help answer these questions.
- With transcripts and video recordings, it is possible to conduct a more objective, detailed, integrated analysis of how and where examiners' rating and/or examining behaviours are different under the two modes. It would also give more information about how some test-takers can take more initiative under the video-conferencing condition, i.e. examiners reported having less control over the direction of the speaking tests in the video-conferencing mode. This may be useful for future examiner training should the Zoom or another computer-mediated video-conferencing mode be introduced at some future time.
- The availability of audio/video data would also allow a focus on a range of conversational features, e.g. length of turn, turn interruptions/overlaps, gaps between turns. These features have been shown to play a role in interaction, both in the Pragmatics and Conversation Analysis literature (e.g. Itakura 2001; Sacks, Schegloff and Jefferson 1974) and in the Language Testing literature (e.g. Berry 2007; Brown 2003; Galaczi 2008, 2014; Gan 2010; Nakatsuhara 2013). At the same time, research in language testing has also indicated that the effect of these features on discourse and on the scores awarded is not necessarily linear and clear-cut (Fulcher 1996). It seems important, therefore, to focus further investigations on interactional features in the two modes to ascertain whether their use – and therefore their impact on scores – differs across the two modes.

Sound quality and test-taker perceptions of performance

- An important issue mentioned in the technical report (Appendix 8) is that some of the test-takers blamed the sound quality for their (poor) performance when the sound and transmission were both fine (as the technical expert recorded and monitored all the sessions in real time, he was able to identify such cases). Using the researchers' field notes and re-visiting video recordings, it would be possible to explore when this is more likely to happen. This may be useful for future examiner training and quality control procedures.
- The use of headphones could also be investigated in the future to ascertain whether they play a significant role in improving sound quality. During the planning stage of this project, the use of headphones for both examiners and test-takers

was considered. Unfortunately, in this instance, it was not feasible as this would have interfered with the video recordings which the research team considered more important for the analysis. However, the use of headphones is something that should definitely be investigated in future studies.

Final comments

It is important to emphasise once again the necessity of building on this small-scale investigation by designing and implementing a further study to train raters in the use of video-conferencing techniques including, for example, looking directly at the camera to simulate eye-contact, while at the same time developing appropriate materials to prepare test-takers for oral interaction through video-conferencing. Carrying out a similar study after mitigating the effects of video-conferencing familiarity on examiners and test-takers would then allow for replication on a larger-scale to confirm and add generalisability to the initial findings.

It is also absolutely essential to address the practical issue of technical problems. Throughout the report, the technical issues with the video-conferencing mode were repeatedly reported regarding the sound quality and speed of video transmission. Should the video-conferencing delivery of the IELTS speaking test be administered in the future, it is critically important that the test venues have stable internet connections, sufficient bandwidth and local technical support, as noted in the technical report (Appendix 8).

Two early indications which emerged from this study, despite its limitations (mostly due to size and technology), are:

- the two modes of delivery, face-to-face and video-conferencing, are likely to result in test-takers achieving essentially the same score, irrespective of familiarity or comfort with the particular delivery mode
- the two modes of delivery seem to be comparable in terms of the underlying construct, despite some differences in language functions elicited from test-takers which may have been caused by the sound quality and delayed video transmission under video-conferencing mode.

These findings therefore support the continued investigation of the tablet/computer-delivery video-conferencing mode as a potentially viable platform for the delivery of high-stakes speaking tests.

References

ALC Press Inc. (2015). The Standard Speaking Test. http://tsst.alc.co.jp/sst/e/index.html (accessed on 21 July 2015)

Atkinson, J. M. and Heritage, J. (eds.) (1984). *Structures of social action: Studies in Conversation Analysis*. Cambridge, New York: Cambridge University Press.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. and Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

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

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt am Main: Peter Lang.

Bond, T. G. and Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences (2nd edition)*. University of Toledo.

Bonk, W. J., and Ockey, G. J. (2003). *A many-facet Rasch analysis of the second language group oral discussion task*. Language Testing, 20(1), pp. 89–110.

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

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), pp. 1–25.

Brown, A. and Hill, K. (1998). Interviewer style and candidate performance in the IELTS Oral Interview, in S. Wood (Ed.) *IELTS Research Reports, Volume 1*. IELTS Australia. Available from http://www.ielts.org/pdf/Vol1Report1.pdf

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, pp. 1–47.

Chapelle, C. and Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge University Press.

Chun, C. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly*, 3(3), pp. 295–306.

Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing*, 5(2), pp. 197–205.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences (2nd edition)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Creswell, J. W. and Plano Clark, V. L. (2011). *Designing and conducting mixed methods research (2nd edition)*. Thousand, Oaks, CA: Sage Publications.

Douglas, D. and Hegelheimer, V. (2007). Assessing language using computer technology. In M. McGroarty (Ed.), *Annual Review of Applied Linguistics* (Vol. 27, pp. 115–132). Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking. Studies in Language Testing*, Volume 30. Cambridge: Cambridge University Press.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13, pp. 208–238.

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

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Computer-based Assessment of Foreign Language Speaking Skills* (pp. 29–51). Luxembourg: European Union.

Galaczi, E. D. (2014). Interactional Competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, 35(5), pp. 553–574.

Gan, Z. (2010). Interaction in group oral assessment: a case study of higher- and lower-scoring students. *Language Testing*, 27(4), pp. 585–602.

Gass, S. M. and Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Hoejke, B. and Linnell, K. (1994). Authenticity in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, 28(1), pp. 103–126.

Inoue, C. (2013). *Investigating the use of language functions for validating speaking test specifications*. Paper presented at Language Testing Forum 2013, Nottingham Trent University, UK (15–17 November 2013).

Isaacs, T. (2010). *Issues and arguments in the measurement of second language pronunciation*. Unpublished PhD thesis, McGill University, Montreal.

Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics*, 33, pp. 1859–1880.

Jamieson, J. (2005). Trends in computer-based second language assessment. In M. McGroarty (Ed.), *Annual review of Applied Linguistics*. (vol. 25, pp. 228–242). Cambridge: Cambridge University Press.

Kenyon, D. and Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other proficiency assessments. *Language Learning and Technology* 5(2) pp. 60–83.

Kramsch, C. (1986). From language proficiency to interactional competence. The *Modern Language Journal*, 70(4), pp. 366–372.

Linacre, M. (2013a). *Facets computer program for many-facet Rasch measurement, version 3.71.2*. Beaverton, Oregon: Winsteps.com.

Linacre, M. (2013b). *A user's guide to FACETS: Rasch-model computer programs*, available online at http://www.winsteps.com/a/facets-manual.pdf

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study*. Jyvaskyla: University of Jyvaskyla.

Luoma, S. (2004). *Assessing speaking* Cambridge: Cambridge University Press.

May, L. (2011). *Interaction in a paired speaking test: The rater's perspective*. Frankfurt: Peter Lang.

McCarthy. M. (2006). *Explorations in corpus linguistics*. Cambridge: Cambridge University Press.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA & Oxford: Blackwell.

Nakatsuhara, F. (2013). T*he co-construction of conversation in group oral tests*. Frankfurt am Main: Peter Lang.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests. Studies in Language Testing,* Volume 13. Cambridge: 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), pp. 33–56.

O'Sullivan, B. and Yang, L. (2006). An empirical study on examiner deviation from the set interlocutor frame in the IELTS speaking paper. *IELTS Research Reports*, Volume 6, pp. 91–118. IELTS Australia and British Council.

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers *Language Assessment Quarterly 6*(2), pp. 113–125.

Sacks, H., Schegloff, E. A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, pp. 696–735.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11, pp. 99–123.

Stansfield, C. (1990). An evaluation of simulated oral proficiency interviews as measures of oral proficiency. In J. E. Alatis (Ed.), *Georgetown University Roundtable of Languages and Linguistics 1990* (pp. 228–234). Washington, D.C.: Georgetown University Press.

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

Taylor, L. (2011). Introduction. In L. Taylor (ed.), Examining Speaking. Studies in Language Testing, Volume 30 (pp.1–35). Cambridge: Cambridge University Press.

Taylor, L. and Galaczi, E. D. (2011). The scoring validity of Cambridge ESOL speaking tests. In L. Taylor (Ed.), *Examining Speaking. Studies in Language Testing, Volume.30*. (pp. 171–233). Cambridge: Cambridge University Press.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3), pp. 325–344.

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wolf, J. P. (2008). The effects of backchannels on fluency in L2 oral task production. *System*, 36, pp. 279–294.

Wright, B. and Linacre, M. (1994). *Reasonable mean-square fit values*. Retrieved 27 March 2012 from http://www.rasch.org

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), pp. 291–300.

Appendices

Appendix 1: Exam rooms

Exam room for face-to-face tests	Exam room for video-conferencing tests (for examiners)
• Examiner • Test-taker • Researcher (FN/CI) • Video recorder + tripod • Audio recorder	• Examiner • Researcher (FN/CI) • IT advisor (JP) • iPad +Bluetooth speaker/microphone • Video recorder + tripod • Audio recorder
	Exam room for video-conferencing tests (for test-takers)
	• Test-taker • Researcher (VB) • iPad +Bluetooth speaker/microphone • Video recorder + tripod • Audio recorder

Appendix 2: Test-taker questionnaire

Test-taker questionnaire

You did 2 speaking tests today. One test was with an interviewer **face-to-face** (f2f) and the other was with an interviewer **via a computer** (COMPUTER). To help us understand the differences between these 2 test formats, we'd like to ask you some questions about your experience of them.

Name: **ID No.:**

For all sections below, tick the relevant boxes below according to the test-taker's responses.

The face-to-face (f2f) test

		1 Never	2	3 Sometimes	4	5 Always
Q1	Did you understand the examiner?					

Additional comments (as appropriate):

		1 V difficult	2	3 OK	4	5 Very easy
Q2	Did you feel taking the test face to face was…					

Additional comments (as appropriate):

The computer test

		1 Never	2	3 Sometimes	4	5 Always
Q3	Did you understand the examiner?					

Additional comments (as appropriate):

		1 V difficult	2	3 OK	4	5 Very easy
Q4	Did you feel taking the test using a computer was…					

Additional comments (as appropriate):

Both tests				
		f2f	Computer	No difference
Q5	Which speaking test made you more nervous – the face-to-face one, or the one using the computer?			
Q6	Which speaking test was more difficult for you – the face-to-face one, or the one using the computer?			
Q7	Which speaking test gave you more opportunity to speak English – the face-to-face one, or the one using the computer?			
Q8	Which speaking test did you prefer – the face-to-face one, or the one using the computer?			

Why?

Any other comments?

Thank you for answering these questions.

Appendix 3: Examiner questionnaire

Examiner questionnaire

Today you administered and rated a number of IELTS Speaking Tests according to two different delivery modes: one mode involved delivering the standard **Face-to-Face** (f2f) approach for the IELTS Speaking Test; an alternative mode involved administering and rating the IELTS Speaking Test **via a computer** (COMPUTER).

To help inform an evaluation of the alternative (COMPUTER) mode of test delivery and rating, and to compare this approach with the standard mode, we'd welcome comments on your experience of administering and rating the IELTS Speaking Test across the two modes.

Background data

Name:

Current examiner role? *(delete as appropriate)*

Examiner Support Coordinator

Examiner Trainer

Examiner

Principal Examiner

Assistant Principal Examiner

Years of experience as an EFL/ESL teacher? ..... years .....months

Years of experience as an IELTS examiner? ..... years .....months

Typical proficiency range of IELTS candidates you examine (e.g. band 5.5–7.0)?

Tick the relevant boxes according to how far you agree or disagree with the statements below.

1a. Administering the face-to-face test

		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Q1	Overall I felt comfortable in administering the IELTS Speaking Test in the standard format					
Q2	I found it straightforward to administer **Part 1** (frames) of the IELTS Speaking Test in the standard format					
Q3	I found it straightforward to administer **Part 2** (long turn) of the IELTS Speaking Test in the standard format					
Q4	I found it straightforward to administer **Part 3** (2-way discussion) of the IELTS Speaking Test in the standard format					
Q5	The examiner's interlocutor frame was straightforward to handle and use in the standard format					

Additional comments?

1b. Rating the face-to-face test						
		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Q6	Overall I felt comfortable rating candidate performance in the standard IELTS Speaking Test					
Q7	I found it straightforward to apply the **Fluency and Coherence scale** in the standard format					
Q8	I found it straightforward to apply the **Lexical Resource scale** in the standard format					
Q9	I found it straightforward to apply the **Grammatical Range and Accuracy scale** in the standard format					
Q10	I found it straightforward to apply the **Pronunciation scale** in the standard format					
Q11	I feel confident about the accuracy of my ratings on the standard format					

Additional comments?

2a. Administering the computer-delivered test						
		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Q12	Overall I felt comfortable in administering the IELTS Speaking Test in the computer format					
Q13	I found it straightforward to administer **Part 1** (frames) of the IELTS Speaking Test in the computer format					
Q14	I found it straightforward to administer **Part 2** (long turn) of the IELTS Speaking Test in the computer format					
Q15	I found it straightforward to administer **Part 3** (2-way discussion) of the IELTS Speaking Test in the computer format					
Q16	The examiner's interlocutor frame was straightforward to handle and use in the computer format					

Additional comments?

2b. Rating the computer-delivered test						
		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Q17	Overall I felt comfortable rating candidate performance in the computer-delivered IELTS Speaking Test					
Q18	I found it straightforward to apply the **Fluency and Coherence scale** in the computer-delivered format					
Q19	I found it straightforward to apply the **Lexical Resource scale** in the computer-delivered format					
Q20	I found it straightforward to apply the **Grammatical Range and Accuracy scale** in the computer-delivered format					
Q21	I found it straightforward to apply the **Pronunciation scale** in the computer-delivered format					
Q22	I feel confident about the accuracy of my ratings on the computer-delivered format					

Additional comments?

3. Comparing the experience of using the standard (f2f) and the alternative (computer) modes for the IELTS Speaking Test				
		f2f	Computer	No difference
Q23	Which mode of speaking test did you feel more comfortable with?			
Q24	Which mode of speaking test did you feel was easier for you to **administer**?			
Q25	Which mode of speaking test did you feel was easier for you to **rate**?			
Q26	Which mode of speaking test do you think gave a better chance for the test-taker to demonstrate their level of English proficiency?			
Q27	Which speaking test did you prefer?			
Q28	Are you aware of doing anything differently in your examiner role across the 2 speaking test modes – f2F and COMPUTER? If yes, please give details below:			

Thank you for answering these questions.

Appendix 4: Observation checklist

(Modified from O'Sullivan et al, 2002 – all modifications are highlighted in red)

1. Informational functions		
Operation	**Gloss: Does a Test taker…**	**For example**
Providing personal information	Give information on present circumstances?	"I'm studying English here in London." "I live…" "I work…"
	Give information on past experiences?	"I studied economics at university" "I've been/ I went to… before/last week"
	Give information on future plans?	"After I go home, …" "I hope to qualify in June." "I'm going/ going to go/ I'll go home next week."
Expressing opinions/ preferences	Express opinions? Expressing preference?	Can be signalled: "I don't like English food." Can be unsignalled: "It would be better if schools were given more funding." Also can be Positive or Negative. "I think this one would be best." "I'd rather have a small one." "I prefer/like this one better."
Elaborating	Elaborate on, or modify an opinion?	Can be signalled: "I mean…" Or "Maybe not that good, but…" Can be unsignalled: "They could reduce class size, or…"
Justifying opinions	Express reasons for assertion s/he has made?	Can be signalled by the test taker: "It's because…" Can be signalled by the other test taker: "Why…" Can be signalled by the examiner: "Well, if they are really interested in the work, that in itself will motivate them and they won't mind how much they are paid." Can be unsignalled: "It's prettier, and cheaper…"
Comparing	Compare things/people/events?	"I think X is more useful" "Both are interesting, but I prefer the style and colours in he smaller one" "This picture shows.. whereas/ while/but this one is busier/more crowded/more interesting"
Speculating	Speculate?	"She must have paid a fortune for that." "I can imagine him spending hours on preparing that." "This might/could/should/would/ can't be must be…"
Staging	Separate out or interpret the parts of an issue?	"So, first I'll talk about…" "So, you think he did it, but it wasn't deliberate, or do you think he was provoked and it was an instinctive reaction?" "But first, we have to… and now. We must choose…"

Describing	Describe ~~a sequence of~~ events / things / people?	Can be marked: "When she first goes to Italy, she is very innocent. Then…" Can be unmarked: "I went to buy a ticket and found that the ticket office had already closed."
Summarising	Summarise what s/he has said?	"So, I think we would choose,…" "So you think…" "So we have decided/chosen…"
Suggesting	Suggest a particular idea?	"We could choose this one." "What about…" "We could (do)…" "Why don't we (do)…" "How about (doing)?"
~~**Expressing preferences**~~[10]		

10 To be combined with 'Expressing opinions'

2. Interactional functions		
Operation	**Gloss: Does a Test taker…**	**For example**
Agreeing	Agree with an assertion made by another speaker? (apart from "yeah" or non-verbal)	Can be marked: "Yes, I agree." "I think you're right." Can be unmarked: "But you can't/ don't mean… do you?"
Disagreeing	Disagree with what another speaker says? (apart from "no" or non-verbal)	Can be marked: "I don't think that's right." "I (don't) agree with you" Can be unmarked: "But you can't/don't mean…, do you?" "Well, that depends on your point of view, but I rather think…"
Modifying/commenting/ adding	Modify/comments on arguments or comments made by other speaker? Or by the test taker in response to another speaker?	"Of course, only is he was forced to go, otherwise…" "Well, (perhaps) not for this but for that…" ~~Other speaker's input may be verbal (Why?), nonverbal (raised eyebrow) or even paraverbal (mmm? –raising intonation)~~
Asking for opinions	Ask for opinions?	"What do you think?" "And you?" "Well?"
Persuading	Attempt to persuade another person?	Can be cued: "Don't you think?" "But don't you think that…?" Can be uncued: "Yes, but he can't spend it all, or he won't have enough left to eat!"
Asking for information	Ask for information?	"What about you? What are your favourite films?" "What are your hobbies/ leisure activities?" "Do you know…"
Conversational repair (only self-repair)	Repair breakdowns in interaction?	~~Can be "other repair" – breakdown during other speaker's turn: "I'm sorry I thought you meant…"~~ → clarification request & responding to requests (negotiating meaning) Can be "self repair" – breakdown during own turn: "What I wanted to say was…" These repairs may be initiated by the person who is speaking (self-initiated) or by the other person (other initiated) and can be verbal ("Pardon.") or non-verbal (quizzical look).

Negotiating meaning	Check OWN understanding?	"So, do I have to (describe all the photographs)?"
	Check OTHER'S understanding?	"OK?" "Is that clear?" ~~"So, do I have to (describe all the photographs)?"~~
	Indicate understanding of point made by partner?	Can be verbal: "Yes, I know what you mean." "OK, yes." Can be non-verbal: head nod Can be paraverbal: mmm (with or without intonational changes)
	Establish common ground/ purpose or strategy?	"Shall we talk about all of them first before deciding?" "But we have to choose three pictures." "So, we both like this one…"
	Ask for clarification when an utterance is misheard or misinterpreted?	"Can you repeat that please?" "What exactly do you mean by wealthy?"
	Correct an utterance made by other speaker which is perceived to be incorrect or inaccurate.	"No, we're already decided not to take that one." "You mean…" (usually a lexical or grammatical correction)
	Respond to requests for clarification?	Can be cued: "What I mean is…." Can be non-cued: "The blue one." The request itself may be verbal ("Which…") or nonverbal (quizzical look)

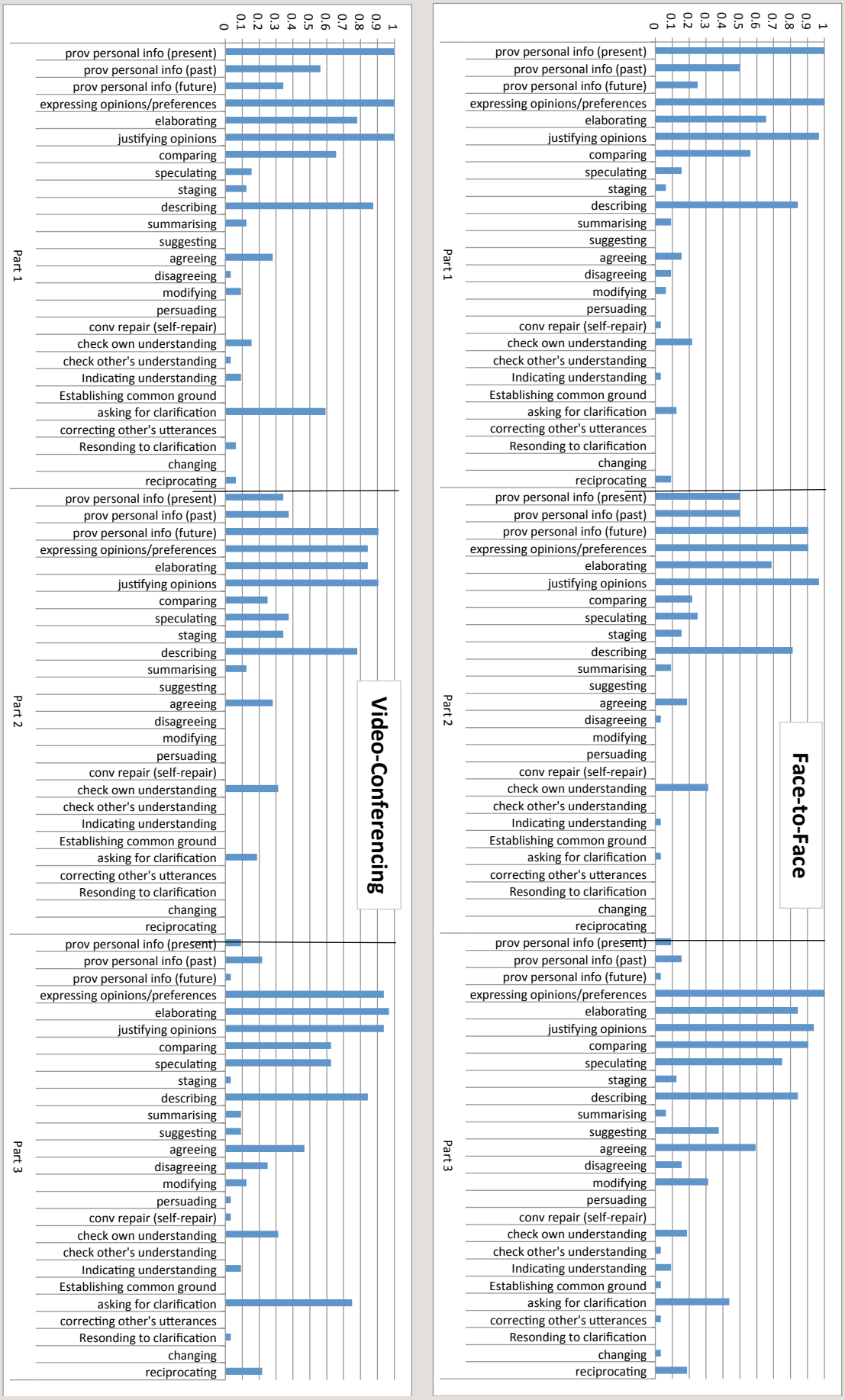
3. Managing interaction		
Operation	**Gloss: Does a Test taker…**	**For example**
Initiating	Start any interactions?	"What do you think?" "Right, so we have to choose the best, what do you think of the blue one?" "But what about …?" "But this one is (much) more, don't you think?"
Changing	Take the opportunity to change the topic?	"Yes, that would be the best, So what about the worst?" "Talking of sizes, did I tell you about those shoes I saw?" "I don't like going to a gym, but I like to go for a walk. Last weekend…"
Reciprocating	Share the responsibility for developing the interaction?	"What do you think we should do?" "Have you ever tried to do it?" May simply consist of verbal ("Yes"), non-verbal (head-nod) or paraverbal (uh huh, mm hmm) support – used to encourage other speaker to continue.
Deciding	Come to a decision?	"So, we have decided…" "You're right, it's easier that way. That will work." "So, let's choose/we've chosen…" "I would choose…" "I think we should choose"

Appendix 5: Transcription notation

(Modified from Atkinson and Heritage, 1984)

Unfilled pauses or gaps	Periods of silence. Micro-pauses (less than .2 second) are shown as (.); longer pauses appear as a time within parentheses. E.g. (.5) represents five tenths of a second
Colon (:)	A lengthened sound or syllable; more colons prolong the stretch
Dash (-)	A cut off, usually a glottal stop
.hhh	Inhalation
Hhh	Exhalation
hah, huh, heh	Laughter
(h)	Breathiness within a word
Punctuation	Intonation rather than clausal structure; a full stop (.) is falling intonation, a question mark (?) is rising intonation, a comma (,) is continuing intonation
Equal sign (=)	A latched utterance, no interval between utterances
Open bracket ([ )	Beginning of overlapping utterances
Percent signs (% %)	Quiet talk
Asterisks (* *)	Creaky voice
Empty parentheses ( )	Words within parentheses are doubtful or uncertain
Double parentheses (( ))	Non-vocal action, details of scene
Arrows (><)	The talk speeds up
Arrows (<>)	The talk slows down
Underlining	A word or sound is emphasised
Psk	A lip smack
Tch	A tongue click
Arrow (→)	A feature of interest to the analyst

Appendix 6: Shifts in use of language functions from Parts 1 to 3 under face-to-face/video-conferencing conditions



Appendix 7: Comparisons of use of language functions between face-to-face (f2f)/video-conferencing (VC) conditions

		Part 1					Part 2					Part 3				
Functions	**Test mode**	**Median**	**Mean**	**SD**	**Z**	**Sig.**	**Median**	**Mean**	**SD**	**Z**	**Sig.**	**Median**	**Mean**	**SD**	**Z**	**Sig.**
Providing personal info (present)	F2F	1.00	1.00	0.00	.000	1.000	0.50	0.50	0.51	-1.387	.166	0.00	0.09	0.30	.000	1.000
	VC	1.00	1.00	0.00			0.00	0.34	0.48			0.00	0.09	0.30		
Providing personal info (past)	F2F	0.50	0.50	0.51	-.535	.593	0.50	0.50	0.51	-.894	.371	0.00	0.16	0.37	-.707	.480
	VC	1.00	0.56	0.50			0.00	0.38	0.49			0.00	0.22	0.42		
Providing personal info (future)	F2F	0.00	0.25	0.44	-.775	.439	1.00	0.91	0.30	.000	1.000	0.00	0.03	0.18	.000	1.000
	VC	0.00	0.34	0.48			1.00	0.91	0.30			0.00	0.03	0.18		
Expressing opinions/ preferences	F2F	1.00	1.00	0.00	.000	1.000	1.00	0.91	0.30	-.707	.480	1.00	1.00	0.00	-1.414	.157
	VC	1.00	1.00	0.00			1.00	0.84	0.37			1.00	0.94	0.25		
Elaborating	F2F	1.00	0.66	0.48	-1.414	.157	1.00	0.69	0.47	-1.508	.132	1.00	0.84	0.37	-2.000	.046
	VC	1.00	0.78	0.42			1.00	0.84	0.37			1.00	0.97	0.18		
Justifying opinions	F2F	1.00	0.97	0.18	-1.000	.317	1.00	0.97	0.18	-1.000	.317	1.00	0.94	0.25	.000	1.000
	VC	1.00	1.00	0.00			1.00	0.91	0.30			1.00	0.94	0.25		
Comparing	F2F	1.00	0.56	0.50	-.775	.439	0.00	0.22	0.42	-.333	.739	1.00	0.91	0.30	-2.714	.007
	VC	1.00	0.66	0.48			0.00	0.25	0.44			1.00	0.63	0.49		
Speculating	F2F	0.00	0.16	0.37	.000	1.000	0.00	0.25	0.44	-1.265	.206	1.00	0.75	0.44	-1.069	.285
	VC	0.00	0.16	0.37			0.00	0.38	0.49			1.00	0.63	0.49		
Staging	F2F	0.00	0.06	0.25	-.816	.414	0.00	0.16	0.37	-1.897	.058	0.00	0.13	0.34	-1.342	.180
	VC	0.00	0.13	0.34			0.00	0.34	0.48			0.00	0.03	0.18		
Describing	F2F	1.00	0.84	0.37	-.333	.739	1.00	0.81	0.40	-.277	.782	1.00	0.84	0.37	.000	1.000
	VC	1.00	0.88	0.34			1.00	0.78	0.42			1.00	0.84	0.37		
Summarising	F2F	0.00	0.09	0.30	-.378	.705	0.00	0.09	0.30	-.378	.705	0.00	0.06	0.25	-1.000	.317
	VC	0.00	0.13	0.34			0.00	0.13	0.34			0.00	0.09	0.30		

		Part 1					Part 2					Part 3				
Functions	**Test mode**	**Median**	**Mean**	**SD**	**Z**	**Sig.**	**Median**	**Mean**	**SD**	**Z**	**Sig.**	**Median**	**Mean**	**SD**	**Z**	**Sig.**
Suggesting	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.38	0.49	-2.714	.007
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.09	0.30		
Agreeing	F2F	0.00	0.16	0.37	-1.633	.102	0.00	0.19	0.40	-1.134	.257	1.00	0.59	0.50	-1.000	.317
	VC	0.00	0.28	0.46			0.00	0.28	0.46			0.00	0.47	0.51		
Disagreeing	F2F	0.00	0.09	0.30	-1.414	.157	0.00	0.03	0.18	-1.000	.317	0.00	0.16	0.37	-.905	.366
	VC	0.00	0.03	0.18			0.00	0.00	0.00			0.00	0.25	0.44		
Modifying/ commenting/ adding	F2F	0.00	0.06	0.25	-.447	.655	0.00	0.00	0.00	.000	1.000	0.00	0.31	0.47	-2.121	.034
	VC	0.00	0.09	0.30			0.00	0.00	0.00			0.00	0.13	0.34		
Asking for opinions	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Persuading	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	-1.000	.317
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Asking for information	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Conversational repair	F2F	0.00	0.03	0.18	-1.000	.317	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	-1.000	.317
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.03	0.18		
Checking own understanding	F2F	0.00	0.22	0.42	-.816	.414	0.00	0.031	0.18	.000	1.000	0.00	0.19	0.40	-1.069	.285
	VC	0.00	0.16	0.37			0.00	0.031	0.18			0.00	0.31	0.47		
Checking other's understanding	F2F	0.00	0.00	0.00	-1.000	.317	0.00	0.00	0.00	.000	1.000	0.00	0.03	0.18	-1.000	.317
	VC	0.00	0.03	0.18			0.00	0.00	0.00			0.00	0.00	0.00		
Indicating understanding	F2F	0.00	0.03	0.18	-1.000	.317	0.00	0.03	0.18	-1.000	.317	0.00	0.09	0.30	.000	1.000
	VC	0.00	0.09	0.30			0.00	0.00	0.00			0.00	0.09	0.30		

		Part 1					Part 2					Part 3				
Functions	**Test mode**	**Median**	**Mean**	**SD**	**Z**	**Sig.**	**Median**	**Mean**	**SD**	**Z**	**Sig.**	**Median**	**Mean**	**SD**	**Z**	**Sig.**
Establishing common ground	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.03	0.18	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Asking for clarification	F2F	0.00	0.13	0.34	-3.638	.000	0.00	0.03	0.18	-1.890	.059	0.00	0.44	0.50	-2.496	.013
	VC	1.00	0.59	0.50			0.00	0.19	0.40			1.00	0.75	0.44		
Correcting other's utterances	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.03	0.18	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Responding to clarification request	F2F	0.00	0.00	0.00	-1.414	.157	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000
	VC	0.00	0.06	0.25			0.00	0.00	0.00			0.00	0.03	0.18		
Initiating	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Changing	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.03	0.18	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		
Reciprocating	F2F	0.00	0.09	0.30	-.577	.564	0.00	0.06	0.25	-1.414	.157	0.00	0.19	0.40	.000	1.000
	VC	0.00	0.06	0.25			0.00	0.00	0.00			0.00	0.22	0.42		
Deciding	F2F	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000	0.00	0.00	0.00	.000	1.000
	VC	0.00	0.00	0.00			0.00	0.00	0.00			0.00	0.00	0.00		

Appendix 8: A brief report on technical issues encountered during data collection (20–23 January 2014) by Jermaine Prince

Black text: statement of problems
Red text: recommendations

Ensuring good internet connection at the venue

- Although we trialled using Zoom and recording with it at the venue on two separate dates a week before the data collection, the internet wireless connection on the actual dates of data collection was not as stable.
- Various factors were at play: room location (i.e. examiner room and test-taker room far from each other; room facing a busy road and fire station which were noisy), usage of internet by the staff and students at the venue on the day (the campus venue was busier and, therefore, the transmission was more delayed than other days).
- Unstable internet connection affected: (1) the quality of video and sound; (2) the recording (i.e. there was a Zoom session which showed severely delayed and out-of-sync transmission of video and sound, but it wasn't observed in real-time); and (3) examiner behaviour (speaking closer to the speaker, holding the speaker, louder etc.).
- Need to ensure that the site location is viewed and tested at least a week in advance.
- The instability of wireless strength on one floor was manageable but even harder to maintain a consistent level of recording over three floors and with rooms situated in different parts of the corridors. Room locations need to be carefully considered and trialled.
- The room needs to be well lit and, if possible, test-taker should be in front of a white background (for good quality video images).

Recommendations on administering tests under Zoom condition

- Have a technical person on-site to handle technical problems and solutions that may arise. Have preparation field work in advance of the examination so the centre can facilitate the room requirements AND system requirements.
- Using Zoom and recording sessions require storage devices with a massive capacity and a laptop to process such large amount of data (which is free from restrictions NOT like some work laptops that may have various restrictions in transmitting/uploading data for security reasons).
- Ensure that the computer being used to host the recordings does not have too much security on it, as it may be difficult to obtain the files afterwards. Alternatively the files can be uploaded to Google drive and then downloaded as a way around this.
- The time it takes to transfer memory (this is crucial because all the sessions may need to be recorded if/when Zoom version is launched for quality assurance purposes (also some test-takers may enquire about their ratings, which results in other rater(s) may need to find the recording and use it to re-rate)).
- In order to transfer all data, an average time of one hour will be required – please note all recordings are done in high definition. Have at least 1tb external hard drive to store all data.

Recommendations (if another research like this is administered with external cameras)

- Ensure that batteries are fully charged, memory cards are empty for a day's worth of recording and that the camera is set to video mode (so that the researchers do not record in a wrong mode).
- To help identify test-takers, it is important that the details are written on a card and placed in front of the camera before each recording.
- As an alternative to using a camera or in emergency of battery failure, an iPad can be used to record the session as it also records in HD, the data is easily transferrable.

Other notes

Equipment

- Apple iPad Air
- Sony SRS BTS50 Wireless Portable Speaker Bluetooth with NFC
- Laptop (host for monitoring and recording Zoom sessions)

Software

- Zoom Cloud meeting – offers the best video, audio and screen-sharing experience across Windows PC, Mac, iOS, Android and H.323/SIP room systems

Moving forward

- There has to be a set specification for location of exams
- Strong and stable internet connection
- On-site rooms have to be situated near one another
- Have personal internet access point (if possible)
- There needs to be a buffer-period for charging devices
- Have a card reader to extract data from memory cards if used for an external camera
- If researchers need to record on an external camera, it is highly recommended that it is recorded upon a camcorder rather than digital SLRs as the battery life/space is better to manage
- Have water bottles ready for examiners (although the venue had a water server, the cups were not self-standing, so examiners could not take water into the room with them. One of them started coughing hard during one of the sessions, and it would have been nice to have water at hand.)

Observation

- A great initiative that will expand and make further examinations available to test-takers
- An examiner's voice levels fluctuated under Zoom
- Examiners tend to look at their papers rather than at the screen with Zoom
- Test-takers can use technology as an excuse for a poor performance