Video-conferencing speaking tests: Do they measure the same construct as face-to-face tests?

Fumiyo Nakatsuhara[a], Chihiro Inoue[a], Vivien Berry[b] and Evelina Galaczi[c]

[a]*University of Bedfordshire, Luton, UK;* [b]*British Council, London, UK;* [c]*Cambridge Assessment English, Cambridge, UK*

Corresponding author:
Fumiyo Nakatsuhara
Fumiyo.Nakatsuhara@beds.ac.uk
Centre for Research in English Language Learning and Assessment (CRELLA), Room 118
University of Bedfordshire, Putteridge Bury Campus, Hitchin Road, Luton LU2 8LE UK.

ORCiD
Fumiyo Nakatsuhara (orcid.org/0000-0002-0429-7417)
Chihiro Inoue (orcid.org/0000-0003-1927-6923)
Vivien Berry (orcid.org/0000-0002-5610-4787)
Evelina Galaczi (orcid.org/0000-0001-8269-8462)

Funding
This project was funded by the IELTS Partners (British Council, Cambridge Assessment English, and IDP: IELTS Australia).

Acknowledgements
This project was funded and supported by the IELTS Partners. We are also grateful to British Council, China, for their contributions to data collection.

Keywords

speaking test construct, video-conferencing, language functions, test scores, online interaction

Ethics

Ethical approval was granted for this research by the CRELLA Research Institute Ethics Committees overseen by the University Research Ethics Committee (UREC) at the University of Bedfordshire, with the Approval Number, *RES2015-05*.

Abstract

This paper investigates the comparability between the video-conferencing and face-to-face modes of the IELTS Speaking Test in terms of scores and language functions generated by test-takers. Data were collected from 10 trained IELTS examiners and 99 test-takers who took two speaking tests under face-to-face and video-conferencing conditions. Many-facet Rasch Model (MFRM) analysis of test scores indicated that the delivery mode did not make any meaningful difference to test-takers' scores. An examination of language functions revealed that both modes equally elicited the same language functions except *asking for clarification*. More test-takers made clarification requests in the video-conferencing mode (63.3%) than in the face-to-face mode (26.7%). Drawing on the findings, as well as practical implications, we extend emerging thinking about video-conferencing speaking assessment and the associated features of this modality in its own right.

Research Background

Online Spoken Communication

Technology developments in internet protocol-based video-conferencing (VC) and the widespread availability of video-enabled mobile phones and tablets in the last decade have made VC interaction more widespread than ever before. This trend has been boosted dramatically in response to the COVID-19 global health pandemic. The physical proximity restrictions, social distancing advisories and remote working/studying triggered by this global phenomenon have brought about a shift in the way we communicate, with VC technology quickly becoming one of the main channels of remote communication in social, professional and educational settings.

Discussions on the nature of online spoken interaction can be traced back to Sellen (1995, p. 407), one of the pioneer researchers in VC technology, who noted twenty-five years ago that "simply giving conversants visual access via video technology does not render the situation similar to being physically co-present". VC technology has advanced significantly in the years since Sellen made that statement, but her observation still holds. More recently, the Council of Europe (2018, p. 96) also acknowledged that online interaction is unlikely to be the

same as face-to-face interaction; "… [in online communication] there may be misunderstandings which are not spotted (and corrected) immediately, as is often easier with face-to-face communication".

Comparability of face-to-face and video-conferencing modes of a speaking test

When test providers consider introducing a video-conferencing (VC) mode of a speaking test as an alternative to the face-to-face test, the comparability of the two test modes needs to be empirically established. The adoption of an alternative test mode, especially one which is novel in a high-stakes test environment, necessitates the gathering of sound validity evidence about its use and associated score interpretations. It is now widely accepted in the L2 assessment field that validity arguments are an ongoing process of forming an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores (Messick, 1989). It is recommended that such validity evidence comes from varied sources and includes quantitative as well as qualitative data (Jang et al., 2014; Moeller et al., 2016; Riazi, 2017), since it is through such mixed-methods designs that the most comprehensive validity evidence can be drawn.

Score comparability of VC and face-to-face tests, which is an essential aspect of comparability between the two modes, has been examined using paired t-tests (e.g. Clark & Hooshmand, 1992; Kim & Craig, 2012) and, more recently, Many-facet Rasch Model (MFRM; Nakatsuhara et al., 2017a). All of these studies concluded that the scores of the face-to-face and VC modes can be considered equivalent. However, Nakatsuhara et al. (2017a) called for further score comparability research with a more robust MFRM research design, since the analysis of their small-scale study could not fully benefit from the potential of MFRM (e.g.

Bond & Fox, 2015; Eckes, 2015). More specifically, in order to confirm the score comparability of the two delivery modes, larger-scale research is needed, exploring potential sources of score variance other than delivery mode (e.g. examiner severity level, multiple test versions).

Other aspects of the comparability between the two modes explored thus far include test-taker perceptions (Nakastsuhara et al., 2016; Clark & Hooshmand, 1992; Craig & Kim 2010), examiner perceptions (Clark & Hooshmand, 1992; Nakatsuhara et al., 2016, Inoue et al., forthcoming) and examiner rating behaviour (Nakatsuhara et al., 2016, Inoue et al., forthcoming). These studies provided evidence supporting test comparability claims, despite some minor differences, which can be addressed through enhancement of examiner and test-taker training on the VC tests. Furthermore, while a direct comparison with face-to-face tests was not made, important insights into test-taker perceptions and the usability and stability of the VC technology have been offered by two recent studies that investigated a VC group test delivered remotely in the US and China (Davis et al., 2018; Ockey et al., 2019). The findings indicated favourable opinions by the test-takers about the tasks and technology, although the reliability of the technology was reported to vary considerably, especially for the China-based cohort, and is therefore an important test administration consideration as part of VC validity arguments.

Investigating Language Functions

Score investigations provide useful evidence for test comparability and can support sound conclusions. Scores are, however, but one lens for gathering evidence. Research arguments based on an integration of qualitative and quantitative data (e.g. language and score data) stand to provide a more comprehensive basis for conclusions. Such a multi-dimensional

methodological approach is no longer a novelty, and in speaking assessment has been the norm since van Lier's (1989, p. 489) now classic appeal to "obtain an insider's view" of a speaking test. The importance of focusing on the language generated in a test is also recognised in validity frameworks: Weir (2005) positions functional language as an element of the context validity of a test. O'Loughlin (2001) and Zhou (2015) have also argued for the importance of going beyond scores, and demonstrated that even if the scores between two speaking test modes are comparable, the features of the elicited language might be different, casting doubts on the equivalence of the construct measured in the two modes.

However, only a few studies have compared the language elicited in face-to-face and VC modes. Cooke (2015) conducted conversation analysis (CA) of the transcripts from face-to-face and VC modes of the IELTS Speaking test[1] in order to explore any similarities and differences in examiner/test-taker discourse under the two delivery conditions. Cooke concluded that the two modes were essentially equivalent, although less smooth turn-taking was observed, and clarification requests and subsequent reformulations were more often present in the VC mode. Cooke's findings echo the problems encountered in VC communication that are reported in the current literature on technology in speaking assessment, such as less obvious cues for turn-taking (Kern, 2014; Wang, 2004, 2006) and fewer back-channellings and interruptions (O'Conaill et al., 1993).

While detailed analysis of transcripts such as Cooke's (2015) is indeed useful, CA is extremely labour-intensive and does not allow for quantification of data (Schegloff, 1993), hindering generalisation of the results. To overcome such shortcomings, O'Sullivan et al. (2002) offered a more practical alternative in examining the elicited language in a test by using a functions checklist for test validation. Building on earlier work by Bygate (1988) and Weir

[1] For information about the IELTS test, see the research methodology section below and https://www.ielts.org/.

(1993), O'Sullivan et al. (2002) reported the development, refinement and successful application of the checklist of various language functions to the recorded performances from the Cambridge English level-based CEFR A2-C2 examinations. The authors described the language of a speaking test in terms of *informational functions* (e.g. providing personal information), *interactional functions* (e.g. asking for information) and *managing interaction functions* (e.g. changing topics). The functions checklist offers a practical tool for capturing the features of the elicited language on a speaking test and for mapping learner language against the list of language functions that test designers intended to assess. Given its advantage of allowing for quicker and more generalisable comparisons among a large number of performances, the functions checklist has been modified and used with various other speaking tests, such as the IELTS Speaking test, Trinity's GESE and ISE examinations, and British Council's Aptis Speaking test (e.g. Brooks, 2003; Inoue & Nakatsuhara, forthcoming; Zhou et al., 2018). For example, Brooks (2003) made some adjustments to the checklist to suit the specific examiner/test-taker discourse of the IELTS Speaking test and provided validation evidence that supported the 2001 revision of that test, as well as demonstrating the usefulness of the checklist for the analysis of the IELTS Speaking test. While these studies demonstrate the practicality and usefulness of the checklist, Green (2012) notes that the broad categorisations of O'Sullivan et al.'s (2002) checklist may not always result in an accurate description of learner language. It may also fail to capture variations in learner language that were used to realise each function and may have limited application in the juxtaposition of functions encountered in different speech events. This highlights the need to combine the use of the functions checklist with a more in-depth analysis of language production, thus allowing for a synergy between macro and micro perspectives.

To compare the elicited language in the face-to-face and VC modes of a large-scale international speaking test, Nakatsuhara et al. (2017a) used a modified version of O'Sullivan

et al.'s (2002) functions checklist. The language functions observed in the test-taker performances across the two modes were found to be mostly comparable, with the notable exception of *comparing* and *suggesting*, which were more frequent in the face-to-face mode, and *asking for clarification*, which was used more in the VC mode. However, as predicted by Green's criticisms of the checklist, and with the use of the functions checklist only, the authors failed to gain further insights into the use of language functions generated in the two modes, and the interpretation of their quantitative comparisons remained speculative.

In order to address the gaps identified from the literature as requiring further investigation, and also to implement recommendations from Nakatsuhara et al. (2016, 2017a), a larger scale study was instituted with the aims of conducting a robust statistical comparison of scores obtained on the two modes of delivery of the test and conducting a systematic examination of elicited language functions, complemented by a discourse analysis, thus providing further evidence of the comparability of the face-to-face and video-conferencing modes of the IELTS Speaking test.

The current study

This is a follow-up study to the exploratory study reported in Nakatsuhara et al. (2016, 2017a), which was conducted in London, United Kingdom with 32 students of mixed nationalities and 4 IELTS examiners. The study reported here aimed to replicate the initial small-scale study and gather additional validity evidence about the IELTS VC Speaking test as an alternative to the standard face-to-face test. In this respect, the study contributes a concrete example of the iterative test development process described by Chapelle and Lee (2021), where the gathering of validity evidence is systematically built into aspects of the research design.

The sample in this study comprised Chinese learners of English, who represent one of the key regions for IELTS test-takers. The use of a one-L1 sample also allowed for variables which could potentially impact on the study findings to be better controlled.

Research questions

In order to compare the construct measured by the face-to-face and VC modes of the IELTS Speaking test, this study addressed two research questions:

RQ1: Are there any differences in test scores awarded between face-to-face and video-conferencing conditions?

RQ2: Are there any differences in the types of language functions elicited under face-to-face and video-conferencing conditions?

Research methodology

Participants

Ninety-nine students from the business school of a Chinese university, 26 males (26.3%) and 73 females (73.7%), participated in this study. They had a wide range of speaking proficiency, ranging from IELTS Band 1.5 to Band 7.0, with the majority of their scores clustered around Bands 5.0, 5.5 and 6.0 (Mean = 5.11, Median = 5.50, SD = 0.97), which broadly cover CEFR B1 and B2 levels[2]. Ten trained and certified IELTS examiners (Examiners A-J) also took part in the research[3].

[2] See https://www.ielts.org/ielts-for-organisations/common-european-framework
[3] Prior to the data collection, all participants received an information sheet and signed a consent form.

Test setting

As is standard with operational IELTS tests in China, suites in a hotel in Shanghai were used as a test centre. The suites were on three floors immediately above each other with 2 suites on each floor (i.e. 6 suites in total). One floor was used for 2 x examiners and test-takers doing face-to-face tests; the other two floors were used for 2 x examiners and test-takers doing VC tests in separate rooms.

Test Materials

The IELTS Speaking test consists of three parts (see Table 1), each part fulfilling a specific role in terms of interaction pattern, task input and test-taker output.

Table 1. Structure of the IELTS Speaking test (UCLES, 2009, p. 55)

Part	Timing (mins)	Task
1	4-5	**Introduction and interview:** The examiner introduces him/herself and asks the candidate to introduce him/herself and confirm his/her identity. The examiner asks the candidate general questions on familiar topics.
2	3-4	**Individual long turn:** The examiner gives the candidate a prompt card with a topic and a few related points to the topic. The candidate is asked to talk about the topic for one to two minutes after one minute's preparation time. The examiner then asks the candidate one or two questions on the same topic to finish this part of the test.
3	4-5	**Two-way discussion:** The examiner asks the candidate further questions which are connected to the topic of Part 2. These questions give the candidate an opportunity to discuss more abstract issues and ideas.

The IELTS Scores Guide (IELTS, 2018, p. 71) lists 14 language functions that are designed to be elicited by the test tasks: *providing personal information, expressing a preference, providing non-personal information, comparing, expressing opinions, suggesting, explaining, conversation repair, contrasting, justifying opinions, narrating, paraphrasing, speculating,*

and *analysing*. This list is informed by Brooks' (2003) study based on O'Sullivan et al.'s (2002) checklist. These function categories, while wording is somewhat different, can therefore be mapped against the checklist.

Data collection

All 99 test-takers took both face-to-face and VC-delivered tests in a counter-balanced order. The VC mode of the test was administered using Zoom technology (http://www.zoom.us). Six versions of the IELTS Speaking test (*Travelling, Success, Teacher, Film, Website, Event*) were used. Examiners were instructed to use the six versions in a randomised order but to use each one relatively equally[4].

Data collection was carried out over five days, with each examiner assessing the same test-takers in each mode of delivery (i.e. each examiner assessed 12 test-takers in 24 test sessions). The face-to-face tests were filmed professionally using external cameras, and the VC tests were video-recorded using Zoom's on-screen recording technology. In total, 198 test sessions were recorded from 99 test-takers.

Examiners in the live tests in both modes awarded scores on the operational IELTS scale of 1-9 on each of the four analytic rating categories: *Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy*, and *Pronunciation* (hereafter referred to as *Fluency, Lexis, Grammar,* and *Pronunciation*). Each test session was then additionally scored by another of the ten examiners using the video-recorded performances. As such, all test sessions were scored by two examiners, and this double scoring followed a carefully developed

[4] The most used task was *Travelling* (38 times) and the least used task was *Event* (26 times). The counter-balancing of the two test modes and the randomisation of the six test versions was found to work well, as evidenced by a 2-way between-groups ANOVA which showed no significant effects of test order or versions (see Nakatsuhara et al., 2017b).

double-scoring matrix, in order to obtain sufficient overlap between examiners to allow for the application of the Many-facet Rasch Model (MFRM).

Data analysis

Examiner ratings (RQ1)

Scores awarded under each condition were compared using MFRM analyses. As will be detailed in the results section, five sets of MFRM analyses were performed using Facets 3.71.2 (Linacre, 2013) to compare all four analytic criteria together and then each criterion individually under the two delivery conditions.

First of all, to gain an overall picture of the research results, a partial credit model analysis was carried out using five facets for potential score variance: i) *test-takers,* ii) *test versions,* iii) *examiners,* iv) *test delivery modes,* and v) *rating scales.* Four additional MFRM analyses were then carried out with four facets: i) *test-takers,* ii) *test versions,* iii) *examiners,* and iv) *test delivery modes on each rating scale*. The reason for conducting the 4-facet analyses was to investigate the performance of each analytic rating scale in each mode as a separate "item" in the 4-facet analysis, allowing us to compare the effect of two modes on each of the four rating categories (e.g. F2F *Fluency* vs VC *Fluency* – i.e. test delivery mode and rating scale were combined to constitute a single facet). For the 4-facet analyses, the rating scale model was used since each rating scale in both face-to-face and VC modes should be interpreted in the same way.

Language functions (RQ2)

Due to resource constraints, 30 test-takers' performances on both test delivery modes (i.e. 60 recordings) were selected from the 99 test-takers' performances (198 performances in total for both modes) for language function analysis to examine whether or not the two modes of delivery elicited comparable language functions. The sample of 30 test-takers' performances was chosen to represent the distribution of scores of the 99 test-takers in the data set of this study while keeping in mind the distribution of scores in the operational IELTS Speaking test[5]. They included one test-taker at Band 7.0, two at Band 6.5, eleven at Band 6.0, six at Band 5.5, six at Band 5.0 and four at Band 4.5. The selected 60 recordings also involved ratings by all ten examiners.

Following the methodology used in Nakatsuhara et al. (2016, 2017a), a modified version of O'Sullivan et al.'s (2002) observation checklist was used. The modifications were informed by Brooks (2003), as well as the research team's discussions during coding training sessions. These changes were limited to clarifying or re-framing each coding category in light of the output specific to the IELTS dataset (for more details, see Nakatsuhara et al., 2016). The full inventory of the 30 functions with our modifications is summarised in Table 2.

[5] See https://www.ielts.org/for-researchers/test-statistics for test-taker statistics.

Table 2. Modified version of O'Sullivan et al.'s (2001) functions checklist

Informational functions	
	Does a Test taker…
Providing personal information (present, past, future)	Give information on present circumstances?
	Give information on past experiences?
	Give information on future plans?
Expressing opinions/preferences	Express opinions? Express preferences?
Elaborating	Elaborate on, or modify an opinion?
Justifying opinions	Express reasons for assertion s/he has made?
Comparing	Compare things/people/events?
Speculating	Speculate?
Staging	Separate out or interpret the parts of an issue?
Describing	Describe events/things/people?
Summarising	Summarise what s/he has said?
Suggesting	Suggest a particular idea?
Interactional functions	
Agreeing	Agree with an assertion made by another speaker? (apart from "yeah" or non-verbal)
Disagreeing	Disagree with what another speaker says? (apart from "no" or non-verbal)
Modifying/commenting/adding	Modify comments on arguments or comments made by another speaker? Or by the test taker in response to another speaker?
Asking for opinions	Ask for opinions?
Persuading	Attempt to persuade another person?
Asking for information	Ask for information?
Conversational repair (only self-repair)	Repair breakdowns in interaction?
Negotiating meaning	Check own understanding?
	Check other's understanding?
	Indicate understanding of point made by partner?
	Establish common ground/ purpose or strategy?
	Ask for clarification when an utterance is misheard or misinterpreted?
	Correct an utterance made by another speaker which is perceived to be incorrect or inaccurate?
	Respond to requests for clarification?
Managing interaction functions	
Initiating	Start any interactions?
Changing	Take the opportunity to change the topic?
Reciprocating	Share the responsibility for developing the interaction?
Deciding	Come to a decision?

Two researchers (who were already familiar with the checklist) watched the videos and used the list to code the elicited language functions. Since the two researchers had been previously standardised and had extensive recent experience of using the checklist, just two performances from this data set were coded as a preliminary standardisation check. Any discrepancies that arose were minor (mostly focused on where the *elaboration* function

commences) and were discussed until agreement was reached. The remaining data set was then divided into two groups and each was coded by one of the researchers.

Following the approach taken by other researchers such as Brooks (2003), Ducasse and Brown (2011), and O'Sullivan et al. (2002), the focus of the coding was on whether each function was elicited in each part of the test rather than on how many instances of each function were observed. 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 in preparation for the more detailed discourse analyses described below. The occurrence of functions in test-taker performances under the face-to-face and VC conditions was then compared using McNemar's tests. Given the purpose of this research, the Bonferroni correction was not applied to the alpha level as we thought that it was more important to avoid potential Type 2 errors.

To interpret and elaborate on the quantitative results for RQ2, the salient and notable segments of recordings identified by the researchers were then transcribed using simplified Conversation Analysis (CA) transcription conventions (Atkinson & Heritage, 1984; see Appendix for the transcription notations). While the discourse analysis carried out in this study did not employ the CA methodology of *unmotivated looking* to discover patterns or phenomena while transcribing recordings (Heritage, 1988), the use of the CA-informed transcription was found to be useful to allow for an in-depth examination of selected talk-in-interaction.

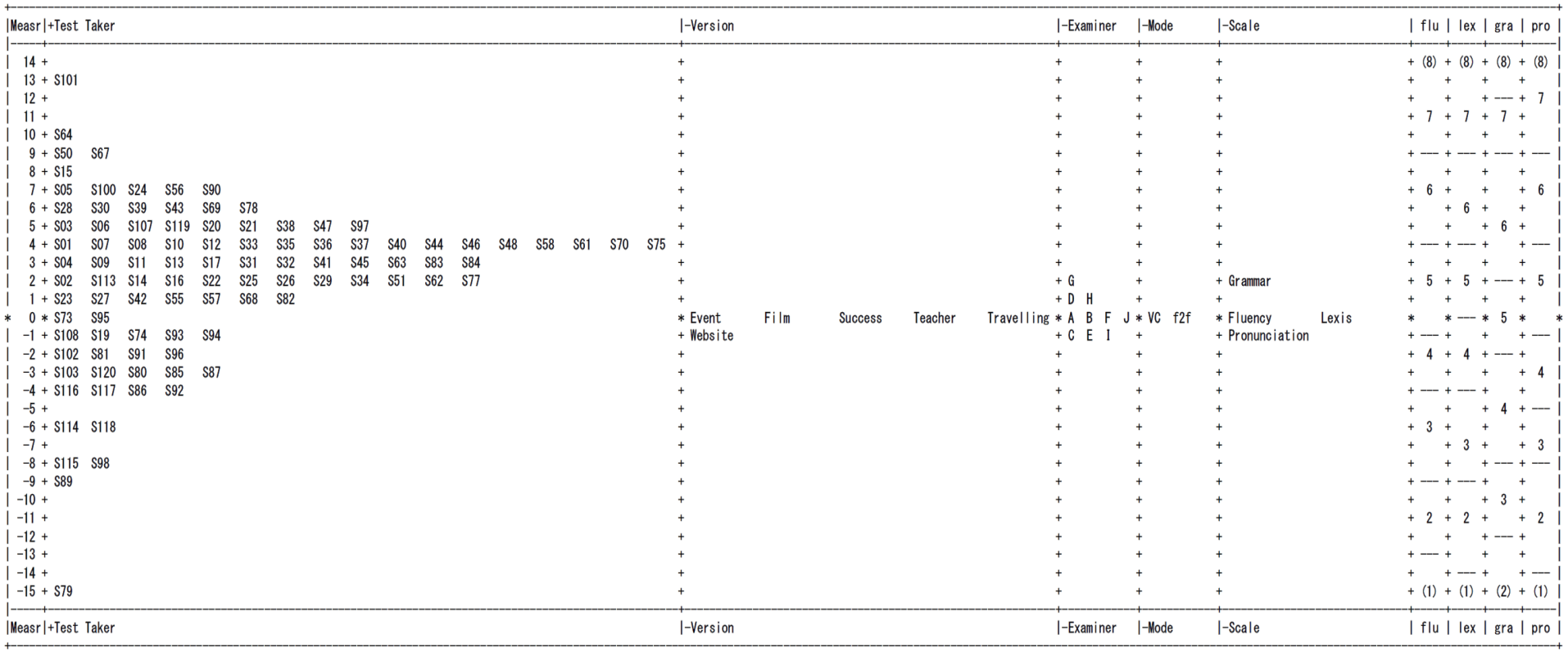
Results

Results of score analysis

Figure 1 shows the overview of the 5-facet analysis, plotting estimates of each of the five facets against a logit scale. In the figure, all facets but test-takers are negatively scaled, and the right-

hand columns (*flu*, *lex*, *gra* and *pro*) refer to the bands of the four analytic IELTS scales. From the figure, we can see that the difficulty levels of the two delivery modes are comparable.

Figure 1. All facet vertical rulers (5-facet analysis with Partial Credit Model)



For each of the facets, infit values were assessed for the model fitness. Infit values in the range of 0.5 to 1.5 are considered to be 'productive for measurement' (Wright & Linacre, 1994, p. 370), and the commonly acceptable range of Infit is from 0.7 to 1.3 (Bond & Fox, 2015). All items in the five facets fell within the acceptable range, except for one examiner (*Examiner G*) who overfitted the model, indicating that his scores were too predictable. Overfit is not productive for measurement but it does not distort or degrade the measurement system. The lack of misfit suggests that the dataset fit the MFRM, and the results were calibrated successfully on a single logit scale.

Among various Facets measurement reports, of most importance for answering *RQ1* is the delivery mode measurement report shown in Table 3. The table shows that the VC mode led to slightly lower scores than the face-to-face mode. Fixed (all same) chi-square indicates that the mode of delivery significantly affected rating scores awarded ($X^2 = 4.8$, $p = 0.03$), but the raw score difference was negligibly small (a fair average difference of 0.04 of a band).

Table 3. Test delivery mode measurement report

Delivery mode	Measure	Real S.E.	Observed Average	Fair (M) Average	Infit MnSq
F2F	-.12	.08	5.17	5.20	.89
VC	.12	.08	5.12	5.16	.89
Fixed (all same) chi-square: 4.8, d.f.: 1, significance: .03					

In the subsequent 4-facet analyses, the measurement report of each facet was assessed in the same manner as the above 5-facet analysis, and no misfitting item in any facet was identified, providing us with confidence in the accuracy of the analysis.

In the interest of space, only the measurement reports for one of the four facets, *test delivery modes on each rating scale*, are presented in Table 4. As seen in the fixed (all same)

chi-square tests, none of the score differences between the face-to-face and VC conditions were statistically significant (*Fluency* $X^2 = 0.8$, $p = 0.38$; *Lexis* $X^2 = 3.1$, $p = 0.08$; *Grammar* $X^2 = 2.1$, $p = 0.15$; *Pronunciation* $X^2 = 1.2$, $p = 0.28$).

Table 4. Four-facet analysis: Measurement reports for *test delivery modes on each rating scale*

	Measure	Real S.E.	Observed Average	Fair (M) Average	Infit MnSq
F2F *Fluency*	-.11	.17	5.11	5.10	.76
VC *Fluency*	.11	.17	5.07	5.08	.76
Fixed (all same) chi-square: .8, d.f.: 1, significance: .38					
F2F *Lexis*	-.20	.16	5.11	5.08	.70
VC *Lexis*	.20	.16	5.04	5.03	.83
Fixed (all same) chi-square: 3.1, d.f.: 1, significance: .08					
F2F *Grammar*	-.20	.20	5.23	5.21	.86
VC *Grammar*	.20	.20	5.17	5.15	.78
Fixed (all same) chi-square: 2.1, d.f.: 1, significance: .15					
F2F *Pronunciation*	-.14	.18	5.24	5.29	.84
VC *Pronunciation*	.14	.18	5.19	5.24	.73
Fixed (all same) chi-square: 1.2, d.f.: 1, significance: .28					

Results of the language functions analysis

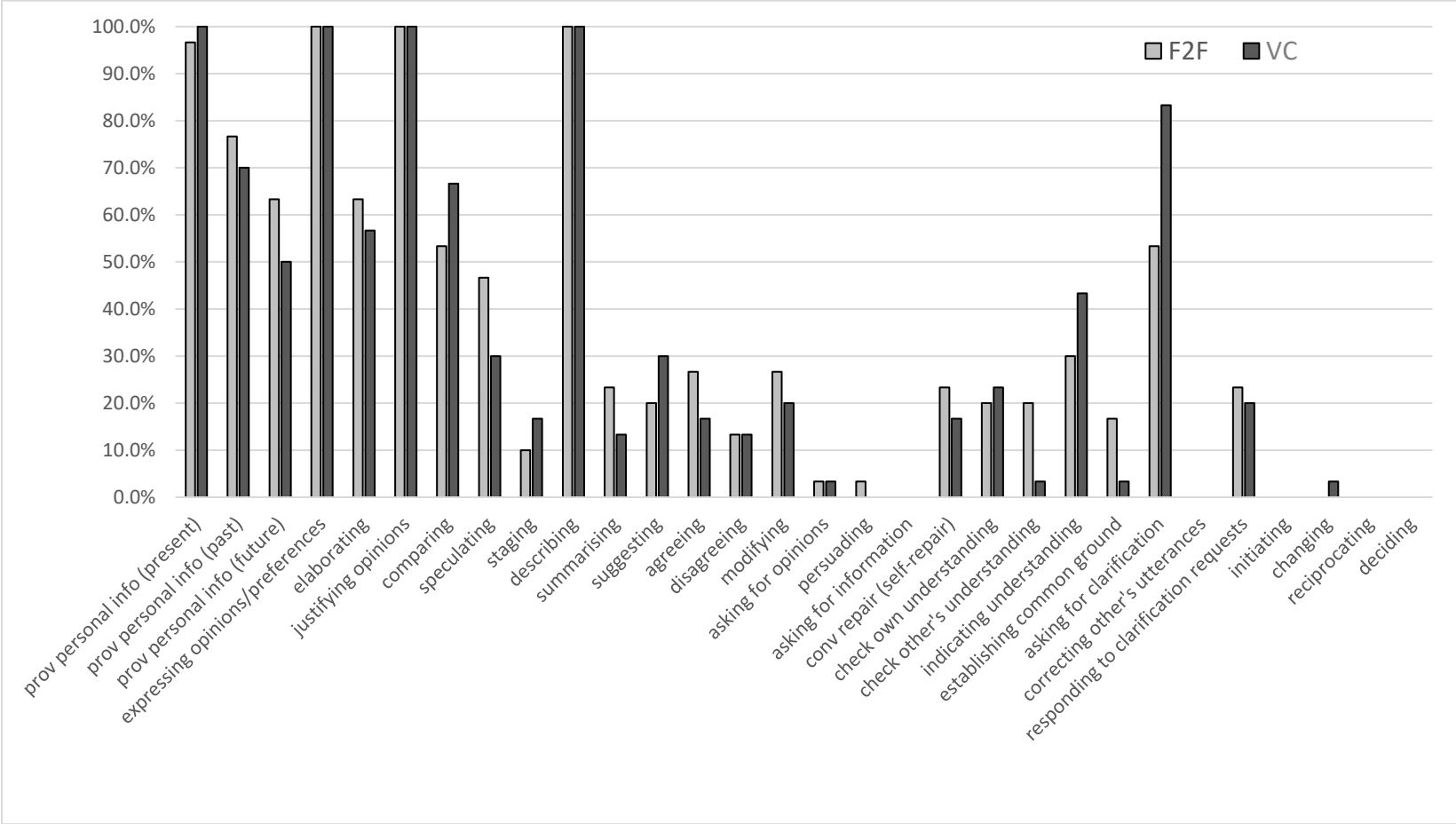
We now report on the analysis of language functions elicited through the two delivery modes, in order to address *RQ2*. Table 5 illustrates the percentage of test-takers who employed each language function at least once under the face-to-face and VC conditions across the three parts of the IELTS test, and Figure 2 visually summarises that information for the entire test. Complementing the findings of Nakatsuhara et al. (2017a), all 14 functions listed in the IELTS Score Guide (IELTS, 2018, p. 71) were observed, although some were employed only by a limited number of test-takers (e.g. *suggesting*, *conversational repair*).

Table 5. Percentage of functions use in face-to-face and VC modes in each test part (n=30 in each mode)

Function	Part 1		Part 2		Part 3	
	F2F	VC	F2F	VC	F2F	VC
provide personal info (present)	96.7%	100.0%	20.0%	16.7%	3.3%	10.0%
provide personal info (past)	36.7%	30.0%	43.3%	56.7%	13.3%	6.7%
provide personal info (future)	16.7%	26.7%	50.0%	30.0%	10.0%	-
expressing opinions/preferences	100%	100.0%	90.0%	93.3%	100.0%	96.7%
elaborating	53.3%	50.0%	50.0%	53.3%	56.7%	50.0%
justifying opinions	96.7%	93.3%	83.3%	86.7%	93.3%	93.3%
comparing	16.7%	33.3%	10.0%	10.0%	46.7%	56.7%
speculating	10.0%	6.7%	10.0%	6.7%	40.0%	23.3%
staging	3.3%	6.7%	3.3%	-	3.3%	10.0%
describing	100.0%	96.7%	100.0%	93.3%	96.7%	96.7%
summarising	6.7%	13.3%	10.0%	-	10.0%	3.3%
suggesting	13.3%	16.7%	6.7%	3.3%	-	13.3%
agreeing	13.3%	10.0%	-	-	16.7%	6.7%
disagreeing	6.7%	3.3%	-	-	6.7%	10.0%
modifying	10.0%	3.3%	-	-	23.3%	16.7%
asking for opinions	-	3.3%	-	-	3.3%	-
persuading	3.3%	-	-	-	-	-
asking for information	-	-	-	-	-	-
conversation repair (self-repair)	6.7%	6.7%	3.3%	6.7%	13.3%	3.3%
check own understanding	13.3%	6.7%	3.3%	3.3%	3.3%	20.0%
check other's understanding	13.3%	-	-	3.3%	10.0%	-
indicating understanding	20.0%	16.7%	6.7%	10.0%	13.3%	33.3%
establishing common ground	3.3%	3.3%	6.7%	-	10.0%	-
asking for clarification	26.7%	63.3%	10.0%	13.3%	36.7%	56.7%
correcting other's utterances	-	-	-	-	-	-
responding to clarification requests	-	3.3%	-	3.3%	23.3%	13.3%
initiating	-	-	-	-	-	-
changing	-	3.3%	-	-	-	-
reciprocating	-	-	-	-	-	-
deciding	-	-	-	-	-	-

Note: Differences greater than 10% are shaded.

Figure 2. Percentage of test-takers who used each language function at least once in the entire test



In Table 5, eight descriptive differences greater than 10% between the two modes are shaded – i.e. *providing personal info: past* in Part 2 (F2F<VC); *providing personal info: future* in Part 2 (F2F>VC); *comparing* in Part 1 (F2F<VC); *speculating* in Part 3 (F2F>VC); *checking own understanding* in Part 3 (F2F<VC); *indicating understanding* in Part 3 (F2F<VC); *asking for clarification* in Parts 1 and 3 (F2F<VC). Figure 2 reveals that *asking for clarification* showed the most distinctive difference (F2F<VC), when the data were combined for the entire test. However, for most of the functions, the percentages were very similar across the two delivery modes.

McNemar's tests showed that there was no significant difference between the two delivery modes in Parts 2 (Individual long turns) and 3 (Two-way discussion) of the test. The only significant difference found between the two delivery modes was for *asking for clarification* in Part 1 (Introduction) of the test (Table 6). As shown in Table 6, while 26.7% of test-takers asked one or more questions to clarify what the examiner said in the face-to-face mode, 63.3% of them asked such questions in the VC mode (mean difference = 36.6%, X^2 = 0.210, p = 0.013). This trend is consistent with the first small-scale study of this project (Nakatsuhara et al., 2017a), where a significant difference was found for *asking for clarification* in both Parts 1 and 3. Although the difference was not statistically significant, Table 5 also shows that a similar trend was observed in Part 3 where 20% more test-takers asked for clarification under the VC condition. The functions of *indicating understanding* (20.0% more in VC) and *checking own understanding* (16.7% more in VC), which are both associated with negotiation of meaning, showed similar patterns across test delivery conditions in Part 3, although these differences did not reach statistical significance.

Table 6. Language functions differently elicited in the two modes (n=30 in each mode)

[Part] Function	Test mode	Descriptive stats			McNemar's test	
		Frequency	Mean	SD	X^2 (d.f.=1)	Sig. (2-tailed)
[Part 1] Asking for clarification	Face-to-Face	8	0.267 (26.7%)	0.450	0.210	0.013
	Video-Conferencing	19	0.633 (63.3%)	0.490		

To further probe the quantitative results, discourse analysis was performed to understand in more detail the increased number of clarification requests observed under the VC condition. A few typical examples of clarification requests in the VC mode are presented below. Excerpt (1) presents the initial set of questions in Part 1, regarding 'where you live'. The test-taker was fairly fluent in answering the first question (concerning a good thing about living in her flat) (lines 1-3). When the second question was posed in line 4, it is possible that she misheard the unvoiced palatal plosive consonant /k/ as the unvoiced labiodental fricative sound /f/ in 'like'. This might have been caused by reduced clarity of the specific acoustic feature in digital communication or by reduced visual support as lip movements play a significant role in understanding pronunciation features, known as the McGurk effect (McGurk & MacDonald, 1976). Alternatively, the test-taker might have thought that 'area life' was a plausible collocation.

Excerpt (1) E: Examiner D, Test-taker: S024 (IELTS Speaking band 6.0), Part 1, Video-conferencing

1	TT:	maybe I can have in my own room in the flat, but not that big
2		but maybe my own room and some of my equipments (.) and my own
3		table, my own computer, my own TV and my own bed. Yeah.
4	E:	what is the area like where you live?
5 →	TT:	area life? ((tilting the head))
6	E:	what is the area <u>like</u> where you live?
7	TT:	oh yeah ((nodding)), ah maybe in the community (.) with uh a
8		small park, and many of block- blocks…

Not all clarification questions were used to repair communication breakdowns as in Excerpt (1). There were other examples where test-takers made clarification requests just to consolidate their understanding of the question they had heard. In Excerpt (2), the examiner asked a series of questions about photographs, and the question in line 4 followed the test-taker's response to a question about how popular photography is in her country. By observing the flow of the conversation and the way in which the test-taker asked a clarification question in line 5, it does not appear that the test-taker completely missed the examiner's question. It seems that she wished to confirm her understanding of the first part of the question 'what kind of' to ensure the relevance of her response.

Excerpt (2) E: Examiner B, Test-taker: S012 (IELTS Speaking band 6.0), Part 1, Video-conferencing

1	TT:	you can see many people will send their .hh image on the
2		WeChat, and the- uh it means that they like photo take, uh
3		self-photo or take photos.
4	E:	what kind of photos do you like (.) looking at?
5 →	TT:	.hhh I: looking at (0.5) emmm (0.5) can you (.) can you speak?
6		((smiling))
7	E:	<what kind of photos (.) do you like looking at?>
8	TT:	.hhh OK, what kind of photos, uh I like uh: photos which uh::
9		are about the:: scenery…

It is possible that the use of the *asking for clarification* function in the VC mode was the result of poor sound quality, although in this study, there were limited sound-video synchronisation problems (cf. the sound quality evaluations by observers in the test-taker rooms were: Very clear = 64.6%, Clear = 21.2%, OK = 11.1%, Not always clear = 2.0%; for more information, see Nakatsuhara et al., 2017b).

Discussion and Conclusions

This study has compared test-takers' scores and language functions elicited between two different delivery modes for the IELTS Speaking test, i.e. the standard face-to-face and VC modes. This section will discuss the implications of our findings while summarising findings relevant to the two research questions of this study.

RQ1: Are there any differences in test scores awarded between face-to-face and video-conferencing conditions?

The two modes generated essentially the same test score outcomes, regardless of the delivery mode. The MFRM analysis of four analytic scores altogether indicated that scores were marginally lower in the VC mode than the face-to-face mode, with the raw score difference being negligibly small (i.e. 0.04 of a band) and not affecting test-takers' final band scores. Four sets of rating scale model analyses also confirmed that none of the analytic categories presented a significant difference although the fair average scores in the VC mode were consistently marginally lower across all categories (e.g. *Fluency*: -0.02, *Lexis*: -0.05, *Grammar*: -0.06, *Pronunciation*: -0.05). The results of the overall difference in scores seem to relate to the effect of accumulating non-significant tendencies of the same direction. We can therefore conclude that this study has provided further corroborating evidence on score comparability between face-to-face and video-conferencing delivery modes of the IELTS Speaking test, thus confirming the results obtained by Clark and Hooshmand (1992), Kim and Craig (2012), and Nakatsuhara et al. (2017a), all of which show the VC mode resulting in non-significantly different, but marginally lower, scores across criteria.

RQ2: Are there any differences in the types of language function elicited under face-to-face and video-conferencing conditions?

In terms of the language produced in the two modes, there was one statistically significant difference in functional output in Part 1 (Interview) of the test (i.e. *asking for clarification*) compared to the three significant differences in Parts 1 (Interview) and 3 (Discussion) found in Nakatsuhara et al. (2017a) (i.e. *asking for clarification*, *suggesting*, and *comparing*). The common significant difference in both studies is *asking for clarification*, with 36.6% more test-takers asking clarification questions in Part 1 of the test in this study. Given the clarity of the sound quality reported in this study, the increased incidence of negotiation of meaning by asking for clarification is unlikely to be attributable solely to audio quality, but perhaps indicates a slight change of construct in communication under the VC mode. This suggests that signalling and solving communication breakdowns and indicating engagement and understanding (i.e. demonstrating 'interactive listening', Ducasse & Brown, 2009) during the enfolding interaction are key to successful communication in the VC mode and that they are potentially more frequent than in face-to-face contexts.

The increased number of clarification requests under the VC condition indicates that this is an attribute of the VC mode, where the sound is transmitted via computer. Although future research is called for to examine the extent to which this can be minimised with better technology, it also seems to be associated with reported difficulties for test-takers to supplement their understanding by the examiner's subtle cues, such as gestures and voice inflection, which might be less salient under the VC condition due to the limited screen size and audio frequency range (Nakatsuhara et al., 2017b; Council of Europe, 2018).

How can the nature of VC communication be reflected in speaking test tasks and administrations? The IELTS examiner script was originally developed for face-to-face tests

only, with limited flexibility allowed for examiner speech. Considering the VC-specific features this study has described and discussed, we argue that the IELTS examiner script should allow for more flexibility, for example with repeating or rephrasing questions, in order to accommodate the VC mode. This is supported by O'Sullivan and Lu's (2006, p. 22) research which determined that examiner deviation from the examiner scripts by paraphrasing questions did not negatively affect the language produced by test-takers.

Such a change in the examiner script, leading to greater flexibility in examiner speech, would allow examiners to provide scaffolding when necessary to help test-takers cope with communication breakdowns that occurred as a result of the technology supporting the VC mode. Furthermore, this would be helpful in retaining 'interactiveness' in both the face-to-face and VC tests. Brown (2007, p.138) offers a cautionary note on the tension between standardisation and interactiveness: "one way [to ensure fairness for test-takers] is to use more constrained and explicit tasks…, but the danger here is the potential loss of communicativeness, or at least interactiveness". Brown's comment on the tension between standardisation and interactiveness is also relevant when discussing further changes in the examiner script for offering a 'standardised' and yet 'interactive' test using VC technology, while noting that the VC test may not offer the same level of subtlety as in face-to-face communication, but its interactiveness seem to be characterised by more explicit and increased instances of negotiation of meaning.

Finally, in order to continue to understand the ever-changing nature of VC communication in real-life contexts, studies that go beyond a comparison between the face-to-face and VC modes of a speaking test to an investigation of the VC mode in its own right are necessary[6]. More in-depth studies focusing on aspects of interactional competence such as turn-

[6] Following recommendations from the current study, a follow-up study located in four countries across Latin America focused exclusively on the VC test, examining the scoring validity, the effect of sound quality

taking management and non-linguistic aspects of interaction would be valuable. These are fundamental mechanisms for the management of interaction and investigations of micro-interactional differences across face-to-face and VC modes would help to better define the new genre of video-conferencing speaking. The knowledge gained from such studies may ultimately define the construct future speaking tests wish to elicit. In light of the COVID-19 global health pandemic, the need to analyse successful VC communication undertaken in educational, social, and professional domains (remote-teaching degree courses, oral examinations taken online, social services, job interviews, business meetings, etc.) is more pressing than ever. Within a cycle of test validation, such an endeavour to better understand continuously evolving constructs would be critical to enable "the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, p. 13), which ultimately would benefit all stakeholders.

on performances, and the perceptions of both test-takers and examiners of the VC mode (Berry et al., 2018). Recommendations from this third study led to a further examination of the VC mode focusing on operational considerations including length of time taken for each section of the test and using a revised interlocuter script (Lee et al., 2021).

References

Atkinson, J. M., & Heritage, J. (Eds.) (1984). *Structures of social action: Studies in conversation analysis.* Cambridge University Press.

Berry, V., Nakatsuhara, F., Inoue, C., & Galaczi, E. (2018). Exploring the use of video-conferencing technology to deliver the IELTS Speaking Test: Phase 3 technical trial. *IELTS Partnership Research Papers, 2018/1*. IELTS Partners: British Council, Cambridge Assessment English, & IDP: IELTS Australia.

Bygate, M. (1988). *Speaking.* Oxford University Press.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.).* Routledge.

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

Brown. A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor, & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 98-138). UCLES/Cambridge University Press.

Chapelle, C. A., & Lee, H. (2021). Understanding validity argument in language testing. In C.A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of argument-based validation research*. (pp. 19-44). Cambridge University Press.

Clark, J. L. D., & Hooshmand, D. (1992). 'Screen-to-screen' testing: An exploratory study of oral proficiency interviewing using video conferencing. *System, 20*(3), 293-304. https://doi.org/10.1080/15434303.2016.1263637

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

Council of Europe. (2018). *Common European framework of reference for languages: learning, teaching, assessment – Companion volume with new descriptors.* htpps://www.coe.int/lang-cefr

Craig, D.A. & Kim, J. (2010). Anxiety and performance in videoconferenced and face-to-face oral interviews. *Multimedia-assisted Language Learning*, *13*(3), 9-32. https://doi.org/10.15702/MALL.2010.13.3.9

Davis, L., Timpe-Laughlin, V., Gu, L., & Ockey, G. (2018). Face-to-face speaking assessment in the digital age: Interactive speaking tasks online. In J. M. Davis, J. Norris, M. Malone, T. McKay, & Y. A. Son (Eds.), *Useful assessment and evaluation in language education* (pp. 115–130). Georgetown University Press. https://doi.org/10.2307/j.ctvvngrq.10

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–43. https://doi.org/10.1177/0265532209104669

Ducasse, A. M., & Brown, A. (2011). The role of interactive communication in IELTS Speaking and its relationship to candidates' preparedness for study or training contexts. *IELTS Research Reports, 12,* 1-26. https://www.ielts.org/-/media/research-reports/ielts-rr-volume-12-report-3.ashx

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments (2nd ed.).* Peter Lang.

Green, A. B. (2012). *Language functions revisited: Theoretical and empirical bases for*

language construct definition across the ability range. Cambridge University Press.

Heritage, J. (1988). Explanations as accounts: A conversation analytic perspective. In C. Antaki (Ed.), *Analyzing lay explanation: A case book of methods.* (pp. 127–144). Sage.

IELTS (2018). *IELTS scores guide*. https://ielts.kz/wp-content/uploads/2019/01/ielts_score-guide_a4_2018_web.pdf

Inoue, C., & Nakatsuhara, F. (forthcoming). Validation of a large-scale task-based test: Functional progression in dialogic speaking performance. In N. P. Sudharshana & L. Mukhopadhyay. (Eds.), *Task-based language teaching and assessment: Contemporary reflections from across the world*. Springer Nature.

Inoue, C., Nakatsuhara, F., Berry, V., & Galaczi, E. (forthcoming). Video-conferencing speaking tests: An investigation of context validity related to test administration. In G. Yu & J. Xu. (Eds.), *Language test validation in a digital age.* UCLES/Cambridge University Press.

Jang, E.E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34,* 123-153. https://doi.org/10.1017/S0267190514000063

Kern, R. (2014). Technology as Pharmakon: The promise and perils of the Internet for foreign language education. *The Modern Language Journal*, *98*(1), 340–357. ttps://doi.org/10.1111/j.1540-4781.2014.12065.x

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning, 25*(3), 257-275. https://doi.org/10.1080/09588221.2011.649482

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

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

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748. https://doi.org/10.1038/264746a0

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). Macmillan Publishing.

Moeller, A. J., Creswell, J. W., & Saville, N. (Eds.) (2016). *Second language assessment and mixed methods research.* UCLES/Cambridge University Press.

Nakatsuhara, F., Inoue, C., Berry, V., & 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 Assessment English, & IDP: IELTS Australia.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017a). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly, 14*(1), 1–18. https://doi.org/10.1080/15434303.2016.1263637

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017b). Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and video-conferencing

delivery (Phase 2). *IELTS Partnership Research Papers, 3*. IELTS Partners: British Council, Cambridge Assessment English, & IDP: IELTS Australia.

Ockey, G. J., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). Exploring the potential of a video-mediated interactive speaking assessment. *ETS Research Report Series, 2019*, 1-29. https://doi.org/10.1002/ets2.12240

O'Conaill, B., Whittaker, S., & Wilbur, S. (1993). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction, 8*(4), 389-428. https://doi.org/10.1207/s15327051hci0804_4

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests.* UCLES/Cambridge University Press.

O'Sullivan, B., & Lu, Y. (2006). The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports, Vol. 6* (pp. 91-117). IELTS Australia and British Council.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56. https://doi.org/10.1191%2F0265532202lt219oa

Riazi, A. M. (2017). *Mixed methods research in language teaching and learning.* Equinox Publishing.

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction, 10*(4), 401-444. https://doi.org/10.1207/s15327051hci1004_2

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction 26*(1), 99-128. https://doi.org/10.1207/s15327973rlsi2601_5

UCLES (2009). *Official IELTS practice materials.* UCLES/Cambridge English.

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23*(3), 480-508. https://doi.org/10.2307/3586922

Wang, Y. (2004). Supporting synchronous distance language learning with desktop videoconferencing. *Language Learning and Technology*, *8*(3), 90–121. http://dx.doi.org/10125/43997

Wang, Y. (2006). Negotiation of meaning in desktop videoconferencing-supported distance language learning. *ReCALL*, *18*(1), 122–146. https://doi.org/10.1017/S0958344006000814

Weir, C. J. (1993) *Understanding and developing language tests*. Prentice Hall.

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

Wright, B. D. & Linacre, J. M. (1994). *Reasonable mean-square fit values.* https://www.rasch.org/rmt/rmt83b.htm

Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia, 5*(2). https://doi.org/10.1186/s40468-014-0012-y

Zhou, Y., Dunlea, J., Negishi, M., & Yoshitomi, A. (2018). *Localisation of an international speaking test for Japanese university admission* [Paper presentation]. The 6th British

Council New Directions in English Language Assessment Conference, Kuala Lumpur, Malaysia.

Appendix: Transcription Notations

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