ISSN 2515-1703
2021/1

IELTS Partnership Research Papers

Development of the IELTS Video Call Speaking Test: Phase 4 operational research trial and overall summary of a four-phase test development cycle

Hye-won Lee, Mina Patel, Jennie Lynch and Evelina Galaczi

IELTS™
BRITISH COUNCIL idp Cambridge Assessment English

Development of the IELTS Video Call Speaking Test: Phase 4 operational research trial and overall summary of a four-phase test development cycle

This is the fourth report in a collaborative project to develop an IELTS Video Call Speaking (VCS) Test. This report investigated issues around the time taken for each part of the test, the interlocutor frame and also Examiner and test-taker perceptions of the VCS test.

Funding

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

Publishing details

Published by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia © 2021.

How to cite this paper

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

Introduction

This is the fourth report in a collaborative project undertaken by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. The very first study was conceived of in 2013 and completed in 2014. Five years later, after rigorous and robust investigation, Video Call Speaking (VCS) has been operationalised.

The previous studies progressed from a small scale exploration of delivering a high-stakes test via video-conferencing by comparing Examiner and test-taker behaviour across the two modes to a larger scale study to confirm the findings of the first study, but also to develop and trial Examiner training for delivering the Speaking Test remotely. The third study then focused solely on the video-conferencing delivery to review, revise and trial the Examiner training and to investigate in more detail technological issues related to the delivery of the test.

This fourth report, following recommendations of the previous study collected data to answer a few outstanding questions about using video-conferencing for a remote, high-stakes speaking test. This study therefore investigated issues around the time taken for each part of the test, the interlocutor frame and also Examiner and test-taker perceptions of the VCS test.

The findings of the study report timings of each part of the test and the test overall to be adequate in the VCS mode. Focus groups with Examiners revealed satisfaction with the interlocutor frame with a few minor changes. Overall test-taker perceptions of the VCS mode of delivery was positive.

This project was conceived with the intention of trying to make the IELTS Speaking Test more accessible for test-takers in areas where an in-person face-to-face test was not always possible, for example, regions made inaccessible by war, disease or simply the lack of infrastructure across vast distances. Through a systematic, iterative and extensive process involving data collection from eight countries over a period of five years the IELTS Partners have operationalised Video Call Speaking, which it is hoped will not only serve its original purpose, but also prove to be a timely innovation as global, regional and even national movements have been restricted indefinitely due to the Coronavirus pandemic.

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

Development of the IELTS Video Call Speaking Test: Phase 4 operational research trial and overall summary of a four-phase test development cycle

Abstract

Explorations into speaking assessments which maximise the benefit of technology to connect test-takers and examiners remotely and which preserve the interactional construct of the test are relatively rare. Such innovative uses of technology could contribute to the fair and equitable use of assessments in many contexts and need to be supported by a sound validity argument. To address these gaps and opportunities, an IELTS Speaking Test administered via video-call technology was developed, trialled, and validated in a four-phase research and development project.

An effort to strengthen parts of a validity argument has guided an iterative process of test development and validation, which included 595 test-takers and 32 examiners from seven global locations participating in a series of mixed methods studies. Each validation phase contributed to updating a validity argument, primarily in terms of the evaluation and explanation inferences, for the Video Call Speaking (VCS) test.

Phase 4, featured in this current report, examined some administration questions raised in the previous phase, such as time allocated in each part of the test and changes in the interlocutor frame, as well as test-taker and examiner perceptions of the VCS test. The average time taken for completion of each test task was recorded for 375 test-takers to investigate how adequate the existing timing is in the VCS mode. Ten examiners, who administered the test in this phase, were asked to respond to a questionnaire and participate in semi-structured focus groups to share their perceptions of the VCS test. Test-takers were also surveyed via a questionnaire, and additionally some of them provided more in-depth perceptions of the test during focus groups.

On the whole, the existing timing for each part was found to be adequate. Examiners perceived using the revised interlocutor frame as straightforward; however, several minor additional changes were suggested. They also perceived test-takers to be comfortable and not intimidated by the video-call mode, they found the overall test delivery quite comfortable, and overall, they perceived their rating experience as positive. A small majority of test-takers agreed that the VCS test allowed them to show their full English ability, and their perceptions about the quality of the sound were generally positive.

The report ends with a summary of the validity evidence gathered throughout the four-phase test development process, contextualised within a validity argument framework.

Authors' biodata

Hye-won Lee

Hye-won Lee is Senior Research Manager at Cambridge Assessment English where she conducts research projects related to new generation assessments and new products. Before joining Cambridge English, Hye-won gained extensive experience developing and validating digital assessments at leading organisations based in the USA. She has also taught and supervised in-service English teachers at TESOL Master's programs in South Korea. Hye-won holds a PhD in Applied Linguistics and Technology from Iowa State University, with specialisations in technology-enhanced language assessment, argument-based validation, and quantitative research methods. Her current work focuses on the use of video call technology in speaking tests and the proficiency model of language ability in data-driven diagnostic assessment

Mina Patel

Mina Patel is Assessment Research Manager with the Assessment Research Group at the British Council. Her background is in English language teaching and training. She has worked in the UK, Greece, Thailand, Sri Lanka and Malaysia as a teacher, trainer, materials developer and ELT projects manager, and has extensive experience working with Ministries of Education in East Asia. Mina has presented at numerous national and international conferences on ELT-related matters. Mina's interests in language testing and assessment lie in the areas of language assessment literacy and the impact of testing and assessment. She is currently a PhD student with CRELLA at the University of Bedfordshire, UK.

Jennie Lynch

Jennie Lynch is Head, Global Examiner Management and Marking at IDP Education. She began engagement with IELTS in 1993 as an Examiner. Jennie has been involved in international education for over 30 years initially in the ELICOS industry and later in universities as a senior lecturer in the disciplines of Academic Development and Student Learning. She was the inaugural Secretary for the Australian national Association for Academic Language and Learning (AALL) and co-editor of the *Journal of Academic Language and Learning* (JALL). Jennie holds a BA, Dip.Ed, B.Ed (TESOL) and M.Ed (TESOL).

Evelina Galaczi

Evelina Galaczi is Head of Research Strategy at Cambridge Assessment English, University of Cambridge, UK, where she leads a research team of experts in language learning, teaching and assessment. She has worked in language education for over 30 years as a teacher, teacher trainer, materials writer, researcher and assessment specialist. Evelina's expertise lies in second language assessment and learning, with a focus on speaking assessment, interactional competence, test development and the use of technologies in learning and assessment. Evelina has presented worldwide, and published in academic journals, including *Applied Linguistics, Language Assessment Quarterly, Language Testing* and *Assessment in Education*. She holds Master's and Doctorate degrees in Applied Linguistics from Columbia University, USA.

Contents

List of tables

1 Introduction

Automation of a speaking test's delivery and/or scoring under the current state of technology limits the test construct, with interactional competence often outside the scope of the underlying construct (Chun, 2006, 2008; Galaczi, 2010; Xu, 2015). In contrast, face-to-face interactional tests tap into a broader construct, but at the expense of practicality and access, due to the logistical necessity for all participants to be at the same location. A remote-delivered speaking test which maximises the ability of technology to connect test-takers and examiners remotely could preserve the interactional construct of the test, contribute to the fair and equitable use of assessments in any context, and ease logistical practicality constraints.

The delivery of direct speaking tests via video call (VC) technology is not a novel idea (Clark & Hooshmand, 1992; Craig & Kim, 2010; Kim & Craig, 2012; Ockey, Timpe-Laughlin, Davis & Gu, 2019). However, an attempt to administer it within an existing high-stakes testing program is new, and requires thorough validation exercises to achieve score and administrative comparability to the standard in-room mode and the stability of the delivery platform to prevent any potential threat to construct validity. The possibility of using VC technology in the IELTS Speaking Test has been examined under an IELTS cross-partner multi-phase research and development project (Berry, Nakatsuhara, Inoue, & Galaczi, 2018; Nakatsuhara, Inoue, Berry, & Galaczi, 2016, 2017a, 2017b), and the findings from each phase have directed the foci of the following ones.

Based on a validity argument built on evidence from the previous three studies, the final Phase 4 was embedded within British Council and IDP operational pilots and looked at specific administration-related questions raised in the previous phase which may impact on validity. One aim of this phase was to investigate whether any changes are needed to the timing of the test and the interlocutor frame (i.e., the script which examiners follow) due to the effect of new delivery mode. These changes were recommended in the Phase 3 findings (Berry et al., 2018). Another aim was to extend the evidence gathered in all three phases about test-taker and examiner perceptions about VC speaking (VCS), in order to inform specific aspects of the test delivery and platform.

As the final report in an initial validation program supporting the current version of the IELTS VCS Test, this report will end with a summary of the findings gathered in all phases of research and development contextualised by argument-based validity. We will provide an overview of how these aspects of validity evidence are woven together to support the validity argument of the IELTS Speaking Test.

2 Gathering validity evidence from operational conditions (Phase 4)

The recommendations from the previous phase guided the research questions of interest in this phase. The Phase 4 questions focused on the administrative aspects of the VCS test, as well as stakeholders' perceptions:

1. Is the existing timing for each part adequate?
2. Do examiners find the minor changes to the interlocutor frame useful?
3. What are the examiner perceptions about the VCS test mode?
4. What are the test-taker perceptions about the VCS test mode?

These administration-related factors – time constraints, interlocutor frames and other emerging ones – were examined in this operational stage of development to seek further evidence to strengthen the underlying validity argument.

2.1 Time allocation in Speaking Test tasks

In assessment task design, as crucial as it is to target the relevant language ability a test intends to measure, it is equally important to offer a setting where a sufficient amount of language can be elicited to infer the ability of a test-taker. This is of relevance for the IELTS VCS test, since decisions had to be made about whether to extend the time allowed for the test to accommodate the technical context and reduce potential unfairness.

In recognition of this importance of task setting, a body of literature has examined the effect of administration conditions such as time constraints on test-taker responses. Additionally, administration conditions have been included in validity frameworks on a par with other task considerations which impact on validity. Weir's (2005) test validation framework positions task administration within context-related validity.

Much of the discussion of time allocation in speaking tests has been focused on the length of pre-planning and its impact on test-taker performance. Whereas accumulated findings in instructed second language acquisition have demonstrated that planning prior to a language task benefits second language (L2) speech production in terms of fluency (e.g., Bui & Huang, 2016) and complexity (e.g., Yuan & Ellis, 2003), mixed findings have been obtained in a testing situation. A few studies have shown that some length of planning time helps in responding to cognitively demanding tasks such as graph description (Xi, 2005, 2010) and improving the quantity, as well as quality, of test-taker responses (Li, Chen, & Sun, 2015), or is positively perceived by test-takers although it does not have an actual impact on their performance (Elder, Iwashita, & McNamara, 2002; Elder & Wigglesworth, 2006; Wigglesworth & Elder, 2010). However, other studies have reported either null or negative effects of planning time. For instance, as part of comprehensive analyses to investigate the relationship between task characteristics/conditions and the level of difficulty and performance, Iwashita, McNamara, and Elder (2001) found that the variable of planning time does not influence task performance. In recent studies with paired/group oral assessment tasks, some found no effects on test scores (e.g. Nitta & Nakatsuhara, 2014) and others, negative effects on the quality of test performance (Nitta & Nakatsuhara, 2014; Lam, 2019).

Perhaps these conflicting results are attributable to an intricate interaction with associating factors such as test-takers' proficiency levels in the target language and the task type they complete. In Wigglesworth (1997), high-proficiency test-takers benefited from planning time in terms of accuracy on some measures where the cognitive demand was high. In contrast, O'Grady (2019) found that it was low-proficiency test-takers whose scores significantly increased as more planning time was given, and increases in scores were larger on the picture-based narrative tasks than on the non-picture-based description tasks.

Compared to the constraints of planning time, little research has been conducted into response time. In one study by Weir, O'Sullivan and Horai (2006), it was found that the amount of speech expected from the time allotted to the task did not have a significant effect on the score achieved by the high and borderline test-takers, whereas reducing the task time produced a lower mean score for the low-proficiency test-takers.

Although these decades of research have produced mixed results, the amount of time allocated to accomplishing speaking test tasks appears to have some impact on the performance of at least some test-takers under certain task conditions. Allocated time, therefore, needs to be considered as one of the important factors to pay attention to in the development of a valid and fair task.

2.2 Standardisation through the interlocutor frame

In response to a body of research highlighting issues with consistency in interlocutor behaviour and its potential impact on test-taker ratings and test fairness (e.g., Brown, 2003; Brown & Hill, 1998; Taylor, 2000), the IELTS Speaking Test was redesigned in 2001 to a more tightly scripted format using an interlocutor frame. In follow-up studies after the change, large-scale questionnaire responses demonstrated that the revision was perceived positively by examiners, but that some concerns about the lack of flexibility in wording prompts were also reported (Brown & Taylor, 2006). Another study by O'Sullivan and Lu (2006) showed that, contrary to examiners' tendency to sometimes deviate from the interlocutor frame (Lazaraton, 1992, 2002), few deviations were noted among the 62 recordings of the IELTS Speaking Test performance included in the analysis, and when deviations did occur, such as paraphrasing questions, the impact on test-taker language appeared to be minimal.

Standardisation across testing events for a fair and equitable test is the main driver behind the introduction of the interlocutor frame. However, the very nature of interaction in oral communication may be incompatible with the rigid control of discourse, as found in the studies summarised above. Interlocutor scripts which reflect the context of interaction as much as possible can minimise this incompatibility dilemma, and this might be even more so when it comes to a test delivered online via video-conferencing technology. Careful consideration, therefore, needs to be placed on potential modification of the existing interlocutor frame to cater for some unique features of tests conducted in the video-conferencing environment.

3 Methodology

3.1 Participants

3.1.1 Test-takers

In total, 375 test-takers participated in the current research study which took place from May to June 2019. The test-takers sat the IELTS Video Call Speaking (VCS) test offered in test centres and delivered on either of the partner-specific test platforms – 126 test-takers from Chongqing, China on the British Council platform and 249 test-takers from Chandigarh, India on the IDP platform.

The ages for the majority of test-takers were between 16 and 25 years old (81.7% for British Council and 96.0% for IDP). Within this range, those between 19 and 21 years accounted for 44.2% of the British Council test-takers and 37.8% of the IDP test-takers, while the younger age group of 16 to 18 years accounted for 20.0% of the British Council test-takers and 30.5% of the IDP test-takers. A larger number of test-takers (65.0%) were female in the British Council trials, whereas 65.9% of the IDP test-takers were male.

The range of IELTS scores on the IELTS VCS test was from Bands 3.5 to 8.5 for the British Council test-takers (M = 5.62, SD = 0.76) and Bands 3.5 to 7.5 for the IDP test-takers (M = 5.80, SD = 0.73). The majority of the scores (82.5% for British Council and 74.0% for IDP) were clustered around Bands 5 and 6.

Since experience with the Internet and VC technology is an important participant variable in this study, information was gathered on the test-takers' use of those technological tools in some of their daily contexts (see Table 1).

Table 1: *Test-takers' experience with the Internet and VC technology (N = 369*), mean and standard deviation (in brackets)*

	1. Never; 2. 1–3 times a month; 3. 1–2 times a week; 4. 5 times a week; 5. Every day	British Council (n = 120*)	IDP (n = 249)	Total (N = 369)
Q1	How often do you use the Internet socially to get in touch with people?	4.79 (0.65)	4.00 (1.42)	4.25 (1.28)
Q2	How often do you use the Internet for your studies?	4.69 (0.66)	3.47 (1.52)	3.86 (1.42)
Q3	How often do you use the Internet for your work?	4.34 (1.17)	2.22 (1.55)	2.91 (1.75)
Q4	How often do you use VC (e.g. Skype, WeChat, FaceTime) socially to communicate with people?	2.97 (1.19)	2.49 (1.22)	2.64 (1.23)
Q5	How often do you use VC (e.g. Skype, WeChat, FaceTime) for your studies?	1.97 (1.16)	2.17 (1.31)	2.10 (1.27)
Q6	How often do you use VC (e.g. Skype, WeChat, FaceTime) for your work?	1.97 (1.21)	1.68 (1.16)	1.78 (1.18)

* Responses from six of the British Council test-takers are missing

Both the British Council and the IDP test-takers reported that they use the Internet, on average, almost every day to socially engage with other people (M = 4.79, SD = 0.65 for British Council; M = 4.00, SD = 1.42 for IDP). For either studies or work, the British Council test-takers are online almost every day as well (M = 4.69, SD = 0.66 for studies; M = 4.34, SD = 1.17 for work), and the IDP test-takers are online a few times a week for studies (M = 3.47, SD = 1.52) and a few times a month for work (M = 2.22, SD = 1.55). With regard to using VC technology specifically, both the British Council and the IDP test-takers use the technology around once a week for a social purpose (M = 2.97, SD = 1.19 for British Council; M = 2.49, SD = 1.22 for IDP) and a few times per month for either their studies or work (means ranging from 1.68 to 2.17 across the two groups). Overall, the test-takers of the current study can be considered to be familiar with using the Internet and VC technology in their daily lives, and therefore not to be negatively affected by this new test delivery mode in their speaking performance.

3.1.2 Examiners

Six certified British Council examiners and four certified IDP examiners were chosen by each partner to administer the IELTS VCS test for the study. The examiners, in general, had extensive experience of teaching English as a second/foreign language: 9.3 years for the British Council examiners and 10.5 years for the IDP examiners. They were also very experienced in examining IELTS, with an average of 7.5 years, ranging from 3.2 years to 15 years for the British Council examiners, and from 1.6 years to 11.8 years for the IDP examiners.

The British Council examiners delivered the test from a test centre in Beijing, China over four days (7–8 and 13–14 May 2019) and the IDP examiners from one in Hyderabad, India over three days (5–7 June 2019). Prior to the trials, they had one day of training (6 May 2019 for British Council and 4 June 2019 for IDP) to understand the differences between the in-room and VCS test and to practise using the technology.

The examiners were also asked about their use of the Internet and VC technology in their social and teaching contexts (see Table 2).

Table 2: *Examiners' experience with the Internet and VC technology (N = 9*), mean and standard deviation (in brackets)*

	1. Never; 2. 1–3 times a month; 3. 1–2 times a week; 4. 5 times a week; 5. Every day	British Council (n = 5*)	IDP (n = 4)	Total (N = 9)
Q1	How often do you use the Internet socially to get in touch with people?	4.60 (0.89)	4.00 (1.41)	4.33 (1.12)
Q2	How often do you use the Internet in your teaching?	1.20 (0.45)	2.25 (1.89)	1.67 (1.32)
Q3	How often do you use VC (e.g. Skype, WeChat, FaceTime) socially to communicate with people?	3.40 (0.89)	3.25 (1.50)	3.33 (1.12)
Q4	How often do you use VC (e.g. Skype, WeChat, FaceTime) in your teaching?	1.00 (0.00)	1.75 (0.96)	1.33 (0.71)

* There was one incomplete questionnaire and therefore his/her survey responses were not included in the data set.

Both the British Council and the IDP examiners reported they use the Internet for social purposes almost every day (M = 4.60, SD = 0.89 for British Council; M = 4.00, SD = 1.41 for IDP) and VC technology a few times a week (M = 3.40, SD = 0.89 for British Council; M = 3.25, SD = 1.50 for IDP). In their teaching contexts, technology use, either the Internet or VC technology, was reported as less frequent – a few times a month for the Internet (M = 2.25, SD = 1.89 for IDP) to 'never' for VC technology (M = 1.00, SD = 0.00 for British Council). Considering the fact that the examiners are quite familiar with the use of the Internet and VC technology in their social contexts, it can be assumed that they can transfer the knowledge and skills to the testing context with the support of a one-day training session.

3.2 Materials

3.2.1 VCS examiner script and test tasks

To reflect slight modifications to the administration setting delivered via video call, a few sentences were revised or added to the standard interlocutor frame used in the in-room speaking test. For the assessed parts, 10 frames for Part 1 and five versions for Parts 2 and 3 were provided by Cambridge Assessment English for the current phase of the study and used during the trials. On the day of the trials, the prompts were randomly chosen by the examiners in each test session.

3.2.2 Partner-specific test platforms

British Council and IDP developed their own IELTS VCS test platform. Although the core platform features and processes such as test content and format are the same between the two test platforms, some minor details slightly differ as a result of the platform interface. The British Council platform is bespoke-built for the VCS test and uses Zoom, a commercial communication software program, as its VC technology to connect the test-taker and the examiner, whereas the IDP platform uses Zoom Rooms. Some of the differences in the platforms and processes are described below in Table 3.

Table 3: *Some key differences in the test platforms and processes during the pilots*

	British Council	IDP
Test-taker login	Invigilator logs in for test-taker	Test-taker name and number appearing on screen at a scheduled time
Headphones	Both examiners and test-takers have headphones on	Examiners and test-takers do not have headphones on*
Interlocutor frame	On screen	On paper
Task 2 card for test-taker	Pushed to test-taker on click of button by examiner	Pushed to test-taker by screen sharing
Task 2 card view	Task card on screen, covering approx. two-thirds of the screen	Task card on screen, covering approx. two-thirds of the screen
Examiner view	Examiner can see him/herself throughout the test	Examiner can see him/herself only during Part 2
Test-taker view	Test-taker cannot see him/herself at all during the test	Test-taker can see him/herself throughout the test
Rating	On screen	On paper

* Shortly after the pilot, IDP ran an additional small-scale study to compare sound quality with and without headphones, and as a result of questionnaire and anecdotal feedback, decided to require both test-takers and examiners to wear headsets in all future VCS test sessions.

3.3 Data collection procedures

3.3.1 Test preparation

Prior to the trial, a one-day examiner training session for administering and rating VCS tests was conducted to explain the differences between the in-room and VCS test and to practice using the technology. Invigilators were given a familiarisation session on the test process and procedures. On the day they were present throughout the test sessions and provided any procedural or technical assistance if needed. On the test day, test-takers were given VCS test guidelines to familiarise themselves with video-call delivered tests.

3.3.2 Timing data

To investigate the first research question regarding the timing for each of the three parts, the time spent for each interview was recorded in minutes and seconds, automatically in the British Council bespoke platform, and manually by trained human timekeepers for the tests conducted in the IDP platform, which does not have an automatic facility to record time. The timekeepers used a mobile phone timer function to keep accurate times. The times for all test-takers were transcribed onto a spreadsheet for analysis. Ideally, timing data would have been collected from in-person tests as well, to allow comparison, but due to resource and time constraints, this was beyond the scope of the study.

3.3.3 Examiner feedback questionnaires

On the completion of all the assigned test sessions, the examiners responded to a questionnaire regarding their perceptions of test administration and rating (see Appendix 1). They were encouraged to note in writing any important points, both positive and negative, in between tests, which became the basis of their responses. The questionnaire consisted of four parts. Part 1 (Q1–Q4) asked about the examiners' general background and experience with the Internet and video-conferencing technology (see Table 2 for the summary of the results). Part 2 (Q5–Q7) concerned their perceptions about delivering each part of the test, including handling Part 2 task prompts on the screen and managing the modified interlocutor frame. Part 3 (Q8–Q10) related to the perceived adequacy of the time assigned to each part, and lastly, Part 4 (Q11–Q13) with regards to applying the IELTS Speaking band descriptors in rating the VCS test.

Parts 1 to 4 were followed by two open-ended questions (Q14–Q15) on any significant differences they noticed between the VCS and in-room test for test-takers and themselves. The entire questionnaire took approximately 15 minutes on average.

3.3.4 Test-taker feedback questionnaires

Each test-taker was asked to complete a brief questionnaire (see Appendix 2) after the test. As English is not used as one of the country's official languages in China, the questionnaire items were translated into Chinese by a qualified British Council staff member to assist the test-takers' understanding and valid responses. The translations were verified by another qualified bilingual colleague at Cambridge Assessment English and presented next to the original English items (see Appendix 3). This bilingual version was used for the British Council trials, and the test-takers were given an option to provide their short responses to the open-ended items in either English or Chinese. The responses given in Chinese were translated into English for analysis by a qualified British Council staff member, and the accuracy of the translations was verified by a bilingual colleague at Cambridge Assessment English.

The questionnaire consisted of eight items. The first two questions (Q1–Q2) asked about the test-takers' experience with the Internet and video-conferencing technology (see Table 1 for the summary of the results). The next five questions (Q3–7) concerned the test-takers' perceptions on the VCS test, ranging from the quality of the sound to the Part 2 prompt on the screen. The last question, Q8, was an open-ended question regarding any other positive or negative points about the VCS test. The questionnaire took from five to 10 minutes to complete.

3.3.5 Examiner focus group discussions

After completing the questionnaire on the last day of the pilot, all the examiners were invited to focus group discussions to elaborate more on their questionnaire responses and share any other reactions to the test. Two cohorts of the three British Council examiners participated each on 9 May and 16 May 2019, and one group of the four IDP examiners participated on 7 June 2019. The discussions were facilitated by one of the researchers for the British Council pilot and by a PSN Manager for the IDP pilot, and semi-structured by the pre-arranged protocol among the IELTS Partners (see Appendix 4). The topics focused on test administration and rating in general, and specifically the timing and the Part 2 task prompts on the screen. In each focus group session, notes were taken by an additional local staff member, and audio-recorded and transcribed for analysis.

3.3.6 Test-taker focus group discussions

As an optional source of data for richer interpretation, British Council conducted focus group discussions with a few of the test-takers who volunteered. Eight semi-structured sessions were held over four days with 42 test-takers in total in eight sessions, using the pre-agreed protocol among the partners (see Appendix 5). The test-takers were asked about their overall experience with the test including the Part 2 task prompts on the screen and interaction with the examiner. One of the researchers facilitated the discussions in English with the presence of a local British Council staff member bilingual in Chinese and English. The test-takers were given a choice of whether to speak in English or Chinese, and when necessary, the facilitator's question was translated into Chinese by the bilingual colleague. The entire sessions were audio-recorded, but only the English parts were transcribed for analysis.

3.3.7 Additional IDP trial with headsets

During the operational pilots, IDP recognised the feedback on sound was not positive, so it conducted a follow-up small-scale pilot with headsets.

Seventeen test-takers took the VCS test twice – once without and once with headsets. There were two examiners, who had also been involved in the original pilot. After the additional trials with headsets, the test-takers responded to an abridged test-taker feedback questionnaire consisting only of the items relevant to sound quality (Q11–Q13 and Q15–Q16 from the original test-taker feedback questionnaire). A new item asking the test-takers' preferences for wearing headsets was also added. The summary of this trial is included in Appendix 6.

3.4 Data analysis

3.4.1 Timing data

The times taken to administer each test were analysed to investigate if the existing timing for each part of the test is adequate (Research Question 1, RQ1). The descriptive statistics, such as means to gauge the overall tendency of the data and standard deviations to understand the variation of the data, were calculated both per part and overall tallying all three parts, and compared to the existing timing – Part 1: four to five minutes, Part 2: three to four minutes (including one minute preparation time and one to two minutes test-taker talking time), Part 3: four to five minutes, and 11 to 14 minutes in total. The averages outside the set range were considered a point of further investigation.

3.4.2 Examiner feedback questionnaires

The ratings on a five-point Likert scale and written responses to the open-ended items were analysed to examine the first three research questions: RQ1 about the timing; RQ2 about the minor changes to the interlocutor frame; and RQ3 about examiner perceptions of the test. The means and standard deviations of the quantitative rating data were calculated to understand the overall trend among the examiners. The qualitative written responses were thematically analysed and used to illuminate and supplement the interpretations of the numeric data.

3.4.3 Test-taker feedback questionnaires

The ratings on a five-point Likert scale and written responses to the open-ended items were analysed to examine primarily the last research question (RQ4): *What are the test-taker perceptions about the video call speaking test mode?* The means and standard deviations of the ratings were calculated to investigate the overall perceptions about the VCS test on a group level; the written responses to the open-ended items were thematically analysed to better understand and triangulate the interpretation of the quantitative findings.

3.4.4 Examiner focus group discussions

The transcripts of all three focus group sessions were carefully read by two of the researchers individually, and analysed thematically for any recurring themes to inform RQs 1 to 3. The two researchers then convened, compared the individually identified themes such as the issue of fiddling with a pencil and paper, and agreed on which points to report as key findings from the examiner focus group data.

3.4.5 Test-taker focus group discussions

The transcripts of all eight British Council test-taker focus groups were also thematically analysed by the same two researchers, first individually and then in pairs, primarily to inform the last research question (RQ4) and to a lesser extent the others. The interpretations were carefully made so as not to overgeneralise the findings, given that some of the comments made by the test-takers may be applicable only to the specific features of the British Council bespoke test platform.

4. Results and discussion

4.1 Length of tasks

The first research question concerned whether the existing timing for each part of the test is adequate under the new delivery mode. This potential concern was raised by the examiners who participated in the previous phase of research because a slight procedural change, such as presenting a Part 2 prompt card on screen may necessitate a little more operating time, possibly requiring more overall test time to be allotted in a VC mode than in the in-room mode.

The test, consisting of three parts, is currently designed to take 11 to 14 minutes in total:

- Part 1 (introduction and interview) four to five minutes
- Part 2 (long turn) three to four minutes (including one minute of preparation time and one to two minutes test-taker talking time)
- Part 3 (discussion) four to five minutes.

Minutes and seconds spent for each part of the VCS test were measured to examine whether the actual time spent, on average, falls within an acceptable range. Table 4 presents the average time spent for each part and for all three parts together.

Table 4: *Average time spent for each part of the test and for the entire test (N = 371*), mean and standard deviation (in brackets)*

	Part 1 (4–5 mins recommended)	Part 2 (3–4 mins recommended)	Part 2: Test-taker talking (1–2 mins recommended)	Part 3 (4–5 mins recommended)	Parts 1–3 (11–14 mins recommended)
British Council (n = 126)	04:49 (00:14)	04:05 (00:19)	02:17 (00:16)	05:16 (00:30)	13:54 (01:25)
IDP (n = 245*)	04:53 (00:11)	03:56 (00:21)	02:00 (00:07)	04:56 (00:16)	13:44 (00:23)
Total (n = 371)	04:52 (00:12)	03:59 (00:21)	02:06 (00:14)	05:02 (00:23)	13:47 (0:53)

*The timing records for four of the IDP test-takers are missing.

On the whole, Part 1 took less than the upper limit of the set time range (04:52 minutes). This was also observed for the timing in both the British Council and IDP platforms. The total time for Part 2 was less than the upper limit of the set time range (03:59 minutes), and the test-taker talking time in Part 2 took six seconds longer than allocated on average. Additionally, the British Council examiners overall went five seconds over in Part 2, and 17 seconds over in terms of candidate talking time. Given that the average test-taker talking time was over the suggested range, it seems that the test-taker talking time was not sacrificed for handling the prompt card online. Part 3, on average, took two seconds longer than five minutes, the maximum time allotted. British Council examiners showed a tendency to spend more time in Part 3 than allocated (05:16 minutes).

The slight time differences between the British Council and the IDP interviews may have been due to operational differences between the two platforms, based on the additional analysis the British Council carried out to further examine the phenomenon. The British Council took a sample of the timing data that went over five minutes for Part 3 (35 cases selected), manually timed Part 3 of those, and calculated differences between the manual timing and the automated one generated by the test platform (see Appendix 7 for raw data). Differences from six seconds to 03:37 minutes were found (M = 00:26, SD = 01:07)

between when the examiners actually finished the test (by following the interlocutor frame) and when they actually ended the test using the button built on the test platform. Based on this finding from the small-scale post-hoc analysis, there are no causes for concern regarding the timing of the test in the VC mode and the potential need to allow longer time. Nevertheless, it is suggested that questions of timing are addressed in future training.

As the reviewed literature has shown that a test-taker's proficiency level may be an intervening factor on the amount of task time needed, the average time spent in each part of the IELTS VCS test was calculated for three proficiency groups. Test-takers were grouped according to their band scores assigned to the VCS test: low (below Band 5, n = 91), middle (between Band 5 and Band 6, n = 216), and high (Band 6 and above, n = 57). Table 5 shows the descriptive statistics and test statistics (H), to compare the three groups.

Table 5: *Average time spent for each part of the test and for the entire test (N = 364*), mean, standard deviation (in brackets), and statistical comparison across proficiency groups*

	Part 1 (4–5 mins recommended)	Part 2 (3–4 mins recommended)	Part 2: Test-taker talking (1–2 mins recommended)	Part 3 (4–5 mins recommended)	Parts 1–3 (11–14 mins recommended)
Low (n = 91)	04:52 (00:11)	04:00 (00:19)	02:09 (00:13)	05:02 (00:15)	13:55 (00:26)
Middle (n = 216)	04:51 (00:13)	03:59 (00:22)	02:05 (00:15)	05:03 (00:28)	13:53 (00:30)
High (n = 57)	04:52 (00:12)	03:56 (00:20)	02:03 (00:08)	05:01 (00:12)	13:49 (00:24)
Kruskal-Wallis Test	$H(2) = 1.081$, $p < 0.582$	$H(2) = 2.464$, $p < 0.292$	$H(2) = 8.594$, $p < 0.014$*, $\eta^2 = 0.013$	$H(2) = 0.429$, $p < 0.807$	$H(2) = 1.431$, $p < 0.489$

* The band scores for six of the British Council test-takers and one of the IDP test-takers are missing.

One statistically significant difference was found among the average time taken for the Part 2 test-taker talking time by different proficiency groups ($H(2) = 8.594$, $p = 0.014$), but the effect size was negligible in magnitude ($\eta^2 = 0.013$). Taken together with the fact that the other parts and the overall test did not yield any group differences, it can be interpreted that the level of proficiency did not have a meaningful impact on timing.

In terms of examiner perceptions about the timing of the different parts, interestingly, the examiners who used the IDP platform perceived the time assigned to Parts 1 and 2 as slightly less adequate (M = 4.25, SD = 0.50 for both) than the time assigned to Part 3 (M = 4.50, SD = 0.58) (see Table 6). For the examiners who used the British Council platform, the existing timing was perceived as fully adequate for all three parts (M = 4.60, SD = 0.55 for Part 1; M = 5.00, SD = 0.00 for Parts 2 and 3). In general, the examiners in both groups together perceived the existing timing as adequate for each part – averaged means ranging from 4.44 to 4.78 (Table 6).

Table 6: *Results of the examiner feedback questionnaire on the timing of the test (N = 9), mean and standard deviation (in brackets)*

1.Strongly disagree – 3.Neutral – 5.Strongly agree	British Council (n = 5)	IDP (n = 4)	Total (N = 9)
The time assigned to Part 1 of the video conference test I just administered was adequate to deliver all the requirements.	4.60 (0.55)	4.25 (0.50)	4.44 (0.53)
The time assigned to Part 2 of the test was adequate to deliver all the requirements.	5.00 (0.00)	4.25 (0.50)	4.67 (0.50)
The time assigned to Part 3 of the test was adequate to deliver all the requirements.	5.00 (0.00)	4.50 (0.58)	4.78 (0.44)

This positive perception, identified in the questionnaire, is corroborated by discussions during the examiner focus groups. Overall, the examiners did not report major problems with the timing of the individual parts of the test or the test as a whole, although a few individual examiners mentioned occasions where they sometimes struggled to finish Part 1 or not having time to ask the rounding off question for Part 2. However, considering the questionnaire and focus group data together with the platform data collected on the timing of individual parts of the test and the test as a whole, it appears that timing was within the allocations stated in the official instructions to the examiners.

4.2 Changes to the interlocutor frame

The second research question concerned whether the examiners found the minor changes to the interlocutor frame useful. As with the previous research phase, some minor functional changes were made to accommodate the medium of the test, such as in Part 2 when the prompt card for the test-taker appears on the screen rather than being handed over by the examiner.

On the whole, the examiners found managing and using the revised interlocutor frame quite straightforward (M = 4.44, SD = 0.53; see Table 7).

Table 7: *Results of the examiner feedback questionnaire on the interlocutor frame (N = 9), mean and standard deviation*

1.Strongly disagree – 3.Neutral – 5.Strongly agree	British Council (n = 5)	IDP (n = 4)	Total (N = 9)
The examiner's interlocutor frame was straightforward to manage and use in the test.	4.40 (0.55)	4.50 (0.58)	4.44 (0.53)

Additionally, the focus group discussion elicited suggestions from all the examiners about further changes to the interlocutor frame, which may improve the test experience. The following points were highlighted by the examiners, some of which were also emphasised in the test-taker focus groups:

- The examiners felt that they wanted a brief linguistic turn before the actual tests began in order to build rapport with the test-taker. This is not in the in-person interlocutor frame. In both the focus group and in the open-ended response to the questionnaire, examiners stated that when they bring the test-taker into the room and greet them, they have a brief opportunity to gauge how the test-taker is feeling, but the VC mode does not allow for that. They recommended something brief and standardised be built into the interlocutor frame. Test-takers in one of the focus groups also mentioned that this might help them to feel more at ease with the mode of delivery.

- Before the VCS test, test-takers were given guidelines of what to expect when they entered the room and a list of 'Dos and Don'ts'. On the guidelines, the test-takers were asked not to touch the pen/pencil on the table until Part 2 of the test when the examiner instructs them to do so. During both the British Council and IDP pilots, there were occasions when test-takers 'fiddled' with the pen/pencil and paper when they should not have. On these occasions, examiners were not sure what to do. There is no guidance for this in the training or in the interlocutor frame, and the recommendation from examiners is that there should be some flexibility in the script to allow them to stop this from happening, which was also echoed in one of the open-ended responses to the examiner feedback questionnaire:
  '…would need some system in place or permission to say something if the candidate fiddles with paper or pencil etc. during the test which might be intrusive'
  (IDP Examiner B, open-ended response).

- The examiners expressed that the end of the interview seemed to be left unfinished. The examiners, like the test-takers, were not always sure what to do. So, perhaps a scripted, 'You may leave the room now' as well as, 'Thank you and goodbye' at the end of the test would provide the necessary formal but polite direction for the test-taker.

4.3 Examiners' perceptions of the VCS test

The examiners' perceptions about the VCS test, including test-taker comfort, test delivery, and rating, were also investigated. The following sub-sections will discuss the findings from the examiner feedback questionnaire and the focus group discussions on these strands.

4.3.1 Test-taker comfort with the VCS test

Both the British Council and IDP examiners, in the focus groups, perceived the test-takers to be comfortable and not intimidated by the VC mode, firstly because the examiner was not in the room and secondly because the test-takers are used to communication via technology.

4.3.2 Test delivery

As shown in Table 8, the examiners found the overall test delivery straightforward (M = 4.22, SD = 0.44).

Table 8: *Results of the examiner feedback questionnaire on test delivery (N = 9), mean and standard deviation (in brackets)*

1.Strongly disagree – 3.Neutral – 5.Strongly agree	British Council (n = 5)	IDP (n = 4)	Total (N = 9)
I found it straightforward to deliver Part 1 (frames) of the video conference test I just administered.	4.60 (0.55)	4.50 (0.58)	4.56 (0.53)
I found it straightforward to deliver Part 2 (long turn) of the test.	4.00 (0.71)	4.50 (0.58)	4.22 (0.67)
I found it easy to handle task prompts on the screen in Part 2 of the test.	4.00 (0.71)	4.75 (0.50)	4.33 (0.71)
I found it straightforward to deliver Part 3 (two-way discussion) of the test.	4.80 (0.45)	4.25 (0.50)	4.56 (0.53)
Overall I felt comfortable in delivering the test.	4.20 (0.45)	4.25 (0.50)	4.22 (0.44)

Open-ended responses to the questionnaire provided insights which corroborated this finding:

- '…clear Audi*/visual link, procedure easy to do' (British Council Examiner C, open-ended response) (note: *typo in the original quote)
- 'Overall…the program is well-designed & user-friendly' (British Council Examiner E, open-ended response)
- '…was comfortable delivering the entire test' (IDP Examiner B, open-ended response)
- 'The overall experience was good.' (IDP Examiner C, open-ended response)

However, a number of issues brought up by examiners needed further consideration.

Exaggerated gestures in the VC mode

Some examiners perceived that interactions with the test-takers were limited in the VC mode, noting several potential issues regarding exaggerating gestures to make them more noticeable. For example:

> 'I felt that the interaction between the examiner and candidate is more subdued in the VC mode; delivering requires more physical effort and strain.'
> (IDP Examiner C, open-ended response)

On a similar note, the British Council examiners reported in the focus group discussion and in their open-ended responses to the questionnaire that interrupting the test-takers was harder during a VCS test. They found the test-takers less sensitive to non-verbal cues, so used more verbal cues to stop test-takers talking and ask questions or develop topics in discussion, or simply made less frequent interruptions than in an in-room test. They found that the strategies discussed during the training, such as hand signals, were not as effective as they thought they would be, and sometimes interruptions were awkward because of delays caused by connectivity. These findings suggest the training would have to address questions on hand movements, gestures and interruptions. An additional implication of this finding is to further examine to what extent slight modifications to the examiners' communication style in the VC mode may, if at all, impact the way test-takers respond.

Headsets

The requirement of headsets during the test was also discussed both in the open-ended questionnaire responses and in the focus groups, and the examiners in the two groups had different opinions. During the pilots, the British Council examiners used headphones and the IDP examiners did not. The British Council examiners said that after a day of testing the headphones felt uncomfortable, and this could potentially be an issue if delivering more than, for example, 11 tests a day. The IDP examiners, however, experienced some audio issues including an echo and suggested that this could be rectified by the use of headphones. After the pilots for the current study, more operational pilots with headphones were conducted, and they were found to be much better in sound quality (refer to Appendix 6).

Alert before each test

Each part of the test was perceived as quite easy to administer (means ranging from 4.22 to 4.56). As for Part 1, both the British Council examiners ($M = 4.60$, $SD = 0.55$) and the IDP examiners ($M = 4.50$, $SD = 0.58$) found it straightforward to deliver the VCS test, although the British Council examiners suggested during the focus group discussion that it would be useful to have some sort of an alert before each session to signal to examiners when a test-taker is ready and waiting. This would allow them to look away from the screen in between sessions and so rest their eyes from the glare of the screen. It would mean that they would be prepared for when the test-taker appears on the screen. As it stands, the examiners felt that they were waiting, sometimes for quite a while, without knowing when a test-taker would appear.

Part 2 prompt card on screen

For Part 2 of the test, the British Council examiners in particular perceived delivering the task, including handling the prompt card, as less straightforward than the other parts (M = 4.00, SD = 0.71 for both questionnaire items). In the focus group, one examiner explained that it felt quite unnatural to put up a prompt card and make the screen of herself available to the test-takers during the one-minute preparation time:

> 'During that one-minute prep time…in a test room situation, they're not looking at an examiner at all, they don't need to. Could we not put the task card full screen? They don't need to see us.' (British Council Examiner 5, Focus Group 2).

This point was repeatedly mentioned during the test-taker focus groups as well.

4.3.3 Rating

The examiners' perceptions about applying the band descriptors to assess candidate performance were overall positive (means ranging from 4.22 to 4.56), but slightly divided between the two groups. As shown in Table 9, the British Council examiners, in general, perceived rating as highly straightforward (M = 4.80, SD = 0.45 for all four aspects of rating) and felt confident about their assigned ratings (M = 4.60, SD = 0.55). On the other hand, the IDP examiners found it relatively less straightforward (means ranging from 4.00 to 4.25) and felt less confident about the accuracy of their ratings (M = 3.75, SD = 0.50).

Table 9: *Results of the examiner feedback questionnaire on rating (N = 9), mean and standard deviation*

1.Strongly disagree – 3.Neutral – 5.Strongly agree	British Council (n = 5)	IDP (n = 4)	Total (N = 9)
I found it straightforward to apply the Fluency and Coherence band descriptors in the video conference test I just administered.	4.80 (0.45)	4.00 (0.00)	4.44 (0.53)
I found it straightforward to apply the Lexical Resource band descriptors in the test.	4.80 (0.45)	4.25 (0.50)	4.56 (0.53)
I found it straightforward to apply the Grammatical Range and Accuracy band descriptors in the test.	4.80 (0.45)	4.25 (0.50)	4.56 (0.53)
I found it straightforward to apply the Pronunciation band descriptors in the test.	4.80 (0.45)	4.00 (0.82)	4.44 (0.73)
I feel confident about the accuracy of my ratings in the test.	4.60 (0.55)	3.75 (0.50)	4.22 (0.67)

In their open-ended questionnaire responses, a majority of the examiners speculated that sound quality may have possibly impacted some of their ratings, as shown in the following examples:

- 'Occasionally, I wasn't sure if it was the microphone or test-taker's English that caused misunderstanding.' (British Council Examiner C, open-ended response)
- 'Disruptions in audio might affect rating especially pronunciation; the audio quality and sound proofing could help this.' (IDP Examiner B, open-ended response)

Similar views were shared in the focus group discussions. All examiners said that generally during the pilots they found rating the VCS tests no different to rating in-room tests. However, there were a few isolated issues during the pilots which the examiners were not sure about, particularly in relation to the quality of the sound. One examiner experienced difficulty when the sound quality deteriorated mid-sentence. Another examiner experienced difficulty because they did not know whether a test-taker was not responding to them because they had not understood the question or because they could not hear the question.

Other examples are related to pronunciation. The examiner provided two examples where she was not sure of the word used by the test-taker and she was not sure whether this was because of the poor audio quality or because the test-taker had used the wrong word. The examiner simply was not able to hear clearly enough. Examiners know that they are rating across the whole test and a slight interference in sound may not impact the overall ratings, but these are issues that they would not face during an in-room test and therefore need strategies to deal with during a VCS test.

4.4 Test-takers' perceptions of the VCS test

The last research question sought to investigate test-taker perceptions about the VCS test mode regarding their overall performance and some of the details specific to the VCS test. Table 10 presents a summary of the quantitative findings from the test-taker feedback questionnaire.

Table 10: *Results of test-taker feedback questionnaire (N = 369)*

Question	British Council (n = 120)	IDP (n = 249)	Total (N = 369)
Did the video conference test you just took allow you to show your full English ability? [1. Not at all, 2. Very little, 3. OK, 4. Quite a lot, 5. Very much]	3.46 (0.84)	3.99 (0.92)	3.82 (0.93)
How clear do you think the quality of the sound in the test was? [1. Not clear at all, 2. Slightly clear, 3. OK, 4. Quite clear, 5. Very clear]	4.31 (0.78)	3.65 (1.07)	3.87 (1.03)
Do you think the quality of the sound in the test affected your performance? [1. No, 2. Very little, 3. Somewhat, 4. Quite a lot, 5. Very much]	1.39 (0.78)	2.64 (1.33)	2.23 (1.32)
In Part 2 (long turn), how clear was seeing the prompt on the screen? [1. Not clear at all, 2. Slightly clear, 3. OK, 4. Quite clear, 5. Very clear]	4.31 (0.88)	4.12 (0.97)	4.18 (0.95)

In the sub-sections below, the quantitative findings from Table 10 will be elaborated with the relevant qualitative findings from the open-ended questionnaire responses and the focus group discussions. The test-taker focus groups were conducted only in the British Council trials.

4.4.1 Perceived test-taker performance

A small majority of the test-takers agreed that the VCS test allowed them to show their full English ability (Total M = 3.82, SD = 0.93; British Council M = 3.46, SD = 0.84; IDP M = 3.99, SD = 0.92). Some test-takers noted that the fact that the examiner was not in the room was less intimidating and for some it was just like talking to friends or family on social media.

The focus group discussions and the open-ended questionnaire responses provided further insights on issues test-takers perceived as impacting their test-taking experience.

Physical distance between test-taker and monitor
Some test-takers mentioned the physical distance between themselves and the monitor was too large, suggesting that this might have made the interaction less natural.

Hand movements
In the British Council test guidelines, the test-takers were asked to keep their hands on the table. This was for security reasons, so that the examiner could see them at all times. The test-takers felt that this was unnatural and made them feel more nervous. Some of them stated that using hand gestures is part of natural conversation and not being able to use their hands made them more nervous.

Control over sound volume

The British Council test-takers welcomed the support given by invigilators before the test with the audio and visual checks. However, during the test, they mentioned that the sound quality would sometimes change (for reasons they were obviously not sure about) and at this point, they questioned whether they would be able to change the volume by themselves. This question was prompted because the guidelines ask them not to touch the headsets at all. It is suggested that the examiners are trained to keep a constant distance from the microphone so that the volume does not fluctuate in the middle of the test.

4.4.2 Sound quality and Its perceived effect on test-taker performance

The test-takers' perceptions about the quality of the sound were generally quite positive, but slightly varied considering a relatively wide range of variance in responses (N = 369, M = 3.87, SD = 1.03; see Table 10). The British Council test-takers, in general, gave relatively higher ratings on the sound quality (n = 120, M = 4.31, SD = 0.78), whereas the IDP test-takers gave slightly lower ratings on average to a varying degree (n = 249, M = 3.65, SD = 1.07). The differing use of headsets in the British Council and IDP pilots may have influenced these differing perceptions.

The perceived effect of sound quality on test performance showed a similar pattern: the test-takers who were tested on the British Council platform were leaning towards the relatively positive end of perception in their ratings (reversed mean = 3.61) and open-ended responses, whereas those who were tested on the IDP platform were towards the relatively negative end in their ratings (reversed mean = 2.36) and open-ended responses (42 comments were made in the questionnaire regarding sound quality affecting performance). These are based only on the test-takers' perceptions, but still suggest some implications for keeping sound quality to an acceptable standard to ensure validity. In the follow-up small-scale pilot with a headset, only one out of 17 test-takers reported that sound quality made a severe impact on their test performance (see Appendix 6).

4.4.3 Prompt card in Part 2

The prompt card shown on the screen in Part 2 seemed to work well for all the pilots (N = 369, M = 4.18, SD = 0.95; see Table 12). Initially for the British Council pilots, the test-takers commented that the script was a little small, but this was rectified for the second week of the study. The test-takers also suggested that the task card should be bigger as during an in-room test their focus is not on the examiner but on the task card. Therefore, for the second week of the study, the British Council increased the font size and enlarged the task card on the screen and for the preparation time, minimised the image of the examiner. After these instant changes between the week-apart pilots, the test-takers perceived the Part 2 prompt card more positively although they still requested to have the card centred, not placed in the left corner of the screen.

4.4.4 Other system-related comments

Additional comments specifically on the operational system of the test platform were made during the British Council test-taker focus groups, some of which were backed up in the open-ended questionnaire responses.

Eye contact

During the focus groups and the open-ended responses, the test-takers mentioned that the lack of eye contact was problematic. Firstly, the examiner was not looking at them, possibly because s/he was dealing with delivery aspects of the test positioned at different parts of the screen and therefore it did not feel like a real conversation. Secondly, they themselves were not sure where to look, whether straight at the screen or the camera.

They felt that it would have been useful had they been told this in the guidelines. It is recommended that British Council build this into the examiner training and also into the test-taker guidelines.

Size of examiner image on screen

The size of the image of the examiner on the screen appeared to be very big. Even though the screen mostly contained the examiner's head, it was not always easy to decipher facial expressions though the quality of the visual was on the whole good. Also, the test-takers felt that because they could not see more of the examiner, they could not use body language to pick up on clues that they might do during an in-person test.

Test-takers able to see themselves on screen

One of the differences between the British Council and IDP pilots was that on the British Council platform, the test-takers were unable to see themselves during the test, whereas on the IDP platform, they could. A popular request during the focus groups and in a few open-ended responses was that the test-takers would prefer to see themselves. This came from a concern that if they moved too much or at all would they move out of the centre circle, which had been used for the visual check prior to the beginning of the test. On popular social interaction platforms such as WhatsApp, Skype, and FaceTime, individuals are able to see themselves, so there is an argument to enable this facility during the VCS test.

Timer

Quite a few of the test-takers said that they would have liked to have a timer during Part 2 of the test, mostly during the preparation time, as they would find that helpful for planning. In the current in-room Speaking test, the test-takers do not have timers for the preparation time or the talking time.

5. Summary of Phase 4 findings

The roll-out of remote delivery and rating of IELTS Speaking has gone through an in-depth four-phase investigation. After undertaking these four phases of the study, ranging from gauging the possibility of video-call technology as an alternative speaking test platform to ensuring comparability with the in-room test, we feel confident that the VCS test would provide wider access for test-takers to be assessed on their speaking abilities, while preserving the crucial interactive nature of communication in IELTS and without presenting serious validity issues. In addition to this broad finding, this investigation has produced a number of specific findings that further strengthen the validity argument. These findings, discussed in Section 4, are summarised in Table 11 for each of the research questions.

Table 11: *Summary of findings*

Research question	Findings
RQ1: Is the existing timing for each part adequate?	• On the whole, Parts 1 and 2 took less than the upper limit of the set time range; Part 3 took two seconds longer than five minutes, the maximum time allotted. • The British Council examiners on average went five seconds over in Part 2, 17 seconds in the Part 2 test-taker talking time, and 16 seconds over in Part 3. • IDP examiners on average were within the set time range. • Both the British Council and the IDP examiners perceived the existing timing as adequate in their questionnaire responses and focus group discussions. • Based on the subsequent ad-hoc analysis by British Council, it was concluded that going over the set time range on average was likely due to some examiners' mistake of ending each part on the test platform and may not be representative of the actual length taken.

RQ2: Do examiners find the minor changes to the interlocutor frame useful?	• In general, the examiners perceived managing and using the revised interlocutor frame as straightforward. • The focus group discussions elicited additional changes which may further improve the test experience: - adding wording prior to the interlocutor frame which is a non-assessed part of the test and gives a brief opportunity to ease any nervousness the test-taker may be feeling; this informal intro to the test is easily achievable prior to the test in the in-room mode, and adding it in the VC mode will bring the test closer to the in-room experience - some flexibility in the script to allow the examiners to stop the test-takers from fiddling with the pen/pencil and paper, which could affect sound quality, and which would meet the general test regulations - adding wording at the end of the test such along the lines of 'You may leave the room now' to explicitly signal the end of the test and provide the formal necessary but polite ending for the test-taker.
RQ3: What are the examiner perceptions about the VCS test mode?	**(i) Test-taker comfort with the VCS test as seen by examiners** • The examiners perceived the test-takers to be comfortable and not intimidated by the VC mode, firstly because the examiner – a potential source of stress – was not physically in the same room and secondly because the test-takers are generally used to communication via technology.
	(ii) Test delivery • The examiners found the overall test delivery quite comfortable. • Some perceived that interactions with the test-takers were limited in the VC mode, which led to less natural interactions and they also had to exaggerate their gestures to make those more noticeable to the test-takers. • The requirement of headsets during the test was seen as being necessary for sound quality but possibly causing some fatigue after wearing for a long time. • Each part of the test was perceived as quite easy to administer with a few suggestions for improvements, such as having some sort of notification, a bell or a sound, in Part 1 to alert the examiners when a test-taker is ready and waiting. • During Part 2 prep time, it was seen as unnecessary to have the examiner visible, since the card prompt was the main focus.
	(iii) Rating • The examiners' perceptions about rating were overall positive, but slightly divided between the two groups. • The British Council examiners, in general, perceived rating as highly straightforward and felt confident about their assigned ratings. • The IDP examiners found it relatively less straightforward and felt less confident about the accuracy of their ratings due to unstable sound quality (before the amendment – see point below). • In their open-ended questionnaire responses, a majority of the examiners speculated that sound quality might have possibly affected some of their ratings. A subsequent small-scale IDP trial focusing on the use of headsets indicated that both the examiners and test-takers perceived the trials with the headsets as more positive.

RQ4: What are the test-taker perceptions about the video call speaking test mode?	**(i) Perceived performance** • A small majority of the test-takers agreed that the VCS test allowed them to show their full English ability. • Some elements made them nervous during the test, such as: - the distance between themselves and the computer monitor - hands required to be on the table all the time (applicable only to the British Council VCS procedure) - not being allowed to change the volume during the test.
	(ii) Sound quality and its effect **• The test-takers' perceptions about the quality of the sound were generally quite positive, but slightly varied considering a relatively wide range of standard deviations in their responses.** • The British Council test-takers, in general, gave relatively higher ratings on the sound quality, whereas the IDP test-takers gave slightly lower ratings on average. • The perceived effect of sound quality on test performance showed a similar pattern: the test-takers who were tested on the British Council platform were leaning towards the relatively positive end of perception in their ratings and open-ended responses, whereas those who were tested on the IDP platform were leaning towards the relatively negative end in their ratings and open-ended responses. As noted above, IDP conducted an additional small-scale pilot with and without headsets, to further explore concerns about sound quality. In their questionnaire responses, both the examiners and test-takers perceived the trials with the headsets as more positive.
	(iii) Prompt card in Part 2 • The prompt card shown on the screen in Part 2 seemed to work well for all the pilots, although the test-takers requested to have the card centred rather than placed in the left corner of the screen on the British Council platform.
	(iv) Other system-related comments • Both the examiners and the test-takers from the British Council focus groups mentioned that eye contact was not natural during the test. The lack of eye contact was seen as problematic, since the examiner was not looking at the test-takers, possibly because s/he was dealing with delivery aspects of the test positioned at different parts of the screen and therefore it did not feel like a real conversation. Secondly, they themselves were not sure where to look, whether straight at the screen or the camera. • The size of the examiner window during the Part 2 preparation time was also perceived as too big as it is the task that should be the focus during the time, not the examiner. • Some of the test-takers mentioned that they would want to see themselves during the test as they do in general video conferencing communications. • Some test-takers wanted a timer during Part 2.

Overall, the IELTS VCS test was perceived positively by both the examiners and the test-takers who participated in the current pilot study. As in the case of integrating new technology into usual practice, however, further improvements in the test platform and associated materials would enhance test experience more and minimise any potential threat to a validity argument to the greatest extent possible. Some suggestions for achieving better quality and enhancing the validity argument are provided below.

The findings of the current study have pointed to six major areas of recommendation:

1. timing
2. further modifications to the interlocutor frame
3. wearing headsets for sound quality
4. minor changes to the test platform
5. additions to guidelines
6. non-verbal communication in the VC mode.

1. Timing

It was found that, on average, the British Council examiners went five seconds over in Part 2, 17 seconds over in the Part 2 test-taker talking time, and 16 seconds over in Part 3. A British Council team revisited the collected timing data through observing videos, and concluded that a potential reason for the British Council timings could be the examiners' mistake of ending each part on the test platform that inflated the averages. This indicates that the timing is adequate and the in-room and VCS test time ranges can be identical. It is, however, still recommended to focus on timing issues through examiner training to make sure that a sufficient amount of test-taker speech can be elicited for each part and that examiners follow the allocated time range.

2. Changes to the interlocutor frame

As suggested by the examiners, some additional modifications to the current interlocutor frame are recommended for the pre-, while-, and post-test phase:

- adding scripted unassessed ice-breaking exchanges prior to the test, which can be easily achievable informally in the in-room test mode, would ensure a smoother VCS test start
- some flexibility in the script is suggested to allow the examiners to stop the test-takers from fiddling with the pen/pencil and paper, which is against the general test regulations and would help with audio quality interference
- add wording to explicitly signal the end of the test and provide a formal but polite ending for the test-taker.

A discussion has already been made among the partners to revise and standardise the interlocutor frame for the VC mode.

3. Headsets

According to the findings of the current pilot and the small-scale follow-up pilot by IDP, it is suggested to require both the examiners and the test-takers to wear headsets during the test to ensure better sound quality and more seamless communication, and therefore achieve a higher degree of validity.

4. Platform changes

It is recommended to implement all or some of the suggestions regarding the test platform (refer to Section 4.3.2 and Section 4.4), through changes to the platform such as:

- having an alert when the test-taker is in the room and ready to start (applicable only to the British Council platform)
- centre the Part 2 card prompt on the screen (applicable only to the British Council platform) for a better test experience
- consider the inclusion of a timer or a signal in Part 2 when the one-minute prep time in Part 2 is about to end: a timer is not used in the current in-room test, but that environment provides more non-verbal cues which can signal the end of the prep time
- consider the test-taker view, so that test-takers can see themselves throughout the test (applicable only to the British Council platform)
- consider the size of the examiner window and prompt card in Part 2, so that the focus is the prompt card (applicable only to the British Council platform).

5. Additions to guidelines

It is recommended to implement all or some of the suggestions regarding guidelines for test-takers and examiners (refer to Section 4.3.2 and Section 4.4).

- The distance between examiners/test-takers and the monitor/camera; this would affect both the availability of non-verbal cues and the sound quality.
- Provide guidance on eye contact and where test-takers and examiner should be looking, i.e., at the screen? At the camera?
- Re-consider the security requirement for test-takers to keep their hands on the table. The test-takers felt that this was unnatural and made them feel more nervous.
- Test-takers mentioned that the sound quality during the test would sometimes change (for reasons they were obviously not sure about) and at this point, they questioned whether they would be able to change the volume by themselves. In the guidelines, they are asked not to touch the headsets at all.

6. Non-verbal communication

It is suggested to further examine to what extent slight modifications to the examiners' communication style, particularly non-verbal, in the VC mode may impact the way test-takers respond and how to further approximate the in-room mode operation-wise. This issue was commented on by the examiners from both the British Council and IDP focus group sessions and by the British Council test-takers. Given the currently available data, we are not sure whether this is just the subjective perception of the examiners and the test-takers, and if it is not, what may cause alterations in video-call communication. It could be a slight but inherent difference between face-to-face and video-call communication that we should include in the extended definition of speaking construct underlying the VCS test, or behaviours that could be remedied simply through operational changes or strategic training. As this was not perceived a critical issue in the past three phases of the study, it might be a peculiar characteristic to the participants or the test platforms of the current study.

It is suggested that this recommendation should not preclude the operational roll-out of the VCS test but should be explored further through live test data and a small-scale study. It would bring the practical benefit of further informing examiner training and test experience from a test-taker point of view.

6. Summary of overall test development and validity argument

6.1 Validity argument built over a four-phase development

In Messick's (1989) seminal chapter on validity, 'validity' is defined as 'an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores' (p. 13). Following Messick's view, validity is seen not as a binary decision, but rather an 'evaluative' spectrum based on a collection of theoretical and empirical underpinnings. Such a perspective has influenced the way language testers conceptualise validity and, crucially, the process of validation, seen as an ongoing process of inquiry that requires accumulation and integration of evidence (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Norris, 2008). Kane's (1992, 2006, 2013) argument-based approach allows a practical implementation of this theoretical perceptive, providing a means to tie a thread of validity evidence into a logical argument.

Validity arguments serve multiple functions, one of which is as an evaluative framework to guide test development processes (Chapelle & Lee, 2021; see also Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010; He & Min, 2017; Pardo-Ballester, 2010; Schimidgall, Getman, & Zu, 2018; So, 2014; Youn, 2015). Validation studies, conducted during test development, not only warrant a chain of inferences with backings in an initial argument but also direct the next moves of development. The IELTS VCS test, developed for wider accessibility in geographically remote or politically unstable areas, was also guided by such an evaluative process in its development.

In particular, the research-informed test development focused particularly on the evaluation inference regarding administration for collecting accurate samples of test-taker output and the explanation inference regarding the comparability of the speaking construct assessed in the in-room and the VC mode. The following sections will provide a brief summary of the findings from each of the previous three phases, as well as the current one, and explain how the cumulative evidence has strengthened the inferential links of validity argument as the development/validation programme progressed.

6.2 Phase 1: Initial evidence to support the evaluation and explanation inferences

In Phase 1 (Nakatsuhara et al., 2016), the possibility of using VC technology in IELTS Speaking was explored in comparison to the standard delivery mode where a test-taker and an examiner are present in the same room. A small set of 32 test-takers and four examiners in London participated in both delivery modes. The participants were observed during, and interviewed after, the tests by the researchers, and they completed a questionnaire about their perceptions of the test delivered in each.

Evidence that supports or refutes the explanation and evaluation inference was sought and is summarised in Table 12. First, in support of the explanation inference, classical test theory (CTT) analysis and many-faceted Rasch measurement (MFRM) analysis showed no statistically significant differences in test-taker scores between the in-room and VC delivery mode. On the other hand, functional analysis of speaking performances found some differences in types of language function elicited from test-takers between the two modes. For instance, the function of asking for clarification was used more in the VC mode than in its counterpart, which might have been attributable to some technological constraints of the delivery platform and potentially weakens the explanation inference. Second, examiners reported some differences in their use of paralinguistic cues such as eye contact and behaviours as interlocutors and raters. In addition, test-takers felt that they were more nervous and had less opportunity to speak in the VC mode, which together were considered counter-evidence to the evaluation inference.

The partially-supported inferential links from Phase 1 indicated that the next phase of development should focus on stabilising technical issues of the delivery platform, devising familiarity training for both examiners and test-takers, and validating assumptions evidenced from a larger sample of test-takers and examiners.

Table 12: *The evaluation and explanation inference examined in Phase 1 and recommendations for Phase 2*

Inference	Assumption	Evidence	Recommendation
Explanation	The comparability of the speaking construct assessed in the in-room and the VC mode	No significant differences in scores found from the CTT and MFRM analyses.	• Replicating the study to confirm with a larger data set.
		Some differences in types of language function elicited from test-takers (e.g., asking for clarifications) used more in the VC mode.	• Providing examiner and test-taker training. • Improving the level of transmission and sound quality of the delivery platform.
Evaluation	Administration for collecting accurate samples of test-taker output	Differences in using paralinguistic cues and behaviours as interlocutors and raters, reported by examiners.	
		Nervousness and less opportunity to speak, perceived by test-takers in the VC mode.	

6.3 Phase 2: Gathering additional support for the evaluation and explanation inferences

Inspired by the findings and suggestions from the previous phase, the study was replicated in Phase 2 with 99 test-takers and 10 examiners in Shanghai, China (Nakatsuhara et al., 2017b). The effect of sound quality on test-taker performance, and new training on test-taker and examiner perceptions and behaviours in the VC mode, was investigated through questionnaires, observations, interviews and focus group discussions.

As summarised in Table 13, the findings for the explanation inference obtained in this phase demonstrated few gaps between the in-room and VC delivery mode in terms of scores awarded and types of language function elicited, and therefore, combined with the findings from Phase 1, largely warranted the explanation inference. When it comes to evidence sought for the evaluation inference, sound quality of the platform in the VC mode was considered adequate by both examiners and test-takers, although higher scores were assigned to lower-level test-takers when sound quality was perceived problematic. The newly added training was also generally perceived useful.

Test-taker training seemed to exert an influence on the level of nervousness and the perceived difficulty of the test delivered in the VC mode; examiner training was considered very effective despite the examiners requesting the need for more opportunities to practice in the new delivery mode.

Table 13: *The evaluation and explanation inference examined in Phase 2 and recommendations for Phase 3*

Inference	Assumption	Evidence	Recommendation
Explanation	The comparability of the speaking construct assessed in the in-room and the VC mode	No significant differences in scores found from the CTT and MFRM analyses.	• Develop independent bespoke platforms to cater for the specific needs of VCS test and enhance technical stability for better sound quality and video transmission.
		Fewer differences in types of language function elicited from test-takers in the VC mode than in Phase 1.	
Evaluation	Administration for collecting accurate samples of test-taker output	Sound quality perceived adequate, in general, by examiners and test-takers.	
		Higher scores assigned to lower-level test-takers when sound quality was perceived problematic.	
		Examiner training perceived very effective despite the requested need for more practice.	
		Test-taker training was generally perceived to positively influence the level of nervousness and the perceived difficulty of the VCS test.	

A collection of evidence obtained in the first two phases supported, to a larger extent, the comparability of the speaking construct assessed between the two delivery modes in view of test scores and linguistic features, and provided a higher degree of confidence in ruling out concerns about discrepancies between the two delivery modes with regards to construct validity. However, technical limitations, although improved compared to the previous phase, still constrained seamless administration for collecting accurate speaking samples of test-takers. To enhance the administrative capability of the technical solution and thus strengthen the evaluation inference further, it was suggested to develop independent bespoke platforms in the next phase.

6.4 Phase 3: Strengthening the evaluation inference further

Phase 3 validated the newly developed bespoke delivery platform in several Central and South American cities by investigating the consistency of scoring procedures under the new platform, as well as its impact on perceived test-taker performance and test administration (Berry et al., 2018). Table 14 summarises evidence gathered from test scores, test-taker and examiner feedback questionnaires, and examiner focus group discussions.

Table 14: *The evaluation and explanation inference examined in Phase 3 and recommendations for Phase 4*

Inference	Assumption	Evidence	Recommendation
Evaluation	The consistency of scoring procedures	Lack of systematic inconsistency in test scores from MFRM analysis.	
	Administration for collecting accurate samples of test-taker output	Sound quality perceived, in general, clear by examiners and test-takers despite some minor technical and/or sound problems.	• Making further efforts to minimise technical problems. • Addressing administration-related issues raised by examiners such as the timing of each part of the test and modifications to the interlocutor frame.
		Lower-proficiency test-takers felt that their performance was slightly more susceptible to sound quality.	
		The test was perceived positively by examiners in terms of both administration and rating.	
		Some concerns raised by examiners about the time required for handling an on-screen prompt for Part 2 (long turn) and potential modifications to the interlocutor frame.	
		The functionality of the bespoke platform was perceived satisfactory by test-takers.	

To highlight key findings, the MFRM analysis found a lack of inconsistency in test scores under the use of the new platform. Questionnaire responses and focus group discussions showed generally positive perceptions from both test-takers and examiners about the overall VCS test, including the sound quality and the functionality of the bespoke platform. Based on these backing data for the evaluation inference, as well as others from the previous phases, the decision was made to begin an operational trial, but with continuing technical platform evaluations and further research inquiries into a few remaining issues in test administration procedures, such as the timing of the test and changes to the interlocutor frame; these became the primary foci of the next validation phase.

6.5 Phase 4: Strengthening the evaluation inference with data from operational conditions

Phase 4 focused on the recommendations from the previous phase through an investigation of the performance of the test in operational conditions, which is summarised in Table 15.

Table 15: *The evaluation inference examined in Phase 4 and recommendations*

Inference	Assumption	Evidence	Recommendation
Evaluation	Timing conditions are adequate	Length of each test part is within the expected limits and examiner perceptions indicated that timing is adequate.	No change to timing.
	Interlocutor frame changes are suitable	Examiner perceptions indicated that changes are straightforward.	Minor changes involving: adding an informal warm-up prior to the start of the test, allowing some flexibility to deviate from the interlocutor frame in cases of bad audio quality, adding an explicit sentence to signal the end of the test.
	Examiners perceptions about aspects of the delivery and scoring the VCS test are positive	• The examiners perceived the test-takers to be comfortable and not intimidated by the VC mode; each part of the test was perceived as easy to administer with a few suggestions for improvements; overall rating was perceived as straightforward and examiners felt confident about their assigned ratings. • Focus group data indicated that a majority of the examiners felt that sound quality might have potentially affected some of their ratings.	Require both examiners and test-takers to wear headsets.
	Test-taker perceptions about the VCS test are positive	• Overall test-takers agreed that the VCS test allowed them to show their full English ability. • Data from the trial and post-trial follow-up indicated that sound quality was perceived to be adequate. • Part 2 prompt card was perceived to work well.	Guidance about eye contact, the size of the examiner window during Part 2 preparation time; test-takers being able to see themselves while speaking; adding a timer.

7. Final remarks

Over the four phases of this project, a total of 595 test-takers and 32 examiners from seven global locations participated in a series of mixed methods studies. Each validation phase informed the subsequent research and development stage, and contributed to updating a validity argument, primarily in terms of the evaluation and explanation inferences, for the test. The earlier validation phases examined issues of test construct in the VC mode and its comparability to the in-room counterpart by using multiple research methods (MFRM analyses, language function analyses, and verbal reports), the results of which contributed to substantiating the explanation inference (Nakatsuhara, et al., 2016, 2017a, 2017b). Some of the findings from the theory-driven construct analyses prompted the necessity for ensuring technical comfort with using the delivery platform and examiner training and test administration tailored to the VC mode. This required an earlier inference of the chain, the evaluation inference, to be revisited, and stronger support for it to be built (Nakatsuhara et al., 2017b; Berry et al., 2018). The latest trial (the current report), which focused on technical analyses and examiner and test-taker perceptions, suggested further fine-tuning of the platform and procedure to minimise any remaining construct-irrelevant variables – more evidence collected for the evaluation inference.

The implementation stage of the VCS test has taken into account these recommendations, and the test became operational in November 2019 in locations in India. Since then, the live tests have been constantly monitored, and an ongoing validation process continues.

As demonstrated in this report, developing and strengthening a validity argument can guide an iterative process of test development and validation. The responsibility of test developers in informing external stakeholders of the research-led decisions made during the process, which is often shared internally only and not reported widely, should be emphasised. Test development framed around an argument-based approach to validation can allow more systematic, targeted activities from the outset and manage a complex, intricate network of test development stages in a coherent way.

References

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly,* 2(1), pp. 1–34.

Bachman, L., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford 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.

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

Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Woods (Ed.), *IELTS Research Reports, Volume 1* (pp. 1–19). Canberra: IELTS Australia.

Brown, A., & Taylor, L. (2006). A worldwide survey of examiners' views and experience of the revised IELTS Speaking test. *Research Notes*, 26, pp. 14–18.

Bui, G., & Huang, Z. (2016). L2 fluency as influenced by content familiarity and planning: Performance, measurement and pedagogy. *Language Teaching Research*, 22(1), pp. 94–114.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. New York: Routledge.

Chapelle, C. A., Chung, Y-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27(4), pp. 443–469.

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

Chun, C. W. (2008). Comments on “evaluation of the usefulness of the Versant for English Test: A response”: The author responds. *Language Assessment Quarterly*, 5(2), pp. 168–172.

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.

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

Elder, C., & Wigglesworth, G. (2006). An investigation of the effectiveness and validity of planning time in part 2 of the IELTS Speaking test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports, Volume 6* (pp. 1–28). Canberra: IELTS Australia & British Council.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, 19(4), pp. 347–368.

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). Luxemburg: European Union.

He, L., & Min, S. (2017). Development and validation of a computer-adaptive EFL test. *Language Assessment Quarterly*, 14(2), pp. 160–176.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, 51(3), pp. 401–436.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, pp. 527–535.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational Measurement* (4th edition) (pp 17–64). Westport: Greenwood Publishing.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), pp. 1–73.

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning*, 25(3), pp. 257–275.

Lam, D. M. K. (2019). Interactional competence with and without extended planning time in a group oral assessment. *Language Assessment Quarterly*, 16(1), pp. 1–20.

Lazaraton, A. (1992). The structural organisation of a language interview: A conversational analytic perspective. *System*, 20, pp. 373–386.

Lazaraton, A. (2002). A qualitative approach to the validation of oral language tests. *Studies in Language Testing,* Volume 14. Cambridge: UCLES/Cambridge University Press.

Li, L., Chen, J., & Sun, L. (2015). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly*, 49(1), pp. 38–66.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd edition) (pp. 13–103). New York: Macmillan.

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), pp. 1–18.

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.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on oral task performance. *Language Testing*, 31(2), pp. 147–175.

Norris, J. M. (2008). *Validity Evaluation in Language Assessment*. New York: Peter Lang.

Ockey, G. J., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). *Exploring the potential of a video-mediated interactive speaking assessment. Research Report No. RR-19-05.* Princeton, NJ: Educational Testing Service.

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing*, 36(4), pp. 505–526.

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 Volume 6* (pp. 91–117). Canberra: IELTS Australia & British Council.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly*, 7(2), pp. 137–159.

Schimidgall, J. E., Getman, E. P., & Zu, J. (2018). Screener tests need validation too: Weighing an argument for test use against practical concerns. *Language Testing*, 35(4), pp. 583–607.

So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly*, 11(3), pp. 283–303.

Taylor, L. (2000). Issues in speaking assessment research. *Research Notes*, 1, 8–9.

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

Weir, C. J., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: An intra-task perspective. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports Volume 6* (pp.119–160). Canberra: IELTS Australia & British Council.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14(1), pp. 85–106.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, 7(1), pp. 1–24.

Xi, X. (2005). Do visual chunks and planning impact the overall quality of oral descriptions of graphs? *Language Testing*, 22(4), pp. 463–508.

Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing*, 27(1), pp. 73–100.

Xu, J. (2015). *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech* (Unpublished doctoral dissertation). Iowa State University, Ames, USA.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, 32(2), pp. 199–225.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and online planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 21(1), pp. 1–27.

Appendix 1: Examiner feedback questionnaire

IELTS Speaking Research Trials: Video Conference Test (Examiner)

Examiner Feedback Questionnaire

Thank you for taking the time to complete the following survey. Obtaining feedback from IELTS examiners is a critical part of ongoing improvement to the test.

It should take between 10 and 15 minutes of your time.

Your responses will be confidential. Responses will not be identified by individual respondents. Your name is not required to complete this survey. All responses will be compiled and analysed as a group. Results of this survey will not be made public.

Background Data

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

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

Your Experience with Technology

3. How often do you use the Internet for each of the following purposes?

	1. Never	2. 1–3 times a month	3. 1–2 times a week	4. 5 times a week	5. Everyday
Socially to get in touch with people					
In your teaching					

4. How often do you use video conferencing (e.g., Skype, WeChat, FaceTime) for each of the following purposes?

	1. Never	2. 1–3 times a month	3. 1–2 times a week	4. 5 times a week	5. Everyday
Socially to get in touch with people					
In your teaching					

Delivering the Test

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

	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
5-1. I found it straightforward to deliver Part 1 (frames) of the video conference test I just administered.					
5-2. I found it straightforward to deliver Part 2 (long turn) of the test.					
5-3. I found it easy to handle task prompts on the screen in Part 2 of the test.					
5-4. I found it straightforward to deliver Part 3 (two-way discussion) of the test.					
5-5. The examiner's interlocutor frame was straightforward to manage and use in the test.					
5-6.Overall I felt comfortable in delivering the test.					

6. If you chose Option 1 or 2 for any questions from 5-1 to 5-6, please explain why. Write the question number(s) and your comment here.	

7. Are there any other positive or negative points that you would like to highlight?

Timing of the Test

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

	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
8-1. The time assigned to Part 1 of the video conference test I just administered was adequate to deliver all the requirements.					
8-2. The time assigned to Part 2 of the test was adequate to deliver all the requirements.					
8-3. The time assigned to Part 3 of the test was adequate to deliver all the requirements.					
9. If you chose Option 1 or 2 for any questions from 8-1 to 8-3, please explain why. Write the question number(s) and your comment here.					

10. Are there any other positive or negative points that you would like to highlight?

Rating the Test

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

	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
11-1. I found it straightforward to apply the Fluency and Coherence band descriptors in the video conference test I just administered.					
11-2. I found it straightforward to apply the Lexical Resource band descriptors in the test.					
11-3. I found it straightforward to apply the Grammatical Range and Accuracy band descriptors in the test.					
11-4. I found it straightforward to apply the Pronunciation band descriptors in the test.					
11-5. I feel confident about the accuracy of my ratings in the test.					
12. If you chose Option 1 or 2 for any questions from 11-1 to 11-5, please explain why. Write the question number(s) and your comment here.					

13. Are there any other positive or negative points that you would like to highlight?

14. Do you see any significant differences between the video conference and in-person test for test-takers?

15. Do you see any significant differences between the video conference and in-person test for examiners?

Thank you for answering these questions.

If you have any questions, please e-mail us at ResearchSurveys@cambridgeenglish.org

Appendix 2: Test-taker feedback questionnaire

IELTS Speaking Research Trials: Video Conference Test (Test-taker)

Test- taker Feedback Questionnaire

Thank you for taking the time to complete the following survey. Obtaining feedback from IELTS test-takers is a critical part of ongoing improvement to the test experience.

It should take between 5 and 10 minutes of your time.

Your responses will be confidential. Responses will not be identified by individual respondents. Your name is not required to complete this survey. All responses will be compiled and analysed as a group. Results of this survey will not be made public.

Your Experience with Technology

1. How often do you use the Internet for each of the following purposes?

	1. Never	2. 1–3 times a month	3. 1–2 times a week	4. 5 times a week	5. Everyday
Socially to get in touch with people					
For your studies					
For your work					

2. How often do you use video conferencing (e.g., Skype, WeChat, FaceTime) for each of the following purposes?

	1. Never	2. 1–3 times a month	3. 1–2 times a week	4. 5 times a week	5. Everyday
Socially to communicate with people					
For your studies					
For your work					

During the Test

3. Did the video conference test you just took allow you to show your full English ability?	1. Not at all	2. Very little	3. OK	4. Quite a lot	5. Very much
4. How clear do you think the quality of the sound in the test was?	1. Not clear at all	2. Slightly clear	3. OK	4. Quite clear	5. Very clear
5. Do you think the quality of the sound in the test affected your performance?	1. No	2. Very little	3. Somewhat	4.Quite a lot	5. Very much
6. In Part 2 (long turn), how clear was seeing the prompt on the screen?	1. Not clear at all	2. Slightly clear	3. OK	4. Quite clear	5. Very clear
7. If you chose Option 1 or 2 for any questions from 3 to 6, please explain why. Write the question number(s) and your comment here.					

8. Are there any other positive or negative points that you would like to highlight?

Thank you for answering these questions.

If you have any questions, please e-mail us at ResearchSurveys@cambridgeenglish.org

Appendix 3: Test-taker feedback questionnaire (English-Chinese bilingual version)

IELTS Speaking Research Trials: Video Conference Test (Test-taker)
雅思口语研究试验：视频口语考试（考生）

Test-taker Feedback Questionnaire
考生反馈问卷

Thank you for taking the time to complete the following survey. Obtaining feedback from IELTS test-takers is a critical part of ongoing improvement to the test experience.
感谢您抽出宝贵时间完成以下调研。雅思考生的反馈对于我们持续改善考试体验是至关重要的。

It should take between 5 and 10 minutes of your time.
本次调研将占用您5到10分钟的时间。

Your responses will be confidential. Responses will not be identified by individual respondents. Your name is not required to complete this survey. All responses will be compiled and analysed as a group. Results of this survey will not be made public.
您的回答会被保密，您也不需要在调研中提供姓名。所有调研参与者的回答不会被单独地分析，而是会被汇编为小组报告进行分析。本次调研的结果不会被公布。

Your Experience with Technology
您在科技方面的经验

1. How often do you use the Internet for each of the following purposes?
在下述活动中，您是否经常使用互联网？

	1. Never 从不	2. 1–3 times a month 每月1–3次	3. 1–2 times a week 每周1–2次	4. 5 times a week 每周5次	5. Everyday 每天
Socially to get in touch with people 与人社交					
For your studies 学习中					
For your work 工作中					

2. How often do you use video conferencing (e.g., Skype, WeChat, FaceTime) for each of the following purposes?
在下述活动中，您是否经常使用视频通话（例如Skype、微信、FaceTime）？

	1. Never 从不	2. 1–3 times a month 每月1–3次	3. 1–2 times a week 每周1–2次	4. 5 times a week 每周5次	5. Everyday 每天
Socially to get in touch with people 与人社交					
For your studies 学习中					
For your work 工作中					

During the Test
考试中

3. Did the video conference test you just took allow you to show your full English ability? 您刚刚参加的视频口语考试是否能够让您全面展示您的英语能力？	1. Not at all 完全不能展示我的英语能力	2. Very little 仅可以很少程度的展示我的英语能力	3. OK 基本可以展示我的英语能力	4. Quite a lot 可以很大程度的展示我的英语能力	5. Very much 可以完全展示我的英语能力
4. How clear do you think the quality of the sound in the test was? 您认为考试过程中的声音质量如何？	1. Not clear at all 完全不清楚	2. Slightly clear 可以听清楚一些	3. OK 基本清楚	4. Quite clear 足够清楚	5. Very clear 非常清楚
5. Do you think the quality of the sound in the test affected your performance? 您是否认为考试中的声音质量影响到了您的考试表现？	1. No 没有影响	2. Very little 有一点影响	3. Somewhat 有一些影响	4.Quite a lot 有较大影响	5. Very much 有很大影响
6. In Part 2 (long turn), how clear was seeing the prompt on the screen? 在考试的第二部分，您可以看清屏幕上的提示吗？	1. Not clear at all 完全看不清	2. Slightly clear 模糊	3. OK 基本清楚	4. Quite clear 足够清楚	5. Very clear 非常清楚
7. If you chose Option 1 or 2 for any questions from 3 to 6, please explain why. Write the question number(s) and your comment here. 如果您在上面的3到6题中选择了1或2，请解释原因。请将题号和您的评论写在这里。					

8. Are there any other positive or negative points that you would like to highlight?
您是否还想强调其他积极或消极的方面？

Thank you for answering these questions.
感谢您回答上述问题。

If you have any questions, please e-mail us at ResearchSurveys@cambridgeenglish.org
如果您有任何问题，请发邮件至ResearchSurveys@cambridgeenglish.org进行咨询。

Appendix 4: Examiner semi-structured focus group protocol

"Welcome, and thank you for participating in this focus group. My name is [name] and I'm your facilitator for this focus group. I'd like to find out who you are, so let's go around the circle and have each person introduce themselves to the rest of the group.

In a minute, I'm going to ask you some questions and I'd like you to answer them. Please share only information with this group you are comfortable sharing. Everything you say is strictly confidential – your real names will not be used at any time during this research project. Please remember that you can leave at any time.

OK, are there any questions or concerns before we begin?"

Turn on recorder

"We will now begin and I will turn on the recorder."

"Again, I would like to extend my appreciation for your participation here today. My first question is …"

Q1-a. How do you perceive test-takers' reactions to the IELTS video conference test you just delivered?

Q1-b. Specifically, how comfortable do you think they felt during the test and why?

Q1-c. Did you notice any difference in test-takers' reaction in the video conference test compared to in an in-person test?

Q2-a. How did YOU find delivering this test in general?

Q2-b. Anything you would do differently?

Q3. How did you find rating the test in general? Why?

Q4-a. How adequate do you think the time assigned to each part is to deliver all the requirements?

Q4-b. Was there any part you think that needed less or more time, and if any, why?

Q5-a. How did you find working with task prompts on the screen in Part 2?

Q5-b. Anything you would do differently?

Q6-a. How natural do you think your nonverbal communication was during the test, particularly eye contact with the test-taker?

Q6-b. Anything you did differently from your usual in-person test administration? Why?

Q7. Are there any elements in the interface that you would want to change for better test administration and/or rating? Why?

"That was my final question. Is there anything else that anyone would like to add or any additional comments concerning what we have talked about here today?"

Allow time for comments

"This concludes our focus group. Thank you for participating. This has been a very successful discussion. Your opinions will be valuable to the study. I hope you have found the discussion interesting. If you have any follow-up questions, please contact [name]. I'd like to remind you that comments will be anonymised and the discussion we have had should be kept confidential."

Appendix 5: Test-taker semi-structured focus group protocol

Introductions

"Welcome, and thank you for participating in this focus group. My name is [name] and I'm your facilitator for this focus group. I'd like to find out who you are, so let's go around the circle and have each person introduce themselves to the rest of the group.

In a minute, I'm going to ask you some questions and I'd like you to answer them. Please share only information with this group you are comfortable sharing. Everything you say is strictly confidential – your real names will not be used at any time during this research project. Please remember that you can leave at any time.

OK, are there any questions or concerns before we begin?"

Turn on recorder

"We will now begin and I will turn on the recorder."

"Again, I would like to extend my thank you for your participation here today. My first question is …"

Q1. How do you feel about the Speaking test you just took?
How comfortable or stressful did you feel during the test, and why?

Q2. How did you find the interaction with the examiner during the test?
What would you change if you could?

Q3. How did you find getting task prompts on the screen in Part 2?
What would you change if you could?

"That was my final question. Is there anything else that anyone would like to add or any additional comments concerning what we have talked about here today?"

Allow time for comments

"This concludes our focus group. Thank you for participating. This has been a very successful discussion. Your opinions will be valuable to the study. I hope you have found the discussion interesting. If you have any follow-up questions, please contact [name]. I'd like to remind you that comments will be anonymised and the discussion we have had should be kept confidential."

Appendix 6: Additional IDP trial: Comparison of test-taker perceptions of using, and not using, a headset

Question 1: Did the video conference test you just took allow you to show your full English ability?

With a headset

Response	TTs	%
Not at all	0	0%
Very little	1	6%
OK	1	6%
Quite much	9	53%
Very much	6	35%
Total	*17*	*100%*
Positive (Quite/very much)	**15**	**88%**

Without a headset

Response	TTs	%
Not at all	0	0
Very little	4	24%
OK	2	12%
Quite much	9	53%
Very much	2	12%
Total	*17*	*100%*
Positive (Quite/very much)	**11**	**65%**

Question 2: How clear do you think the quality of the sound in the test was?

With a headset

Response	TTs	%
Not at all	0	0%
Very little	1	6%
OK	1	6%
Quite much	5	29%
Very much	10	59%
Total	17	100%
Positive (Quite/very much)	**15**	**88%**

Without a headset

Response	TTs	%
Not at all	1	6%
Very little	6	35%
OK	7	41%
Quite much	3	18%
Very much	0	0%
Total	*17*	*100%*
Positive (Quite/very much)	**3**	**18%**

Question 3: Do you think the quality of the sound in the test affected your performance?

With a headset

Response	TTs	%
Very much	0	0%
Quite much	1	6%
Somewhat	4	24%
Very little	4	24%
Not at all	8	47%
Total	*17*	*100%*
Positive (very little/not at all)	**12**	**71%**

Without a headset

Response	TTs	%
Very much	3	18%
Quite much	4	24%
Somewhat	6	35%
Very little	2	12%
Not at all	2	12%
Total	*17*	*100%*
Positive (very little/not at all)	**4**	**24%**

Question 4: If you chose Option 1 or 2 for any questions from 1 to 3, please explain why. Write the question number(s) and your comment here.

With a headset	Without a headset
Question 2:initially it was somewhat disturbing but quite good afterwards. Would recommend with headset	Question2: the sound quality is not up to mark, it is very distracting
3 as I can hear her voice very clearly	2 as the voice wasn't clear. I had to make them repeat the questions
Sound quality is good	Voice is not audible
There is disturbance in headphones which effects the quality	3, there is disturbance in voice which is not much audible
	sound not clear
	Sometimes the sound was not clear and there was humming sound in between the question

Question 5: Are there any other positive or negative points that you would like to highlight about the test you just took?

With a headset	Without a headset
Error in the survey: No comments saved	The experience of using speakers can be better
	Distrubence in sound
	I was not getting the clear voice
	Voice quality is poor in video conference
	Quite good in building confidence
	Voice clearity. Echo
	Couldnt hear it properly .It affects our performances
	Sound disperse
	Sou d quality needs to be improve
	Yes, because of the virtual presence of the examiner lil bit prob in voice clearity
	Indeed, the quality of the sound was not clear in between otherwise it was good overall

Question 6: Would you prefer to take the video conference test with a headset or without it?

Response	TTs	%
With a headset	*13*	76%
Without a headset	*2*	12%
I have no preference	*2*	12%
Total	*17*	*100%*
With a headset	**13**	**76%**

Appendix 7: Additional British Council data analysis: Difference between manual and automated timing of Part 3

Date of test	Candiate L1	Examiner ID	Interview ID	Manual timing part 3	Time difference	Length of Part 3
07/05/2019	Chinese	201902	829457813	05:45	00:11	5:56
07/05/2019	Chinese	201901	343654276	04:58	00:12	5:10
07/05/2019	Chinese	201902	849661008	05:09	00:13	5:22
07/05/2019	Chinese	201907	244786334	05:28	00:09	5:37
07/05/2019	Chinese	201909	395550703	05:24	00:06	5:30
07/05/2019	Chinese	201908	551480545	05:06	00:10	5:16
07/05/2019	Chinese	201907	532839269	05:30	00:11	5:41
07/05/2019	Chinese	201908	745531617	05:52	00:12	6:04
07/05/2019	Chinese	201908	848584589	05:09	00:11	5:20
07/05/2019	Chinese	201908	984941510	05:29	00:11	5:40
08/05/2019	Chinese	201902	148696546	05:32	03:37	09:09
08/05/2019	Chinese	201909	812269396	04:53	00:22	05:15
08/05/2019	Chinese	201902	205771590	05:06	00:16	05:22
08/05/2019	Chinese	201902	227606489	05:11	00:12	05:23
08/05/2019	Chinese	201902	599530618	05:12	00:16	05:28
13/05/2019	Chinese	201906	434800198	06:12	00:09	06:21
13/05/2019	Chinese	201904	625135921	05:18	00:12	05:30
13/05/2019	Chinese	201906	998143375	05:22	00:20	05:42
13/05/2019	Chinese	201905	360227285	05:37	00:08	05:45
13/05/2019	Chinese	201905	256399712	05:18	00:07	05:25
13/05/2019	Chinese	201906	156307321	05:16	00:07	05:23
13/05/2019	Chinese	201905	835073363	05:48	00:06	05:54
13/05/2019	Chinese	201905	543470696	05:05	00:09	05:14
13/05/2019	Chinese	201906	704610370	05:09	00:11	05:20
13/05/2019	Chinese	201905	929719286	05:13	00:09	05:22
13/05/2019	Chinese	201906	475037282	04:33	00:08	04:41
13/05/2019	Chinese	201905	349123605	05:05	00:16	05:21
13/05/2019	Chinese	201905	607976840	05:05	00:09	05:14
13/05/2019	Chinese	201906	486908007	05:29	00:08	05:37
14/05/2019	Chinese	201911	853958999	05:36	00:06	05:42
14/05/2019	Chinese	201918	791202453	05:07	00:10	05:17
14/05/2019	Chinese	201917	689707076	05:17	00:07	05:24
14/05/2019	Chinese	201918	790599535	05:12	00:07	05:19
14/05/2019	Chinese	201924	179972213	05:42	00:08	05:50