ISSN 2201-2982
2021/2

IELTS Research Reports Online Series

Towards new avenues for the IELTS Speaking Test: Insights from examiners' voices

Chihiro Inoue, Nahal Khabbazbashi, Daniel Lam and Fumiyo Nakatsuhara

IELTS™
BRITISH COUNCIL idp Cambridge Assessment English

Towards new avenues for the IELTS Speaking Test: Insights from examiners' voices

This study investigated examiners' views on all aspects of the IELTS Speaking Test – the test tasks, topics, format, interlocutor frame, examiner guidelines, test administration, rating, training and standardisation, and test use. The report provides suggestions for potential changes in the Speaking Test to enhance its validity and accessibility in today's ever-globalising world.

Funding

This research was funded by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Grant awarded 2017.

Publishing details

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

How to cite this report

Inoue, C., Khabbazbashi, N., Lam, D., and Nakatsuhara, F. (2021.)
Towards new avenues for the IELTS Speaking Test: Insights from examiners' voices,
IELTS Research Reports Online Series, No. 2. British Council, Cambridge Assessment English and IDP: IELTS Australia.
Available at https://www.ielts.org/teaching-and-research/research-reports

Introduction

This study by Inoue, Khabbazbashi, Lam and Nakatsuhara was conducted with support from the IELTS Partners (British Council, IDP: IELTS Australia and Cambridge Assessment English), as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge Assessment English, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with over 120 empirical studies receiving grant funding. After undergoing a process of peer review and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing series* (www.cambridgeenglish.org/silt), and in the *IELTS Research Reports*. Since 2012, to facilitate timely access, individual research reports have been made available on the IELTS website immediately after completing the peer review and revision process.

The research study featured in this report explores examiner perspectives on the speaking section of the IELTS test. Because the assessment landscape continually evolves, periodically revisiting key perspectives on the test – and making any necessary adjustments based on these views – is important for IELTS to keep pace with contemporary developments. Understanding stakeholder perspectives has formed an important part of the *IELTS Research Report Series* since its inception, and the input of examiners is as relevant as the views of students, teachers and recognising institutions. The global pool of IELTS examiners operates on a huge international scale; involving a highly trained, qualified and experienced cohort – all of whom have solid pedagogical expertise.

The large examiner group involved in this study covered a range of years of experience, locations and training history. Using a mixed method approach, examiner views on all aspects of IELTS Speaking were investigated – including topics, examiner guidelines, task types and training. Large-scale questionnaire data was initially gathered, the results of which were used to conduct follow-up interviews, probing examiner responses in further depth.

So what were the findings of this study? Examiners viewed IELTS Speaking in a positive light overall, which was encouraging to note. As expected, there were aspects signalled as requiring development for optimal contemporary use. Examples included suggestions for a broader range of topics, more flexibility in the use of the interlocutor frame, adjustments in the rating criteria or potential additions to examiner guidelines. The report contains an extensive discussion of these in detail, and recommendations in response are outlined.

In addition to these findings, the fact that research in this mould (conducted by independent academics and peer reviewed) is shared in the public domain indicates the importance IELTS places on transparency. Commissioning research to critique aspects of major tests – and publishing the results – is the only legitimate approach to ensure that assessment standards are sufficiently scrutinised for high-stakes use. For stakeholders such as recognising institutions, this type of transparency should be central in guiding decision-making about which test to use for their context and purposes. It demonstrates the important role that academic research must play in providing support for test users, and the continued need to reflect on best practice for evidence-based decision making in the contemporary assessment domain.

Tony Clark
(with acknowledgement to the
British Council Research Team for their involvement)
Senior Research Manager
Cambridge Assessment English

Towards new avenues for the IELTS Speaking Test: Insights from examiners' voices

Abstract

This study investigated the examiners' views on all aspects of the IELTS Speaking Test, namely, the test tasks, topics, format, interlocutor frame, examiner guidelines, test administration, rating, training and standardisation, and test use.

The overall trends of the examiners' views of these aspects of the test were captured by a large-scale online questionnaire, to which a total of 1203 examiners responded. Based on the questionnaire responses, 36 examiners were carefully selected for subsequent interviews to explore the reasons behind their views in depth. The 36 examiners were representative of a number of different geographical regions, and a range of views and experiences in examining and giving examiner training.

While the questionnaire responses exhibited generally positive views from examiners on the current IELTS Speaking Test, the interview responses uncovered various issues that the examiners experienced and suggested potentially beneficial modifications. Many of the issues (e.g. potentially unsuitable topics, rigidity of interlocutor frames) were attributable to the huge candidature of the IELTS Speaking Test, which has vastly expanded since the test's last revision in 2001, perhaps beyond the initial expectations of the IELTS Partners.

This study synthesised the voices from examiners and insights from relevant literature, and incorporated guidelines checks we submitted to the IELTS Partners. This report concludes with a number of suggestions for potential changes in the current IELTS Speaking Test, so as to enhance its validity and accessibility in today's ever-globalising world.

Authors' biodata

Chihiro Inoue

Dr Chihiro Inoue is Senior Lecturer at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, UK. She specialises in the assessment of L2 speaking and listening, particularly in the task design, test-taker processes and features of learner language. She has carried out numerous test development and validation projects around the world, including IELTS, Cambridge English Qualifications, ISE, TOEFL iBT, Oxford Test of English, GEPT in Taiwan, GTEC and EIKEN in Japan, and the National English Adaptive Test in Uruguay. Her publications include a book (2013, Peter Lang), book chapters in *The Routledge Handbook of Second Language Acquisition and Language Testing* (2021) and journal articles in *Language Assessment Quarterly* (2021; 2017), *Language Learning Journal* (2016) and *Assessing Writing* (2015).

Nahal Khabbazbashi

Dr Nahal Khabbazbashi is Senior Lecturer in Language Assessment at CRELLA, University of Bedfordshire. Her research interests include the assessment of speaking, the effects of task and test-taker related variables on performance, the use and impact of technology on assessment, and new constructs in the digital age. Nahal has led the research strands of a number of high-profile test development and validation projects in different educational and language testing contexts from school settings in Uruguay to higher education institutes in Egypt. Her work appears in journals such as *Language Testing* (2020; 2017), *Linguistics and Education* (2019), and *Language and Education* (2019).

Daniel Lam

Dr Daniel Lam is Lecturer in Language Learning and Assessment at CRELLA, University of Bedfordshire. His research interests include assessing interactional competence, the role of feedback in learning-oriented assessment, and use of language test scores in university admissions. He has worked on various funded research projects related to Cambridge English Qualifications, IELTS, and TOEFL iBT. Daniel's work appears in journals such as *Applied Linguistics* (2021), *Language Assessment Quarterly* (2019) and *Language Testing* (2018; 2020).

Fumiyo Nakatsuhara

Dr Fumiyo Nakatsuhara is Reader in Language Assessment at CRELLA, University of Bedfordshire, UK. Her main research interests lie in the nature of co-constructed interaction in various speaking test formats, the impact of test-taker characteristics on test performance, task design, rating scale development, and the relationship between listening and speaking skills. She has carried out a number of international testing projects, working with ministries, universities and examination boards. For example, she led a series of research projects for the use of video-conferencing technology to deliver the IELTS Speaking Test (2014–2018). Fumiyo's publications include the books, *The discourse of the IELTS Speaking Test* (with P. Seedhouse, 2018, CUP) and *The co-construction of conversation in group oral tests* (2013, Peter Lang). Her work also appears in journals such as *Language Testing, Language Assessment Quarterly, Modern Language Journal* and *System.*

Table of contents

1 Rationale

Like its predecessor ELTS, the IELTS test was developed to be 'a non-static instrument' which would continue to gather research evidence and information to engage in 'the dynamic process of continuing test development' (Hughes, Porter and Weir, 1998, p. 4). Since its inception in 1989, prominent events related to this philosophy include the start of the IELTS joint-funded research program in 1995, which generated over 90 external studies involving more than 130 researchers all over the world[1] and the IELTS Revision Projects (Reading, Listening and Writing in 1993–1995 and Speaking in 1998–2001). The Speaking Test Revision Project (1998–2001), which resulted in a number of significant changes, was informed by studies funded by the IELTS joint-funded research program (see Taylor and Falvey, 2007) and research within/in collaboration with Cambridge (e.g. Lazaraton, 2002) as well as cutting-edge research in language testing and applied linguistics at that time (for a summary of the ELTS and IELTS development, see Nakatsuhara, 2018).

At the inception of this current study in 2016, fifteen years had passed since the last major revision in 2001. Given the number of IELTS Speaking studies carried out since 2001, together with innovations in digital technologies which advanced the field of speaking assessment more rapidly than ever, it was considered timely to undertake a study that looked into possibilities for future revisions.

In the series of test development and revision projects conducted for ELTS and IELTS in the last half a century, we have learnt what sources of information, in addition to empirical investigations into speaking test designs and different aspects of validity (e.g. Brown, 2007; Brown and Hill, 2007; Lazaraton, 2002), could be useful to inform test revisions. For instance, Davies (2008, p. 90) critically evaluates the ELTS Revision Project (1986–89) in which stakeholder questionnaires and interviews, despite an expenditure of £200,000, did not result in any major advances in understanding innovative test systems for the original IELTS test. Davies is particularly critical about the input from the applied linguists, arguing that "the applied linguists' responses were varied, contradictory, and inconclusive, and provided no evidence for a construct for EAP tests on which we could base the test" (ibid.).

One critical point to reflect for possible reasons for this unfortunate outcome is that the targeted stakeholders were too varied, and the way in which data was collected was not sufficiently focused. On the contrary, a questionnaire survey focused only on examiners carried out prior to the 2001 Speaking test revision (Merrylees and McDowell, 2007) was found to make a valuable contribution to the IELTS Speaking Revision Project. The survey by Merrylees and McDowell was conducted in 1997, gathering IELTS examiners' views on various aspects of the test. The results informed the development of the analytic rating scales, interlocutor frame and training and standardisation procedures for IELTS examiners (Taylor, 2007).

In more recent years, examiners' voices have been regarded as one of the most important sources to inform speaking test validity (e.g. Ducasse and Brown, 2009; Galaczi, Lim and Khabbazbashi, 2012; May, 2011; Nakatsuhara, Inoue, Berry and Galaczi, 2017a). All of these studies used structured data collection methods, such as stimulated verbal recalls while showing video-recorded test sessions, focused surveys with both selected-response and open-ended questions, and structured focus group discussions. The researchers of this study have recently been involved in four IELTS Speaking projects funded by the IELTS Partners (Nakatsuhara, Inoue and Taylor, 2017; Nakatsuhara, Inoue, Berry and Galaczi, 2016; 2017a, 2017b; Berry, Nakatsuhara, Inoue and Galaczi, 2018) which elicited IELTS examiners and examiner trainers' views on the test administration procedures and rating processes as a part of these mixed-methods studies. The research teams have always been impressed by the usefulness of their voices in specifying the construct(s) measured in the test and in suggesting potential

1. https://www.ielts.org/teaching-and-research/grants-and-awards

ways forward to improve the IELTS Speaking Test. Based on our recent experience with gathering examiners and examiner trainers' voices in a systematic, focused manner, it is our firm belief that their views will be highly valuable in guiding the directions of possible IELTS Speaking Test revisions in the future.

2 Research question

This research addresses the following question.

> What are the IELTS examiners' and examiner trainers' views towards the IELTS Speaking Test and their suggestions for future improvement?

3 Research design

The design of this study involved two main phases: 1) conducting an online questionnaire for a wider participation from examiners around the world; and 2) follow-up semi-structured interviews (via video-conferencing or telephone) with selected examiners who are representative of different regions and examining experiences.

3.1 Phase 1: Online questionnaire

3.1.1 Questionnaire

In order to construct the online questionnaire, firstly, three experienced IELTS Speaking examiners were invited to participate in a focus group session where they discussed various aspects of IELTS with the researchers in May 2017. The aspects discussed included the test tasks, topics, format, Interlocutor Frame (i.e. examiner scripts), the Instructions to Examiners (i.e. examiner handbook), administration, rating, examiner training and standardisation, and the construct and use of IELTS Speaking Test.

After the focus group discussion, the researchers put together a draft questionnaire and sent it to the three examiners for their comments, based on which the questionnaire was revised. The revised version was then sent to the British Council to be reviewed by the Head of Assessment Research Group, the IELTS Professional Support Network Manager and the Head of IELTS British Council.

The final version of the questionnaire (see Appendix 1) was put online using SurveyMonkey (https://www.surveymonkey.com/) in November 2017, and emails with the link were sent out to the Regional Management Team in the British Council[2] to distribute to test centres administrators who then forwarded it to examiners. The questionnaire was open until the end of January 2018.

3.1.2 Participants

Through the online questionnaire, a total of 1203 responses were collected. The respondents, on average, had taught English (Q1) for 18.9 years (SD=10.04, N=1198), and have been an IELTS Speaking Examiner (Q2) for 7.49 years (SD=5.68, N=1152). Of the 1203 respondents, 404 (33.64%) identified themselves as an IELTS Speaking Examiner Trainer[3].

2. Although this project focused on the examiners managed by the British Council, we believe that the results and implications discussed in this report apply to the examiners managed by IDP Australia (another IELTS Partner), as both pools of examiners follow exactly the same training and standardisation procedures.

3. However, this number (n = 404) may not be entirely accurate, as we found some respondents who were not actually examiner trainers during the interviewee selection stage. Eighty-one respondents identified themselves as an examiner trainer on Q47 where they selected their current main role concerning examiner training and standardisation, and this (n = 81) is the number that we used to stratify data for analysis and discussion in Section 4.8.

In terms of the regions where they were working as an examiner or examiner trainer, 1179 respondents answered: 35% were based in Europe; 16% in the Middle East and North Africa; 14% in East Asia; and 13% in Northern America. A smaller percentage of examiners were in South Asia (8%), Southeast Asia (6%), Africa (3%), Latin America (3%), Australia and New Zealand (1%), and Russia and Central Asia (1%).

3.1.3. Data analysis

Responses to the closed questions on the online questionnaire were analysed using descriptive statistics in order to capture the general views of the IELTS Speaking examiners towards the test. The comments on the open questions were used for selecting participants for the follow-up semi-structured interviews in the second phase of the study. Some of the written comments were also quoted wherever appropriate, to interpret quantitative results or to support the interview data.

3.2 Phase 2: Semi-structured interviews

The second phase of the study involved semi-structured interviews with a small sample of examiners (N=36) in order to gain a more in-depth understanding of the views expressed in the online questionnaire. There was a call for volunteers at the end of the questionnaire inviting examiners to share their contact details should they wish to participate in a follow-up interview.

3.2.1 Participants

From a total of 1203 respondents of the online questionnaire, approximately one-third (n=418) provided their contact details. We first used a stratified sampling approach with (a) region and (b) examining experience as the main strata for an initial interviewee selection. We subsequently reviewed the individual questionnaire responses of these examiners (both closed and open-ended) to select participants with diverse opinions. This was to ensure that the interview data was representative of examiners' voices. Fifty (50) examiners were included at this stage from which the final 36 were selected based on availability. There were 30 examiners and six (6) examiner trainers (Examiner ID: E01–E36; for the six examiner trainers, their IDs start with ET, e.g. ET08).

Participants in the second phase had a range of examining experience (M=7.12; SD=6.76) from new examiners (with less than six months of experience) to highly experienced examiners (with 23 years of experience). The countries in which these examiners reported being active included Australia, Canada, Chile, China, Germany, Greece, Hong Kong, Indonesia, Ireland, Japan, Nigeria, Pakistan, Qatar, Romania, Russia, Singapore, South Africa, Spain, Sweden, Thailand, the United Arab Emirates, the United Kingdom, and the United States.

Two of the researchers conducted interviews. Both were familiar with the IELTS Speaking Test and previous research on it, and one was also a former IELTS Speaking examiner.

3.2.2 Interviews

The interviews generally followed a similar structure by focusing on the main themes covered in the questionnaire and specifically, those areas where the quantitative results pointed to a need for a more in-depth examination. Interview questions, nevertheless, were tailored to the individual examiners drawing on their specific responses to the survey.

To illustrate, the results of the survey showed that more than half of the respondents disagreed with or felt neutral about the statement 'the topics are appropriate for candidates of different cultural backgrounds'. Following up on this trend, we formulated different interview questions for interviewees who had expressed contrary views on the questionnaire:

- Q: *You had selected 'disagree' in your responses regarding appropriateness of topics in particular in terms of culture/gender. Can you give us a few examples of some of these topics? In what ways do you think these were inappropriate?*
- Q: *Overall, you were happy with the topics and their appropriateness in terms of culture/gender. What do you think makes for a good topic? In your experience have there been any instances of culturally sensitive or unfamiliar topics?*

It was not possible to address all survey areas in each interview; however, we ensured that all areas of interest were covered across the interviews through a judicious selection of themes and questions tailored to each individual examiner. Samples of interview questions can be found in Appendix 2.

Data collection took place over a period of four months from March 2018 to June 2018. We sent out an invitation email (Appendix 3) to the selected examiners and asked them to confirm their interest by return email, after which we scheduled individual interviews via video or audio-conferencing according to participant preferences. Upon completion of all interviews, participants were sent an Amazon gift voucher for the value of £25 as a token of appreciation for their time.

The first three interviews were jointly conducted by both researchers in order to establish a common approach for guiding the interviews. The remaining interviews were independently conducted by the two researchers. Interviews were designed to last between 30–40 minutes, although this varied from individual to individual. A degree of flexibility was built into the scheduling to allow participants sufficient time to express their views and at their own pace. All interviews were audio-recorded with the consent of participants. Researchers took detailed notes as they were conducting the interviews and also wrote a summary report for each interview. This was done to facilitate the transcription of interviews at a later stage by identifying the most relevant parts to transcribe.

3.2.3 Data analysis

All audio recordings were carefully examined. Researchers' detailed notes were helpful when listening to the audio files in identifying the most relevant parts for transcription. A thematic analysis of the transcriptions was subsequently carried out by the two researchers. Given that the interviews were structured around the survey, coding the responses was generally straightforward, with themes closely aligning with the survey categories and subcategories. Since the dataset was relatively small and the coding was mostly straightforward, it was not necessary to use qualitative analysis software such as NVivo; instead, all coded data were simply tabulated for easy reading and comparison. For those instances where a response fit into multiple codes or did not fit in neatly with specific survey code(s), the researchers, in joint discussion, either created a new code or included it under an 'other' category for further analyses.

4. Results and discussion

This section presents findings on the nine different aspects of the IELTS Speaking Test, following the structure of the questionnaire: test tasks, topics, format, Interlocutor Frame (i.e. examiner scripts), the Instructions to Examiners (i.e. examiner handbook), administration, rating, examiner training and standardisation, and the construct and use of the test. Each aspect first describes the general trends of the responses on the closed questions in the questionnaire (see Appendix 1 for the descriptive statistics for each closed question), and then links the relevant themes found in the interviews to the questionnaire data, so as to identify the issues in the current IELTS Speaking Test and discuss potential ways forward.

4.1 Tasks

The first section of the questionnaire asked about the tasks in the IELTS Speaking Test. Responses to Q1–Q6 showed that the majority of examiners:

- found the *language samples elicited* in each part of the test either *often useful* or *always useful*: Part 1 (60.0% [Q1]), Part 2 (87.2% [Q3]) and Part 3 (92.1% [Q5]).
- felt that the *lengths of each part* are *appropriate*: Part 1 (80.6% [Q2]), Part 2 (82.6% [Q4]) and Part 3 (72.1% [Q6]).

Although the responses were generally positive towards the tasks in the current IELTS Speaking Test, the percentages or degrees of agreement from the examiners varied among the different parts and aspects of the test. The results from the follow-up interviews, which aimed at unearthing the reasons and issues behind such variations in examiners' views, are presented below.

4.1.1 Part 1

The questionnaire results above showed that 40% of the examiners did not find the language samples elicited in Part 1 to be often or always useful (Q1). This is a considerably higher percentage compared to those for Parts 2 and 3. Also, approximately 20% of respondents did not find the length of Part 1 appropriate (Q2). When asked to comment on the above findings as well as general observations regarding Part 1 in the follow-up interviews, examiners touched on a variety of issues with this part of the test.

Length of Part 1 and the requirement to cover all three topic frames

Some examiners commented that it was not always possible to go through all frames and that they had to make the decision to leave out questions in the rubrics. While the *Instructions to Examiners* booklet states that all questions within each of the three frames should be covered and asked one-by-one in the order in which they appear in the frame, examiners are allowed to skip questions if four specific circumstances apply – and one of them is when they run out of time and need to move on. However, in practice, some examiners appear to have difficulties deciding whether and when to skip questions.

Some examiners also commented that 'you can only go so far' (E01) with the questions. Part 1 was also seen as somewhat limited in assessing different aspects of speaking: 'Part 1 is typically short responses sometimes less than a sentence so we can't, for example, access cohesion or coherence' (E13).

Appropriateness and relevance of questions

Examiners provided examples of instances where questions/rubrics were irrelevant, inappropriate, or had already been covered by the candidate: 'sometimes they have answered the questions already so it's a bit ludicrous to ask the questions again' (E14). As a result, the requirement to have to go through the rubrics regardless of its appropriateness was negatively perceived.

Particular concerns were raised in relation to the first frame where questions were found to be too prescriptive and not necessarily applicable to all candidates. Some examiners pointed out that the wording of the question asking candidates whether they are studying or working requires categorical responses, whereas some candidates are neither studying or in work (e.g. the period of time right after school, or those who have finished studies and applying for postgraduate degrees). Another examiner believed this question to be insensitive to spouses or stay-at-home mothers, some of whom were not even allowed to have jobs due to spouse visa restrictions. In the words of our examiners 'people in transition' (E05), 'young adults of 15–17' (E16) and 'home makers' (E09) are not necessarily covered in these questions as currently framed.

This issue actually seems to relate to the need for enhanced examiner training and clarification of guidelines. The above two issues are actually addressed in the *Instructions to Examiners*. Examiners are allowed to skip a question if the candidate already answered it, and in the first Part 1 frame only, they can change the verb tense of questions as appropriate, i.e. past tense to ask about previous work/study experience. Therefore, these issues indicate the problems where examiners (or the trainers who monitor their performance) may have interpreted the guidelines too rigidly.

Other examiners pointed to how questions such as where are you from, where is your hometown, or tell us about your culture may not necessarily be appropriate in an era where there is increased mobility or in contexts where most people are immigrants:

> *A lot of students do not necessarily live in one place so their 'hometown' can be confusing/cause problems.* (E15)

> *We are largely a country of immigrants; these questions are becoming more and more unpopular.* (E09)

> *First part of Part 1 needs to be seriously revised. The first couple of questions should not create ambiguity and they definitely do that. You are asking someone who has lived in the US for 10 years about their cultures! They are Americans…they are hyphenated Americans! The first few questions should be totally clear and relax the candidate and not have their brains tell them 'what the hell does that mean' – it has to be rethought.* (E12)

> *Give us starting off questions or areas to touch on rather than tangential questions which 9 times out of 10 will not be in the natural flow of conversation.* (E03)

The *Instructions to Examiners*, however, does offer some flexibility where parts of the frames referring to one's country may be changed to 'in [name of town]' or 'where you live' as appropriate, so that the question would elicit speech on familiar, immediate surroundings. These examiner voices indicate that this may also be one of the areas that need to be emphasised in the training.

Memorisation of responses

There were some comments on the questions in Part 1 being too familiar or general and lending themselves easily to memorisation or 'learning by heart' (ET08), thus giving a false impression of fluency.

Differential performance across test parts

Linked to the theme of memorisation is candidates' differential performance across test parts. As one examiner commented 'someone's fluency for questions can be very different in Part 1 than in other parts...Part 1 can be easily prepared' (E18). Another examiner commented that 'quite often candidates whether their language is weaker will produce the same response' (E07). These can explain why 40% of examiners do not find this part useful in eliciting language samples.

Perceived inflexibility in Part 1

One of the recurrent criticisms of Part 1 was the strict examiner frame which, according to some examiners, does not offer any flexibility in making even minor amendments to the wording of questions and rubrics and/or to skip questions where deemed appropriate. We will return to this theme in Section 4.4 on the interlocutor frame.

> *Part 1 is always a bit artificial and I'm not allowed to ask my own questions.* (ET08)

> *Worse thing about Part 1 is the obligation to ask all the questions under each of those questions.* (E06)

Although Part 1 in the IELTS Speaking Test, by design, offers less flexibility than Part 3, the *Instructions to Examiners* specify that the verb tense of questions can be changed as appropriate in the study/work frame (but not in other frames). There are other instances where questions can be omitted, for example, due to sensitivity of the topics or for time management purposes. The examiners' interview responses may therefore be indicative of a tendency of over-imposition of the examiner guidelines in certain testing contexts, an aspect perhaps worth emphasising in examiner training. However, there remains the issue of examiners having no choice but to move on after posing a question that has turned out as sensitive; more often than not it is about being able to flexibly shift after questions have been posed and candidates have signalled a problem.

4.1.2 Part 2

Part 2 was viewed favourably by 87.2% of the examiners, in terms of eliciting useful samples of language for rating (Q2) in the questionnaire. However, the follow-up interviews revealed two issues with this part regarding the response time and memorised responses.

Response time

The prescribed response time of two minutes was considered the main problem in Part 2. Two minutes may sometimes be too long for both proficient, native-speaker-like candidates and weaker candidates alike.

> *It depends on the candidate…whether too weak or too strong: if they are weak, two minutes can be too painful. In extreme situations, it can get too long and you just have to sit there for two minutes.* (E07)

> *It's difficult even for native speakers to talk for two minutes on a subject that they have not had much time to prepare for.* (E27)

> *In Part 2, depending on the fluency of candidate, you can get a huge amount of language in 2 min, and the candidate can dry up before that, and think 'I've blown it'. They look at the examiner to seek help, but the examiner has been trained to leave a long gap, which would be unnatural in conversation, in the hope that they come up with something. I think 1.5 minutes is adequate.* (E26)

Some examiners suggested the need for clearer wording in the instructions to candidates to set up the right expectations:

> *Examiners instructed to tell candidates to talk for 1–2 minutes, which could give a misleading impression that the candidates have the option, whereas they do need to continue until the two minutes is up.* (E27)

> *The instruction says talk for 1–2 min. So the candidates thought they have reached the 1 min mark and 'I'm good' – and the examiner asks them to keep on talking. It's the expectation. In the training, examiners were told that candidates need to speak for the full two minutes. So I use gesture to ask them to continue. For lower level candidates, there are problems, not because of nature of questions but their ability.* (E32)

In fact, the instruction to candidates is worded as '1 to 2 minutes' so as to mitigate the anxiety that candidates – especially weaker ones – may feel by being told to speak for two minutes. However, the discrepancy between the stated duration in the instructions to candidates and what examiners are trained/instructed to do (cf. E32's comment above) might be worth addressing.

One examiner trainer also linked the length issue to the topic and the nature of the task, and suggested an alternative for Part 2.

> *Part 2 is rather artificial – when in life would you be required to give a two-minute monologue on a personal experience? Moving this to a presentation, if done well, could make this closer to a real-life situation.* (ET21)

Nevertheless, ET21 also recognised that changing Part 2 task into a presentation would still have unauthentic element because 'in any real-life situation, you wouldn't have to give an extremely important, 'life-changing', presentation when you're given only one minute to prepare.

Memorised responses

Examiners from certain regions discussed how Part 2 (together with Part 1) is prone to elicit memorised, pre-fabricated responses. This is particularly evident in cases where candidates give a response that seems to touch on several potential topics but not entirely relevant to the assigned topic. Several examiners (e.g. E19, ET21, E22, E24, E27, E30) discussed the issue of candidates giving memorised responses to questions and shared insights on how to deal with the issue.

> *As the examiner, I would take them out of the topic, ask them in Part 3 something I know they cannot give memorised answers to. I would like to see that a bit more spontaneity in the IELTS test, where examiners are allowed to deviate a bit sometimes, to get a bit more out of the candidates, or get them back onto the topic, take them out of their memorised lines.* (E30)

In contrast, another examiner supported the current Part 3, which is designed to elicit more spontaneous use of language.

> *If people have rehearsed and prepared topics, you'll know about it, as Part 3 will reveal it, so you can revise or adjust the rating as you go along.* (E24)

At times, combating memorised responses might mean that examiners have to interrupt the candidate in the middle of their (memorised) responses, in order to ask follow-up questions and elicit more spontaneous responses.

> *Students around [Bands] 5 and 5.5 deliver very good memorised answers to Qs, using quite archaic language. The answer goes on for such a long time that it is difficult to interrupt them, but you often have to interrupt them, interrupt their memorised responses – it's the only way to get them to give a natural response.* (E19)

As illustrated in other sections of this report, a recurring theme (suggestion) among examiners' comments has been to introduce more flexibility to the interlocutor frame, as a way of creating more natural, smooth-flowing interaction and reducing redundancy, easing up the cognitive load for examiners, among other purposes. The examiners' insights here points to another important cause – to combat candidates' memorised responses.

Other suggestions for combating memorised responses by one of the examiners (E22) include the following, with a common theme around devising means to make available more frames or test versions for examiners to choose from at any one time:

- go back and recirculate older frames/booklets; over time, item writers have produced numerous frames, but only a fraction are in circulation at a time
- booklets can go out of circulation more quickly (1–2 months)
- randomise questions using tablet devices.

I'd suggest more topics – partly as giving more choice to examiners, partly for preventing rehearsed answers. Currently, a test booklet is used for eight months, then it retires. Generally, in China, the new test booklet is available pretty much publicly, not quite accurately, after a week it has come out. If there're more questions or topics, it makes it harder to prepare a rehearsed answer. Instead of getting a booklet to retire, why not add a new booklet to the number? (ET21)

4.1.3 Part 3

Questionnaire results showed the highest agreement rate of 92.1% for the usefulness of language samples elicited in Part 3 compared to the other test parts (87.2% for Part 2 and 60.9% for Part 1). The length of this part, on the other hand, was found to be *a bit too short* or *too short* by over 20% of questionnaire respondents. The examiners were asked to elaborate on these findings in the interviews, from which the following themes emerged.

Overcoming memorisation

As mentioned in Section 4.1.2, Part 3 was viewed positively in allowing examiners to 'push them [candidates] out of their comfort zone' (ET08) while noting that this part did not lend itself to learning by heart and rehearsing.

Cognitive load and time pressures

Examiners, particularly new ones, viewed the requirement to paraphrase and think about spontaneous questions in Part 3 as challenging and a struggle within the given time restrictions, particularly in taking focus away from candidate speech. Illustrative comments are provided below.

Challenges of paraphrasing, inventing new questions, and worrying about timing distracts from listening to their language. (E10)

I'm used to the Cambridge exams [where you have a separate rater present during the test] and I'm getting used to Part 3 and it's difficult to manage everything by yourself. We have to think more, we have to be careful in how we formulate those questions and we are not supposed to exactly use the words in brackets and I don't see why we cannot use those specific words. (E14)

Opportunities to demonstrate higher proficiency

Some examiners expressed concerns over the limited opportunities in the speaking tasks for candidates to display features of higher speaking proficiency characteristic of Bands 8 and 9 on the rating scale.

Topic of the questions was listed as a contributing factor. One examiner, E35, in particular, views the task questions as often incompatible with the rating criteria for higher bands.

The test requirements and the questions do not match. The test does not have face validity in expecting candidates to use complex language and idiomatic speech when talking about fruit or a happy childhood experience.

A candidate could give a perfect answer but wouldn't get a high rating due to the linguistic criteria, for example, using unreal condition[al] or other complex structures.

You'll get a candidate who is [Band] 6 or 6.5, but nothing so far [Parts 1 and 2] will push them to use higher level language; and if you follow the task 3 questions, you're not likely to get that.

Part 3 follow-up questions are only for clarification. They are not really argumentative. There's no time for them [candidates] to express their opinions – the questions are unable to push them.' (all quotes from E35)

ET21 presents a similar view:

In Germany, there are some Band 8 and 9 candidates in every session – and the questions are too basic for these candidates. (ET21)

On the other hand, some examiners reported that, at times, the topics are too abstract or unfamiliar to the candidates, such that they have little to say about them and are therefore unable to display their proficiency.

The final part of Part 3 questions – sometimes the questions can be extremely cognitively demanding. For example, talking about the design of buildings, it feels quite specific, and you need a pretty keen interest [on the topic] to answer the question. Sometimes you have a very good candidate, but who just doesn't know a lot in that area, so it's difficult to keep on topic. It's frustrating because we are trying to push the candidate to express as much as possible. (E32)

Moreover, the length of Part 3 was another issue examiners related to limited opportunities to display higher proficiency levels.

For [Band] 8 or 9 students, they need the space to demonstrate that level of ability. More time is needed in Part 3 to really distinguish between 8 and 9. (E28)

You sometimes run out of time to get candidates to really use higher level language, for example, vocabulary, in the final frame. Some less experienced candidates would go on to talk about themselves, when they are expected to talk about the topic in general. If we're given 2–3 minutes more, there can be more language sample, with the possibility of displaying higher level language. (E20)

Examiners also referred to the nature of the tasks as a factor – they elicit a relatively narrow range of language functions.

It's rather one-sided, difficult to do with a one-on-one examiner-candidate setup; but there's a little real interaction with the examiner being asked Qs or having any operational discussion between the examiner and the candidate. There's a little attempt to do that in Part 3, but still basically questions from examiner, answer from candidate. It's not a real discussion. You try and make it feel like one, but it's still not. (ET21)

4.1.4 Range of task types

While a strong majority of questionnaire respondents (91.4% [Q7]) believed the *number* of test tasks to be appropriate, the percentage was lower (71.9% [Q9]) for the range of tasks. More than one in four examiners felt that the range of tasks was *a bit too narrow* or *narrow* (27.5%) on Q9, and nearly half of these examiners (i.e. 13.3% against the whole 1203 respondents) wanted to include a free discussion task (Q9a). In the subsequent free comment box, there were a number of examiners who wished to have increased flexibility in what they can say and ask, rather than suggest different tasks to be included. These findings, again, seem to point to the need for increased flexibility in the interlocutor frames.

The follow-up interviews found examiner preferences for task types that: (a) allowed for *finer distinctions to be made* between candidates; (b) did not lend themselves to *memorisation* due to their familiar nature – a recurrent theme touched on in the previous section; and (c) aligned more closely with *academic settings*. Another examiner viewed the different tasks as 'repetitive' (E07) and expressed a preference for more variation.

This is in line with Seedhouse's (2018) observation that despite the test designers' original intention to measure three distinct types of interactional patterns in the three parts of the test, essentially only two types of interaction are elicited: 'topic-scripted Q-A adjacency pairs' in Parts 1 and 3 and monologic speech in Part 2. Contrary to the intended design for Part 3, more flexible turn-taking opportunities were not observed in Seedhouse (2018).

4.1.5 Sequencing of tasks

In the questionnaire, most examiners *agreed* or *strongly agreed* that the *sequencing of tasks* was appropriate (84.7% [Q8]). In the follow-up interviews, examiners viewed the sequencing of tasks as appropriate and in particular, the thematic linking of Parts 2 and 3 of the test was positively commented on in creating *a sense of purpose* and providing a degree of *authenticity* by allowing for 'follow-up on a conversation' (E09). Moving on to a different topic was considered to be 'too distracting' (E02) and potentially making candidates 'feel lost' (E02).

However, it should also be noted that in the interviews, there were examiners who were not always positive about the thematic link between Parts 2 and 3 because even if a topic does not work well for a candidate in Part 2, that has to continue in Part 3. This is further discussed as one of the topic-related issues in Section 4.2.5.

4.2 Topics

The second section of the questionnaire was on the topics in the test. It was found that the more than 60% of examiners **agreed** or *strongly agreed* with the following statements.

- Overall, the topics in the test tasks are appropriate. (61.7% [Q11])
- The topics are appropriate for candidates of either gender. (67.7% [Q12])
- The range of topics (task versions) which examiners can choose from in the Booklet is adequate. (74.5% [Q14]).

However, as the percentages above indicate, up to 40% of questionnaire respondents for each question were less positive towards the topics given in the IELTS Speaking Test. For Q11, almost 40% of the respondents *disagreed* or felt *neutral* about the overall appropriateness of the topics. Moreover, when this statement was further narrowed down, results showed that topic appropriateness was particularly problematic in terms of candidates' background and gender. Specifically, over half of the respondents *disagreed* or felt *neutral* about the statement, 'The topics are appropriate for candidates of different cultural backgrounds' (Q13) and over one-third for topic appropriateness for either gender (Q12). These negative responses echo the findings of Brown and Taylor's (2006) survey with 269 IELTS Speaking examiners, which reported examiners' concerns about topic choices either in terms of inappropriateness for candidates from different age groups, cultural backgrounds, rural areas, or with different levels of world experience.

The results of our survey also highlighted the topics as one of the main areas to explore in more depth in the interviews. Here we summarise the main emerging themes from the interviews. Examiners touched on several problematic features of topics, which are presented in the following section.

4.2.1 Issues raised about topics

Topic and cultural background

The inappropriateness of topics, particularly in Part 1 of the test, was frequently brought up by some examiners who used adjectives such as 'frightening' (E02), 'mundane' (E10), 'silly' (E16), 'outdated' (E09), or 'trivial' (E12) to describe topics/frames. However, the incongruity of a given topic within a specific cultural context was a recurrent theme with some examiners highlighting this point with illustrative examples from their experiences.

> *I find it [topic of Boats] interesting but if you live in central China they think it is hilarious and they are baffled by it.* (E15)

> *Bicycles can be problematic for example in Saudi Arabia…full of sand and blistering hot and people become more and more indignant and it would be good to be able to ask them something else.* (E15)

> *Some topics are just so English! Greece is a bus country and not a train county. Taxis are considered public transport. These kinds of things.* (E36)

> *Sometimes I worry about the culture-specificity of the topics.* (E18)

Examiners, most notably those working in the Middle East and the Gulf countries, listed topics such as music or pop stars as topics not necessarily culturally appropriate and acting as 'stumbling blocks' (E01) for candidates.

The position of the IELTS Partners as communicated to the research team is that examiners are able to choose appropriate frames and topics. However, the *Instructions to Examiners* booklet explicitly requires examiners to vary the topics from candidate to candidate, with no guidelines stating whether or not examiners can intentionally decide not to use certain topics that they find unsuitable for a particular group of candidates. It may therefore be necessary to add a caveat to the *Instructions to Examiners* that individual examiners have discretion to avoid certain topics should they identify any inappropriateness for a particular cohort of candidates, though this should not be overly used for test security reasons.

Affective nature of topics

Examiners observed the potential for some topics to be too emotional for candidates and even causing breakdowns, which may in turn affect their performance.

> *I've had people break down when you ask them to recollect the past; we don't need any Marcel Proust prompts to go back to your childhood and think about Madelaines!* (E12)

> *Some topics in Part 2 can disarm them enough to get them frazzled.* (E11)

> *A student can perform badly because of the topic they are talking about; family members…for example. It's a recipe for disaster. Some can handle it and some can't. You see the tears well up but you are not supposed to intervene. So the examiner is in a very difficult situation.* (E12)

The inability to 'intervene', as pointed out by E12, ties in with the theme of inflexibility which was touched on earlier in restricting examiners to take appropriate actions when the test does not proceed smoothly and as intended. However, the *Instructions to Examiners* do state that if candidates break down and become emotionally distressed, examiners can stop the test and give them a few moments to recollect themselves. This might be another area for the attention of both examiners and examiner trainers. However, there remains another issue that, even if examiners know (or learn) that they can pause the test in such a situation, it may be difficult to make that decision because pausing a test means the need to move on to the next part upon resuming, and therefore, the loss of a valuable two-minute sample of candidates' language from Part 2, which could become a scoring validity concern.

The role of socio-economic background

Issues of class and socio-economic status were raised by several examiners who referred to some topics as too 'middle class' (E02, E09, E36) or outside the experience of candidates from lower socio-economic backgrounds. See below for illustrative examples.

> *For example, a car journey… a lady from a rural area…she probably hasn't stepped foot in a car so she would find it very challenging to tell a story.* (E02)

> *Most of the topics are urban-centric and upper-class oriented.* (E34 open comment from questionnaire)

> *Not everyone is from an English middle-class background.* (E36)

> *We don't need to talk about pieces of arts in museums. We should have them balanced off with other kinds of questions, e.g. transactional things that might be more common or useful.* (E09)

> *Topic of boats…maybe it's appropriate at the Cote d'Azure or the Riviera. I mean even in Qatar when the boat show is on, out of the Qatari population, maybe only 20 are at the boat show. I can't think of anyone/anywhere talking about love of boats as a teenager.* (E03)

The reference to 'teenagers' (E03) in the above quote can be linked to another theme from the interviews which is the extent to which topics are age-appropriate and/or within the realm of experience of candidates as discussed below.

Age-appropriateness of topics

Some examiners pointed to their experiences of examining younger candidates – high school leavers or those aged between 17–18 year – as a 'new generation of candidates' (E03, ET25) who may display high levels of language ability and fluency, yet lack the world knowledge and experience necessary for addressing some topics.

> *The new generation of candidates are very fluent, Band 8, but when you give them the topic of consumerism, or international relations they struggle. They may have the language, but don't know how to answer or add things to the response. They lack the general knowledge, probably dealing with these issues only at university. So they may be disadvantaged. They can do well with technology, youth etc. but not food production, checking quality of food, or transporting food from one place to another – probably something they've never thought of, or not really relevant to them.* (ET25)

> *Topics about business you get for very young candidates (17–18 year olds) or about married life; international relations which is way out of their depth. I worked in China for a long time and I worked in Uzbekistan and they are sometimes limited talking about topics outside their experience.* (E16)

Lack of interest/familiarity with topics

Some examiners also commented on the problem of assigning topics that candidates have little interest in or are not familiar with. They raised concerns about how this might disadvantage candidates and expressed a preference for having the flexibility to switch topics or provide support in such cases.

> *As an examiner, I'm bound by the rubric. It's happened when someone said they have no interest in sports. But I've started the rubric, so I had to continue. But I do worry it is a bit of a disadvantage for example when a candidate is a movie fan but has no interest in sport.* (E18)

> *Some candidates cannot relate to certain topics…and they should be able to ask for suggestions from us.* (E02)

Given the range of topics, it seems perverse to ask them about the one thing they don't like to talk about. (E15)

If the candidate is not familiar with the topic, there is no option to change the topic. This puts the candidate at a serious disadvantage. Some candidates, especially intermediate and undergraduate students, are extremely good at grammar and pronunciation but are not exposed to certain cultures or are not aware of certain lifestyles. They lack world/topic knowledge. (E34)

Giving examiners the option to change topics, however, is controversial. According to the guidelines for materials development (shown to the research team by the IELTS Partners), topics and frames are designed and trialled so that they do not disadvantage candidates without much background knowledge (see more details in the section 'Topic equivalence in Part 2' below).There are also test security and fairness reasons. The *Instructions to Examiners* state that examiners must not change the topics unless a legitimate request is made by the candidate. This is to minimise malpractice and ensure that candidates do not choose to avoid topics that they feel less prepared for, which is a crucial consideration for a standardised test like IELTS Speaking.

Gender and topics

The appropriateness of topics in terms of gender was perhaps one of the most controversial aspects of the interviews with some examiners who challenged the way some questions were phrased as reinforcing 'gender stereotypes' (E09, E11) with illustrative comments below.

Sometimes Part 3 questions almost seem like getting the candidates to be sexist – women should stay at home and men should be at work. (E13)

Part 3 questions have a sexist framing. (E26)

So, for example, 'Do you think boys are better at sports than girls or better at maths than girls?' It implies that one is better than the other…so it's harder for someone with low confidence or weaker skills to assert themselves to say why should one be better than the other or I don't like the way this question is phrased. (E09)

Note that in the third quote, the examiner hints at a possible interaction between language proficiency, confidence levels, and the ability to challenge such questions. Another examiner pointed out that asking questions like 'who is more supportive in a workplace: males or females?' may put pressure on the candidates to find 'the right answer' (E11) that does not offend the examiner.

There were, however, other examiners who did not view gendered topics as problematic pointing to the cultural backgrounds in which they were examining where gender roles are 'quite traditional' (E07).

In my examining experience in Russia and Kazakhstan there haven't been many problematic topics. They are quite traditional in terms of gender roles. (E07)

An examiner from South America acknowledged an interaction between gender and performance and went on to describe her own procedure for ensuring gender-appropriate topic selection:

As I'm going through the booklet in the beginning, I have a procedure and I take different colour tape and look at the list, and this one is a boy and he is better suited for questions about music, electronics, media, whereas I'll choose families, keeping yourself fit, etc. for girls. I see, especially in [examiner's country] the gender identity is a traditional Latino culture and gender identity is much more defined. Girls know more about certain topics than boys. (E10)

Examiners from the Middle East referred to topics such as handbags and jewellery as not necessarily working well with their male candidature observing that they are 'not often confident to elaborate' (E01) on these topics.

Topic equivalence in Part 2

Furthermore, concerns were raised regarding the extent to which the topics are equivalent in terms of difficulty, particularly in Part 2.

> *Talking about a healthy lifestyle is easier than describing a wild animal. You need more specific vocabulary.* (ET08)

Different topics require different sets of vocabulary, and it is arguable whether the necessary vocabulary to describe a healthy lifestyle is more or less 'specific' than those for describing a wild animal. However, examiners' concerns about topic equivalence seemed legitimate.

Upon gathering these voices from examiners regarding the neutrality and appropriateness of topics, the research team made a request to see the guidelines for materials development from the IELTS Partners, and learned about the development processes and guidelines for the IELTS Speaking test materials. According to these guidelines, the topics used in the test are designed to be accessible and of interest to a wide candidature, and not assume any particular background knowledge or socio-economic background. There is also a list of unsuitable topics in order to avoid imposing biases and conflicts, such as religion, politics, distressing topics (e.g. death, divorce) and stereotypes (sexism, racism, cultural clichés, etc.). Draft topics and frames go through a number of iterations, internal and external reviews, as well as trials at various proficiency levels and types of educational institutions, before selection for the operational tests. The accessibility of different topics and the equivalence of test versions are ensured through trials and multiple levels of expert judgement.

Although all the topics and frames are developed based on these rigorous guidelines and go through the reviews and trials, examiners have still found some topics problematic. This is partly because trials are, by nature, done on a smaller scale than in the live tests, so there are inevitably issues that only emerge when tests go live with the wider IELTS candidature. A potential addition to the cycle of materials development that could safeguard against this is obtaining feedback from examiners regarding the suitability of the live topics and frames. As communicated by the IELTS Partners at the time of writing this report, there is a form available that centres can complete as and when they see fit, although currently there is no compulsory system in place for completing and sending feedback on topics and frames on a regular basis. Examiners' awareness could be raised about this feedback form, so that potentially unsuitable topics, should there be any, are flagged early and either removed from the pool or revised.

4.2.2 Impact of topic-related problems on performance

Linked to the above theme is the perceived impact of topic-related issues on candidate performance, and examiners commented on how unfamiliar topics can lead to response that are too short, or may not always give candidates 'a chance to talk' (E02). Examiners' comments suggested that they viewed the main role of topics as generating sufficient samples of speech and believed that problematic topics ran the risk of eliciting speech that was not necessarily representative of candidate abilities.

> *The responses for unfamiliar topics can be too short and that is the main problem.* (ET25)

> *Problematic topics can raise affect and decrease language production.* (E09)

A portion of their interview is less representative of their ability; unless they are really flexible and get tangential. (E01)

The reference to 'unless they are really flexible' by E01 can be taken to hint at certain candidates' adeptness at using communication strategies – an arguably construct-relevant factor – or other candidate attributes such as 'confidence' (E09) or 'outspokenness' (E02) perceived to be important by examiners in 'navigating the situation' (E11). One examiner thought that candidates' ability to handle 'bizarre or astonishing' (E15) topics could be a good measure of their performance. Nevertheless, the extent to which some of these skills or attributes are construct-relevant or irrelevant is open to debate.

Overall, these examiner comments seem to be in line with recent IELTS Speaking research by Khabbazbashi (2017), who reported on the effect of different levels of candidates' background knowledge on their speaking performance. The score differences were statistically significant (i.e. not obtained by chance), but they were not large enough to move candidates across adjacent bands. Although Khabbazbashi's findings indicate the presence of some topic-related bias in the test, as examiners in this study suspected, it does not actually lead to changes in the test results.

4.2.3 Examiner strategies for dealing with 'problematic' topics

Related to the above issue, some examiners mentioned specific strategies for dealing with such problematic instances; for example, using body language or gestures to communicate that they are not 'pressing' (E01) candidates for a response or brushing off the topic and quickly moving on subsequent topics – particularly in Part 1 of the test. Others used body language to invite candidates to continue talking. Topic-related problems were found to be less pronounced in Part 3 where there is a degree of flexibility for examiners to formulate their own questions.

Examiners also reported adopting a selection strategy (also mentioned in Section 4.2.1) by, for example, assigning topics they deem appropriate to males/females while others select those they deem to be appropriate to everyone from the list of available frames. It is worth emphasising that these are not standard operating procedures for IELTS, but those adopted by examiners as a measure to circumvent topic-related issues.

In terms of possible solutions, examiners once again expressed a preference for more flexibility in topic selection, particularly drawing on their experiences of what works well or not within their context. As suggested earlier in Section 4.2.1, this again seems to highlight the necessity to add a caveat to the *Instructions to Examiners* that individual examiners have discretion to avoid certain topics should they identify any inappropriateness for a particular cohort of candidates, though this should not be overly used for test security reasons..

4.2.4 Content or language?

Related to the theme of topics is the difficulty experienced by some examiners in separating language from content or ideas – features not explicitly measured in the IELTS Speaking Test. Others problematised the fact that some questions 'seem to be too much about ideas' (E36) which may advantage or disadvantage those with better ideas. Illustrative comments are presented below.

I know language is used to express yourself to the world, but is it our concern whether people spend time reading or watching the national geographic? (E02)

Some questions seem to be too much about ideas, but this is not an exam about ideas and it is difficult to separate the two. (E36)

I know that what is being tested in IELTS is language, but should we not care whether the content of candidate speech makes sense or not? (E02)

Although some descriptors regarding the content and relevance are present in the IELTS Speaking rating scale, they only appear in Bands 8 and 9 in Fluency and Coherence. Contrary to some of the expressed opinions above, recent research suggests the content of speech to be an important criterion closely attended to by language experts in academic domains and linguistic laypersons in general communicative settings alike (Eckes, 2009; Sato, 2014). The explicit addition of a content-related criterion – which currently is not included in the test – might therefore be an area of consideration by the IELTS Partners.

4.2.5 Topic connection between Parts 2 and 3

In the online questionnaire, the topic connection between Part 2 and Part 3 was perceived to be a positive feature by 79.5% of the respondents (Q15). However, when asked about potentially given the choice to change topics in Part 3 (Q16), the responses were more mixed; with 40.8% of the respondents disagreed, 22.7% stood neutral, and 36.5% agreed. This was further investigated in the follow-up interviews, where examiners exhibited different views on whether Part 3 should stay on the same topic as Part 2 or should change to a different topic.

Preference for topic change

The use of the same topic in Parts 2 and 3 is so as to extend the topic area to a more abstract level in Part 3, but some examiners expressed preference for Part 2 and Part 3 to be on different topics, citing fairness to candidates as a main reason – 'fairness' in the sense of how much a candidate knows about the topic and can say something about it:

[If there is topic connection between Parts 2 and 3,] then so much of the test depends on whether you're lucky with the topic. The candidate may be put off and feel really unlucky: they have barely managed to put something together for something they have little or nothing to say about in Part 2, and then in Part 3 they have to do the same again. (E27)

Part 1 goes pretty quickly, as the topics are pretty well-prepared for. And then if you just pick one topic for two thirds of the test [Parts 2 and 3], it just doesn't make too much sense for me. For weaker candidates – they just don't understand some of the words, then Part 2 is a nightmare, and then 'well, let's continue [that in Part 3]… Very occasionally, with young or inexperienced candidates – they don't have anything to talk about, for example, an 11-year-old in China, or a candidate in North Korea. Their life experience is really quite limited. In Part 2, the candidate just doesn't know what to talk about. So being able to switch topics would be a positive thing. (ET21)

ET21's comments on the candidates who are young or with limited life experience echo the discussion presented earlier in Section 4.2.1. Although the speaking item writing guidelines require that test materials should be accessible to different ages and cultural backgrounds, it is unrealistic to expect the topics, especially when expanded to a more abstract level in Part 3, to be easily accessible to candidates at such a young age and without much life experience. This issue is also related to the vastly expanded use of the test over the last two decades, which will be discussed later in Sections 4.9 and 5.4.

Not changing topics between Parts 2 and 3 could work against some candidates, but it could also give some candidates an unfair advantage:

If the candidate is very familiar with the topic or is in that profession, for example, the candidate is a scientist and was asked questions about science and they feel very lucky that all questions [in Parts 2 and 3] are on their topic of interest. (E27)

These examiners are therefore in support of the principle of giving candidates 'fresh starts' in different parts of the test. Indeed, ET21 has such a strong preference that, in explaining why he 'strongly disagreed' that examiners should be given the option to switch topics between Parts 2 and 3 (Q16 in the survey), he argued that changing topics should be made compulsory, citing another reason in terms of the lexical range assessed of candidates.

> *Part 1 covers 3 topics rather superficially, then Parts 2 and 3 go deeply into one general area of lexis. To gain a wider range of lexis, separate Task 2/ [Task] 3 topics would help. It should not be option – it should be forced.* (ET21)

Preference for topic connection

Other examiners cited reasons for keeping the topic connection between Parts 2 and 3, including counter-arguments to the issue of unfamiliar topics, and practical (and cognitive) disadvantages of starting a fresh new topic in Part 3.

ET28 argued that the Part 3 topics are 'abstract enough' for candidates to 'have an opinion' even if they haven't had the relevant experience; and E33 viewed the logical progression of speaking about the same topic from a personal to a more general level a positive feature of the test.

> *I don't see a problem with the current practice. There might be potential difficulties for people who don't have TV if the topic was your favorite TV show, but the Part 3 related topics could be abstract enough to have an opinion on even if you don't have much experience with it.* (ET28)

> *I 'strongly disagreed' [option to change topics in the questionnaire] because it's logical to progress from talking about a topic at a personal level to a more general level. I would like the entire topic in Part 2 to Part 3 to be coherent. I like some sort of a logic as the speaking test progresses.* (E33)

In a similar vein, E24 viewed the topic connection as providing opportunities for candidates to develop an understanding of an otherwise unfamiliar topic and be 'warmed up' enough to talk about it.

> *I think if we try to bring them back to same topic, there're a lot of opportunities, with the different difficulty levels [in Part 3 frames]. Often, they are able to talk about the topic once they've understood the topic...The majority of candidates can speak a bit on the topic. They may have had some exposure to it [in Part 2], and can therefore now go on. They may be thinking about it, they have done their one to two-minute long turn, so it's easier for them to continue.* (E24)

Conversely, if the candidate has to start afresh and think about ideas on a different topic, this, in E19's view, would place more cognitive demand on the candidate. The extra thinking time the candidate needs may in turn cause delays to the test:

> *The candidates would need more thinking time – which makes the exam longer. e.g. movie to education system in your country, a big shift. Need time to process and get the language together. Disadvantage to a lot of candidates.* (E19)

4.2.6 Positive views on topics

Thus far, Section 4.2 has presented and discussed various issues related to topics, as viewed by the examiners. At this point, it is worth reiterating that more than 60% of the examiner respondents to the online questionnaire felt that the topics in the IELTS Speaking are *appropriate* (61.7% [Q11]) and the range of topics to select from is *adequate* (74.5% [Q14]. Nearly 80% of the examiners found the topic connection between Parts 2 and 3 to be a positive feature (79.5% [Q15]).

It is important to highlight that, in the interviews, not all examiners viewed topics to be problematic; take, for example, the following comment:

> *Actually I haven't come across any problematic topics and it has worked fine. Topics work very well.* (E14)

Examiners were also sympathetic to the challenges and difficulties of selecting topics for the international candidature of IELTS and viewed the 'diversity of generic topics that we can all weigh on' (E01) as a positive aspect of the test. Others also found that the selection of topics displayed IELTS designers' concern for 'affect' (E09) and showed sensitivity to the candidature by avoiding 'controversial, political or obvious panic button kinds of questions' (E09), which closely correspond to the guidelines for materials development presented in Section 4.2.1.

A number of issues raised in the interviews and discussed above are likely to have stemmed from the discrepancy between the intended candidature and test use at the time of test design in 2001 and the hugely expanded candidature and varied test use in the last two decades. This issue of test use is discussed in more detail in Sections 4.9 and 5.4.

4.3 Format

In the questionnaire, the vast majority of examiners felt positively about the current format of the IELTS Speaking Test:

- 87.6% of the examiners *agreed* or *strongly agreed* that the one-to-one interview format should be kept in the IELTS Speaking Test (Q18)
- 95.0% of the examiners *agreed* or *strongly agreed* that the face-to-face examiner-candidate interaction mode used in the current test is a suitable delivery for the test, as compared to a computer-delivered mode. (Q19)

Given the increasing popularity of online/computer-delivered tests, we decided to explore this theme in more depth in the interviews and present the main emerging themes below. It should be noted that while the questionnaire specifically asked about the delivery format of the test, our interviewees brought up other issues related to technology and assessment (e.g. automated assessment) and these were at times conflated in the same discussion.

Authenticity

Examiners were careful to acknowledge that there is an artificial element to any kind of assessment; nevertheless, they believed that a face-to-face test is more authentic, closer to the target language use domain and 'at least…adds a more natural element' (E01) while computer-delivered testing was viewed as 'one more step removed from what language is about' (E01).

> *It is unnatural so authenticity is a big problem. Rarely ever do we talk to computers. Because it's easier to manage for the testing board it's popular. The face-to-face interview is not the perfect way, but a good way to accurately gauge their spoken proficiency.* (E06)

> *A test that tests human interaction is a marker of what we need to do in the real world. IELTS sets them up in a much better way.* (E09)

> *Generally speaking, I base my views on my students and they way prefer face-to-face because talking to a computer is not a particularly natural thing even for modern kids who talk to parents with computers.* (E12)

One examiner considered remote testing as a viable option and 'a second best thing' but asserted that 'we lose a lot' by opting for 'an anonymous, and particularly without authentic interaction, computer voice'. (E09)

It is believed that these positive comments towards (remote) face-to-face tests over computer-delivered tests described here and in the next two sub-sections are of particular interest to the IELTS Partners, given a series of recent research into the use of video-conferencing to deliver the IELTS Speaking Test (Berry et al., 2018; Nakatsuhara et al., 2016; 2017a; 2017b).

Construct of speaking

Linked to the above theme is examiners' voiced concerns about a narrowing of the speaking construct with the removal of interactive features of speaking and elements of natural communication in computer-delivered tests; features that are otherwise elicited in an interview format.

Computers can't replace human interactions; gestures, eye contact, etc. are all parts of language ability. The purpose of the speaking test is to test candidates' ability to speak in a natural communicative environment. (ET25)

We have an interview because we are interested in communicative abilities and skills that you cannot get from other things. It's like you are cutting your nose to spite your face! In essence you have an interview because you can't test in a computer. (E12)

Answering questions on a computer is not enough. What about body language? Intonation? And also responding to what has been said? People need to be able to talk to a person. (E15)

One examiner pointed to the potential for computer-delivered assessment to simulate certain real life conditions – for example, giving a timed lecture or speech – by imposing time restrictions, although he still maintained that a face-to-face format is stronger.

Affective factors and provision of support

Drawing on their professional teaching and testing experiences with other computer-delivered tests such as TOEFL, examiners associated more stress and anxiety with computer-delivered assessment and believed that a face-to-face format helps in reducing stress, allows for better supporting of candidates, can help elicit candidates' best performance, and is also better value for money.

When face-to-face with another person you have lots of options to support a candidate whether it is facial gesture like a smile or a hand to say continue but the computer does not do that. (E05)

I see a lot of pitfalls and lots of stress with the speaking part of the TOEFL – they are worried about so many things and having to talk into the computer, you've got the timing issue that IELTS doesn't have and that's a good thing for candidates. (E10)

I think it would make it easier for the candidates if there is a human touch – you can put them at ease and be friendly. (ET08)

I have taught TOEFL preparations…and they are very different. TOEFL does not give leeway for emotional reactions, or being sick running out of the room but a face-to-face interaction makes the student much more relaxed. With IELTS, you can skip questions or take your time and go as slow and fast as you like. Face-to-face in general is much more calming but computer-based can be very jarring. (E11)

We have to remember that most people are very nervous and a human voice can be very reassuring and having someone face-to-face can be really helpful. Someone can feel much more reassured with a smile and we can put them at ease. You can probably elicit all forms of language in computer but you can help them to perform to the best of their ability in a face-to-face test. (E14)

My students hated TOEFL and talking to a computer. They prefer the interaction and it puts them at ease. They feel to be taken more seriously. Also they are putting so much money why just use a computer? If I am paying 200 Euros, I'm not paying 200 for speaking to a computer. (E36)

Scepticism towards technology

Examiners raised concerns about the reliability of technology and automated assessment, and pointed out risks such as biasing against particular language backgrounds or candidates memorising responses and cheating such systems.

Until technology is good enough, a human has to be in charge of it. Otherwise, you'll be messing around with the kids. Even BBC [British Broadcasting Corporation], that is probably using the best technology [in sound recognition for subtitles], gets words wrong. A scenario where the software has difficulty with exact words by British native speaker on the news, how do you expect it to work for our guys from Pakistan? From the Philippines? Or our friends from Scotland? Case closed! (E03)

A computer will never understand the nuances and subtleties of someone. If you have people with difficult pronunciations, a computer will have a bigger problem. I have a good ear for different accents and computers might not be able to tune in like that. (E36)

I used to work in China where there was a TOEFL test with a computer speaking component and they are good at working out what the questions are. They used to prepare, and memorise and the kind of answers was completely rote so those same students in real life…their speaking skills were [so low] and they just memorised. And you can challenge them better with the face-to-face test. (E16)

Overall, examiners were very supportive of the face-to-face format of the IELTS Speaking Test, which has more advantages than the computer-based format in simulating human interactions, provision of support, flexibility to understand various accents and nuances, as well as combating memorised responses.

4.4 Interlocutor frame

For Q20 to Q22 on the online questionnaire, over 70% of the examiners felt that the interlocutor frames (i.e. examiner scripts) for Part 2 (72.1%) and Part 3 (80.3%) were *appropriate*. However, for Part 1, more than half of them felt that it was *too rigid* (62.1%).

Responses to Q24 indicate in what ways the interlocutor frame for Part 1 could be modified. The values in brackets show the percentage of examiners (N=1109) who ticked each option:

- an optional extra question in Part 1 frames should be provided (37.4%)
- there should be an optional extra topic in Part 1 in case the candidate completes the first two topics quickly (50.0%)
- in Part 1 frames, there should be the option to ask the candidate 'tell me more' instead of 'why/why not'. (83.4%)

In contrast to Part 1, the flexibility in Part 3 was appreciated and exploited by nearly all of the examiners. The responses to Q23 of the questionnaire indicated that 79.7% of examiners either *frequently* or *always* ask their own follow-up questions in Part 3.

With these questionnaire results, we explored examiners' views and suggestions regarding different aspects of the interlocutor frame in the interviews, and present our findings as related to each test part below. Although some new examiners found the interlocutor frame as 'something not to worry about' (E10) because it facilitates test management and helps reduce cognitive load, we found, once again, that some examiners felt that increased flexibility would perhaps enhance test performance. Note that some findings overlap with those reported earlier in Section 4.1.

4.4.1 Part 1

Examiners in the interviews expressed negative views towards the interlocutor frame in Part 1, using adjectives such as *inflexible, too specific, too strongly scripted, too stilted*, and *heavily structured* to describe the frames, and they subsequently discussed its adverse impact on their rating behaviour and on candidate performances.

Yes/No and Why/Why not questions

Several examiners criticised the use of these binary questions in Part 1 on the grounds that: (a) such questions do not necessarily fit the preceding interaction; (b) answers to these questions may have already been given by the candidate; (c) questions may not follow up smoothly from what was previously said; and most importantly (d) they do not necessarily elicit responses that help examiners distinguish between different ability levels.

> *'Why' is sometimes the only word we are allowed to utter to generate a response although it sometimes does not fit at all.* (E01)

> *Some rubrics simply don't work at all. We don't have the freedom and hope we can string it out long enough. Sometimes candidates cover all the options within the first answer but we have to ask all the questions again anyway.* (E05)

> *Also the ways prompts are introduced, we are forced to read exactly what is in the booklet. It's too banal and they have to rethink transition in a prompt. It sets us up and lets the examiner seem less credible to the candidate.* (E12)

> *I prefer 'tell me more' to 'why/why not' partly for variety – asking why 12 times in a row. Also candidates may feel they already answered the why question*. (E13)

In fact, according to the *Instructions to Examiners* (2011), the why/why not questions are only optional. As such, comments regarding (b) may indicate individual examiners' (or trainers') unfamiliarity with some specific aspects of the *Instructions to Examiners*, as they are indeed allowed to skip questions where answers have already been provided by the candidates.

More flexibility

The need for more flexibility in the rubrics and the interlocutor frame was once again emphasised by the interviewees:

> *In Part 1 I feel like a robot because I have to ask exactly the questions written down and only why/why not – I'd like to say 'when was that' or 'why was that' and follow up questions could be more flexible.* (ET08)

> *Sometimes the candidate might say something interesting and you'd want to follow up…but the script doesn't allow you to.* (E07)

From these comments, it seems worth considering including 'Tell me more' as one of the follow-up prompts in Part 1, which would allow examiners to follow the threads from the previous responses of the candidates. In fact, over 80% of the questionnaire respondents *agreed* that 'Tell me more' should be an option (Q24).

It was also pointed out how some candidates may go off track or forget to respond to a part of the question, but that the frame does not allow examiners enough flexibility to re-direct the conversation as illustrated in the following example.

> *Sometimes questions are misunderstood, e.g. fruit/food and candidates go along a different path and they misunderstand all the questions and it just goes completely haywire.* (E05)

It is worth noting that, regardless of their criticisms, examiners were sensitive to the need for standardisation in the IELTS Speaking Test and one examiner suggested a balance of the two.

> *I would love it if I could formulate the questions myself, but I understand that I don't have the choice because of standardisation. So, instead, the test makers can provide us with more options.* (E02)

4.4.2 Part 2

Rounding-off questions

The rounding-off questions in Part 2 appeared to be the most problematic aspect of the interlocutor frame with several examiners finding them redundant or unhelpful. The following quotes shed light on these findings.

> *The problem with the rounding-off questions is that those who can talk will probably end up talking more and you have to cut them short. And other times, the questions have already been answered so they end up being redundant.* (E18)

> *We always have a conflict on whether we have time to pose the question or not… If you have a weak candidate and you ask the question, by the time they have thought of the answer you'll run over time. They need the question, reflection, formulation time.* (E05)

The solutions suggested by examiners included making these questions optional, providing extra time for weaker candidates, including a variety of questions for examiners to select from, closing the frame with a simple indication of engagement such as 'thank you for telling me about x', and giving examiners the flexibility to formulate their own follow-up questions in order to relate them to what the candidate has said.

It should be noted here that the rounding-off questions were indeed introduced as optional in 2001, and remains so at the time of writing this report. The *Instructions to Examiners* (2011) specify that examiners do not always have to ask round-off questions in Part 2. Also, each examiner frame for Part 2 clearly states that examiners can choose to use, or not to use, any of the rounding-off questions. Therefore, it seems that the need for rounding-off questions has been over-interpreted over the years, making some examiners feel that asking at least one rounding off question is obligatory. Thus, we recommend measures to ensure that the right guideline regarding rounding off questions is known to examiners; not only new examiners, but also experienced ones.

Freedom to paraphrase

One examiner expressed a preference for having the 'freedom to paraphrase' (E06) key words for Part 2 topics. This was based on the following observation:

> *Sometimes when you give an individual long turn, the whole response relies on an understanding of one or two words (e.g. 'leisure time' or 'past time') and if you don't understand it then there is not a lot of context to help you. A good student will take a guess, but I'd like to have the freedom to paraphrase if the student does not understand a key term. (E06)*

It should be noted that examiners are allowed to provide the meaning of a word in Part 2 if the candidate specifically asks for it. The reason why examiners cannot do so spontaneously is presumably because spontaneous provision of support can vary to the extent that it would pose a serious threat to the uniformity of test administration. This issue could be addressed better by raising candidate awareness that they can ask for clarification.

4.4.3 Part 3

Compared to Parts 1 and 2, examiners' views on the interlocutor frame of Part 3 were more positive owing to a less rigid frame and the potential to challenge higher ability candidates.

> *In Part 3, you have the freedom to change direction.* (E05)

> *High level students can be challenged in Part 3.* (E12)

Nevertheless, increased flexibility was again suggested by our interviewees in terms of the number and types of questions asked with the view to enhance the naturalness of the interaction and adapt to candidates from different proficiency levels.

> *I like Part 3, but I find it difficult to ask all four questions under each of the two; again, I'd like the flexibility not to ask all the questions.* (E06)

> *The way Section 3 [Part 3] is constructed…it's like somebody has recalled a conversation and they picked out the most pertinent points…but conversations can go in multiple different directions, so it would be better to have the opportunity to develop the conversation more naturally and let us [examiners] pick up the points. It would be nice to be able to draw them out a bit more about what it is they were saying rather than move them on to something they wouldn't touch on.* (E03)

> *In Part 3 you can go off script and that's a good thing…though some lower level candidates cannot handle the complexity of more abstract questions.* (E16)

4.4.4 General comments on benefits of increased flexibility

Some examiners commented more generally on the interlocutor frame across the IELTS Speaking Test rather than in relation to a specific part. The main emergent themes are presented below.

Enhancing interaction

Examiners expressed a preference for exercising more freedom in managing the test in terms of various features of timings, the development of the interaction, the types of questions asked, and the sequencing of questions in order to enhance interaction and help create a more positive test experience. Illustrative comments are reproduced here.

> *Your level of interaction is limited because you have to follow that script.* (E03)

> *Sometimes the candidate might say something interesting and you'd want to follow up, but the script doesn't allow you to.* (E07)

> *I would like more freedom with time. Also, in terms of developing conversation so that not everything is scripted.* (ET08)

> *It's meant to be a conversation or a dialogue, but you don't get that.* (E36)

> *I'd like to decide what to say.* (ET08)

> *Sometimes the wording of questions and prompts sound like you are interrogating the candidates.* (E02)

You keep interrupting and I always feel so impolite. (ET08)

Have more 'tell me about' type of questions or 'describe for me'…these types of questions. (E10)

Main issue that I've had is the timing of things. You consistently have to do timings in your head [while interacting and listening to candidates]. (E11)

I have no problems with pre-scripted questions but a bit more freedom to ask 'how did you find it'. (E36)

A set of questions to choose from and not having to take them in the order stipulated. (E14)

Helping candidates understand the questions

Besides a more natural development of interaction, another main reason underpinning examiners' call for increased flexibility in the interlocutor frame is their perception of the need to help ensure that candidates understand the questions. According to the *Instructions to Examiners* (2011), in Part 1, examiners can only repeat a question once and no rephrasing is allowed. In Part 2, repeating more than once is permitted if candidates do not understand some of the vocabulary in a given topic, but examiners can explain the meaning of a word only if explicitly asked by the candidate. In Part 3, examiners can rephrase proactively, and as often as deemed appropriate. Some examiners expressed a sense of helplessness in not being able to explain a difficult vocabulary item and help the candidate understand the question in Parts 1 and 2.

Sometimes the ways the questions are put are confusing, and examiners' hands are tied – examiners can't clarify even if the candidates don't understand. (E30)

In Part 1, you have to ask questions as they are scripted. If the candidate doesn't understand the question, the only option is to repeat. Lower proficiency candidates may miss a question – if you repeat and they still don't get it, then you have to move on. They [candidates] panic if there's too much language. They just switch off. (ET25)

If they say I don't understand the question, examiners can only repeat the question. The only thing examiners are allowed to do is to give a short gloss of the word if the candidate asks. It's quite frustrating sometimes. For example, if you ask, 'Do you like gardening?' The candidate probably wouldn't say 'What does gardening mean?', but only say 'I don't understand', and then you can only repeat the question, and move on. But why can't we help? (ET21)

The only way [to assess lower bands more effectively] is to be allowed to simplify the input material, or language of the prompts. If the prompts are too difficult to understand, they can't answer, especially for Part 2: I've had candidates who just looked at the questions for a minute just give it back with the word 'sorry'. Unless they explicit ask about a word, you can't say anything…In that situation, examiners can usefully be allowed to give some hints or simplify the question even if the candidate hasn't explicitly asked. (ET21)

ET21 further added that there might be instances where candidates' performance drops in Part 2 because they did not understand the prompt.

From these comments, it could be inferred that examiners regard candidates' comprehension of the prompts not as part of the construct being assessed in the current speaking test, but a requisite condition for candidates to respond meaningfully and produce adequate and appropriate language sample for assessment of their speaking ability.

Examiners therefore suggest relaxing this particular aspect of the interlocutor frame in Parts 1 and 2 in terms of 1) allowing provision of glosses for difficult vocabulary items, and 2) paraphrasing and simplifying the language of the prompt where necessary.

Thus far, quite a few examiner comments in relation to relaxing the interlocutor frame have been presented, and it is essential to consider these comments in the light of the history of the IELTS Speaking Test and IELTS research in the past 15 years. One of the rationales for the 2001 revision of the IELTS Speaking Test was to standardise examiner input in terms of the organisations of turn-taking and sequence, and topic and repair management, in order to promote valid and fair assessment of English speaking proficiency. The changes in examiner input introduced in 2001 (as well as more structured task format) was driven by discourse-based studies (e.g. Brown, 2003; Lazaraton 2002) which clearly showed that examiner variability allowed in the 1989 version of the IELTS Speaking Test could lead to advantaging or disadvantaging some candidates.

However, a number of discourse-based studies on IELTS Speaking in the past 15 years seem to suggest that the standardisation of examiner input in the current post-2001 version of the test is overly implemented (O'Sullivan and Lu, 2006; Nakatsuhara, 2012; Seedhouse and Morales, 2017; Seedhouse and Nakatsuhara, 2018). Based on the findings of such literature, the IELTS Speaking Test might benefit from striking a balance between 'the need to standardise the test event as much as possible (to ensure that all test-takers are examined under the same conditions and an appropriate sample of language is elicited) against the need to give examiners some degree of flexibility so that they (and the more directly affected stakeholders) feel that the language of the event is natural and free flowing' (O'Sullivan and Lu, 2006: 22). It is particularly important to note that O'Sullivan and Lu's (2006) study on the current IELTS Speaking Test showed that the impact of a certain type of examiner deviation (i.e. paraphrasing questions) seemed to be minimal on the resulting scores. Considering O'Sullivan and Lu's study was conducted before the establishment of the examiner support network with stricter code of practice, their findings provide further support for allowing some more flexibility in the current interlocutor frame for the rule of paraphrasing questions in the future, as doing so would not affect the scores.

Given the way in which IELTS Speaking had been shaped in the current 2001 form and in light of the IELTS literature in the past 15 years, the examiner comments provided in this study about increasing flexibility in the interlocutor frame need to be interpreted in the spirit of striking an optimal balance between standardised examiner input and a natural interaction, making the best use of the face-to-face IELTS Speaking format.

4.5 IELTS Speaking Test: Instructions to Examiners

The vast majority (85.4% [Q25]) of the questionnaire respondents found the examiner handbook, *IELTS Speaking: Instructions to Examiners*, helpful for administering the test; however, a lower percentage (68.2% [Q26]) believed that it covered all necessary guidelines and questions. We therefore asked examiners to elaborate on aspects of the guidelines that could be improved and we discuss these below.

Special circumstances

Examiners highlighted a need for guidelines that facilitate dealing with special circumstances – and not just 'clear-cut cases' (ET25) – for example when candidates break down due to stress or a sensitive question or topic.

> *Part 2 often elicits an emotional response and candidates might start crying. Managing that is a bit tricky. More guidelines would be good – in those situations, it's a conflict between your human responses versus your examiner responses.*

The whole thing is so streamlined and regimented. I get it because it's for reliability and consistency. I'm not suggesting I want to change it. But just need acknowledgement that it's two humans in a room. (E18)

An examiner from Germany talked about the experience of examining refugee candidates which highlights the need for careful and sensitive handling of such cases with necessary guidelines.

Sometimes we have candidates who have spent time in a refugee camp; and they didn't have a 'childhood toy' to describe and this can be quite insensitive. (E05)

An emerging theme from the above comments is the human element of the Speaking Test with examiners at times experiencing as mentioned above a 'conflict' or tension between two roles – an examiner on the one hand and empathetic listener on the other – particularly in some of the special circumstances described above. The desire to make the test a bit more human is captured in the comment below.

Would be nice to have a little more scope to be human, e.g. a candidate coming from the same town as my wife…I'd like to say something like 'That's nice', 'I've been there', or anything at all. You just have to suppress it all, like if the candidate says their parents have died. You will say next, 'Now let's talk about your favourite park.' (E18)

This comment seems to relate to the rule that examiners must refrain from using response tokens such as 'good' and 'excellent' which candidates may misinterpret as an evaluative comment on their performance (Taylor 2007, p. 189). However, since some empathetic comments are not evaluative, for example 'I'm very sorry to hear that', there is likely to be room for allowing such short non-evaluative phrases to facilitate smoother interaction. Nevertheless, it is important to restrict such additional examiner comments to a selected set of phrases.

Candidates with special requirements

Related to the needs for an increased degree of accommodation for special circumstances, it is also worth noting that several examiners (although not available for interviews) left comments on the online questionnaire, which requested explicit guidelines and procedures for candidates with special requirements, such as those with speech impediments.

More info or training on candidates with special requirements (this is obviously individual for each candidate, but there could be more guidance in terms of timing when dealing with people who stammer, stutter, etc.) (Respondent 318)

Clearer standards for candidates with special requirements should be written and discussed in training. (Respondent 449)

While the *Instruction to Examiners* (2011) has a dedicated page of guidelines for assessing candidates with special requirements regarding the provision of extra time, the use of access technology and modified materials (e.g. in larger print), continuous improvement with more specific guidelines may be helpful.

Native speaker candidates

Referring to the range of English varieties used by native speakers around the world, some examiners requested guidelines on what might be considered 'characteristic of native speakers' (E01). An illustrative comment is presented below:

In India, the present continuous is acceptably used a lot more compared to my context. Is that a 'mistake'? And there is always this kind of nagging questions. (E01)

Linked to this, is the problematic notion of 'a native speaker', the understanding of which might differ from one context to the next.

> *We are always listening for the native speaker that we are familiar with and that is not very fair.* (E01)

While noting that language tests and language benchmark standards nowadays no longer make reference to Native Speaker competence (Taylor, 2006), it is essential to remind ourselves 'the importance of the construct of a test and its score usage when considering what Englishes (rather than 'standard' English) should be elicited and assessed, and when/how we can reconcile notions of 'standard' English with local language norms without undermining the validity of a test or risking unfairness for test-takers' (Nakatsuhara, Taylor and Jaiyote, 2019, p. 188). What is equally important to note is that incorporating every single variety of English in large-scale international examination contexts and guaranteeing fairness to all candidates is unrealistic, since language testing is 'the art of the possible' (Taylor, 2006: 58). As such, it is important for examination boards to select the variety (or varieties) of English in principled and justified way, in order to best sample and assess candidate language that is in line with the construct of the test and make the construct transparent to the users of the test (Harsh, 2019).

Other areas that would benefit from more guidelines

Echoing the needs for increased flexibility in the interlocutor frames that was discussed earlier (Sections 4.1 and 4.4), examiners mentioned the relevant areas for which they would require more guidance and like to have emphasis in the handbook and training, namely:

- how to facilitate eliciting more speech from candidates 'drying up' in a test part
- follow-up questions
- timings of different test parts
- how to deal with candidate misunderstandings within interlocutor frame restrictions.

4.6 Administration of the test

For the overall length of the test (Q28), 86.8% of the examiners felt that it was *appropriate*. For other aspects of test administration, over 60% of the examiners *agreed* or *strongly agreed* with the statements below.

- The task of keeping time for each part of the test is manageable. (67.5% [Q29])
- The examiner's dual role of being the interviewer and the rater is easy to manage. (66.1% [Q30])
- It is easy to adhere to the guideline of administering test sessions for no more than eight hours a day. (66.2% [Q31])
- It is easy to adhere to the guideline of taking a break at least once per six test sessions. (70.4% [Q32])
- It is easy to adhere to the guideline of conducting no more than three test sessions per hour. (69.3% [Q33])

In the interviews, some examiners commented on the challenges of having to both administer the test and rate candidates. This was also linked to a need for more practice and training for new examiners, which is further discussed in a later section (Section 4.8). Some illustrative comments are reproduced below.

> *The role requires a lot of handling of materials, questioning, rephrasing questions. It requires a lot of mental stamina for examiners. Inexperienced examiners find it very difficult to rate candidates immediately.* (ET25)

Directing their attentional resources on one task has often come at the expense of another. One examiner, for example, comments on the tension between keeping to the time limit and maintaining interaction with the candidate.

> *This is a one-to-one situation, so you have to keep an eye on all three things, and it's psychologically difficult. You sometimes have to withdraw temporarily from the interview [interaction]. The focus of the candidate is completely on the examiner, so it can break up the relationship. It's not so much a struggle now after examining for a long time, but in early years it's very tricky. It feels so rude – towards the end of exam, with a very nice dialogue developing, and then you have to say, thanks, the exam is over.* (E26)

Another examiner reported having difficulty managing time-keeping and evaluating the candidate's response simultaneously.

> *We need to be strict with time-keeping – constantly keeping watch of the timer, which takes some focus away from listening to the test-taker's response. I realise that I don't rate them only by that segment, more to whole response. But when sometimes I'm really listening to the test-taker, I find myself five seconds over. Maybe there's a better way to do this.* (E31)

Such cognitive demands in multi-tasking seem particularly challenging for new examiners. This was alluded to by E26 above. One relatively new examiner (with less than 1.5 years' examining experience) also reported:

> *Managing the dual role was more challenging at first, managing just the timing of the whole test with six minutes in between tests — there's not enough time to think about candidate performance before the next test and give the rating. There's still some challenge in having to mentally pin down the candidate score, while keeping the discussion flowing.* (E32)

A more experienced examiner commented that it took nearly a year until she got used to multi-tasking and managing the dual role in the test.

> *To be comfortable doing IELTS, I needed about a year and the first couple of sessions were pretty nerve-wrecking.* (E05)

This examiner (E05) also referred to her experiences of other exams – where the assessment and interlocutor roles are separated (e.g. Cambridge General English examinations) – as easier and less demanding. This issue is again reported in Section 4.7.8.

Mental fatigue

Although there is a strict Code of Practice for test centres for scheduling tests, which prevents examiners from conducting more than three tests per hour and examining for more than eight hours a day, examiners working in certain regions reported suffering from mental fatigue that stemmed from conducting many tests per day, and/or having candidates with very similar proficiency levels or repetitive questions and responses.

> *In a place where most candidates are [Band] 5.5, it's difficult at the end of an eight-hour day to pick up somebody who may be a bit weaker or stronger – you think all of them are 5.5. This is a very natural human thing, as the exam requires a lot of concentration and focus. Even for experienced examiners, it's very tiring.* (E19)

> *Eight hours a day is manageable, but not five or six days in a row. Mental fatigue does come in. Three days a week is manageable and used to be the case.* (E22)

Repetition is a problem for examiners in [country name where examiner is based], with such huge volume of candidates. Repetition of the same questions over and over again has a negative effect on examiners...From a psychological perspective, with a high examining load for examiners, and responses being so repetitive, you stop listening to what the candidates say before you hear the complete response. (E22)

However, it is worth noting again that, judging from the online questionnaire responses, over 60% of the examiners (66.2% [Q31]; 70.4% [Q32]; 69.3% [Q33]) regarded the current test scheduling to be manageable. There are clear requirements in place for test centres and examiners in order to ensure that examiners are not overworked. Nevertheless, as some examiners commented in the interviews, it may be necessary to consider adding to the requirements (e.g. limiting scheduling tests for eight hours per day to three days a week) in regions with high volumes of candidates.

4.7 Rating

For the rating of the IELTS Speaking Test, the online questionnaire included four areas to collect the examiners' views (i.e. rating scales, bands, use of audio-recordings and examiner handbook). The first area asked about how easy it is to apply the four rating scales, for which the responses were as below:

- Fluency and Coherence (74.8% [Q35])
- Grammatical Range and Accuracy (78.5% [Q36])
- Lexical Resource (80.0% [Q37])
- Pronunciation (53.1% [Q38]).

For the first three scales (Fluency & Coherence, Grammatical Range and Accuracy, and Lexical Resource), nearly four in five examiners found the descriptors in each rating category easy to apply. However, the Pronunciation scale had a much lower agreement rate in comparison to the other scales. Different aspects of rating were explored in more detail in the interviews and are discussed in Sections 4.7.1 to 4.7.4.

The second area involved the number of bands and how different bands are measured by the test. Of the examiners, 84.7% *agreed* or *strongly agreed* that having nine bands (as IELTS currently does) is appropriate (Q39). Among the nine bands, the middle bands (i.e. Bands 5.0 to 7.5) were perceived as being assessed more accurately by most examiners (77.9% [Q40]), followed by the higher bands (i.e. Bands 8.0 to 9.0: 69.5% [Q41]). This is in line with the original IELTS Speaking Test design that aimed to most reliably differentiate candidates at the middle bands for various decision-making purposes for which IELTS might be used (Taylor & Falvey, 2007).The results of the follow-up interviews are presented in Sections 4.7.5 to 4.7.7.

The third area in the Rating section of the questionnaire asked about the use of audio-recordings for the test. Of the examiners, 84.8% *agreed* or *strongly agreed* that it is appropriate for second-marking (Q43), and 80.7% of the examiner trainers did so for monitoring purposes (Q44).

The fourth area explored the use of the examiner handbook. The frequency of reviewing the *Instructions to Examiners* at the start of an examining day (Q45) varied among the examiners: 2.5% answered Never; 11.7% Seldom; 27.2% Sometimes; 25.5% Frequently; and 33.1% Always. This question was developed from the focus groups for constructing the questionnaire with the three experienced examiners, who suggested it might be useful, subject to time availability, to review the examiner handbook at the start of an examining day, but there is usually little time to do so. The questionnaire responses seem to indicate otherwise, with a total of 85.8% of examiners being able to review it 'sometimes' or more often. Still, 14.2% of examiners either 'seldom' or 'never' review it,

so it may be useful for test centres to allocate a dedicated time slot before the start of the examining day to review the *Instructions to Examiners*.

Below are the themes that emerged from the interviews regarding rating in the IELTS Speaking Test. Various issues were raised, and we believe that providing further guidelines and more illustrative sample performances at different bands would be helpful in addressing them. If difficult to incorporate in the current certification and re-certification processes due to limitations of time and resources, making a fuller use of, as well as expanding the pools of self-access materials at test centres may be hugely beneficial.

4.7.1 Fluency and Coherence (FC)

Conflating two criteria

One the main issues raised about the FC scale was the conflation of the fluency and coherence criteria into one category. Examiners observed overlap between bands or descriptors; for example, slow pace of speech but frequent use of discourse markers, or fluent speech but problematic pacing. Some examiners highlighted the need for more guidelines and training.

> *Sometimes you see jagged performances.* (E36)

> *More standardisation and training on trade-offs between fluency and coherence would be good.* (E02)

> *Sometimes there are people who speak fluently but the pace isn't right. Sometimes we have speakers from India and they are very fluent and they talk a mile a minute and there is no effort but there might be loss of coherence.* (E05)

Although there are guidelines for rating such candidates in the *Instructions to Examiners*, the examiner comments above highlight the need for raising awareness among the examiners of the available materials, as well as potentially including benchmarked samples of candidate performances with an uneven profile across the four rating criteria in the standardisation.

Subjective performance indicators

Examiners referred to the subjectivity of some of the FC descriptors such as speed or comprehensibility, which made assessment more challenging compared to some of the other criteria.

> *FC is not as easy to measure compared to some of the other criteria; for example, grammatical mistakes…or a complex sentence is a complex sentence but questions about how comprehensible something is or the speed of an utterance can be rather subjective.* (E01)

Indeed, Brown's (2007) verbal protocol study on the IELTS Speaking Test indicated that the examiners in her study found the FC scale the most difficult to interpret. Galaczi et al.'s (2012) large-scale IELTS examiner survey with 1142 respondents from 68 countries also reported that more clarification and exemplification for terms used in the FC scale, such as 'cohesive devices', 'discourse markers', and 'connectives', are needed. Additionally, some respondents in their study also commented on how speech rate as a measure of fluency can take into account personal speaking styles of some candidates.

4.7.2 Grammatical Range and Accuracy (GRA)

For the GRA criterion, E20 commented on the difficulty in applying the descriptors for GRA to candidates in specific L1 or learning contexts. In her context, candidates seem to have a profile of grammar development different to the profile reflected in the descriptors:

[Rating] GA in this part of the world, it's very difficult, as many candidates have fossilised features, and there are many Band 4/5/6 candidates. The descriptors say: able to use complex sentences, and basic sentences should be accurate. But candidates here are fossilised in basic sentences. They do use complex ones, but a lot of the basic grammar is inaccurate. (E20)

Examiners also cited difficulty in evaluating range / complexity of syntactic structures. E27 raises the challenging question of how to balance in rating the trade-off between accuracy and complexity:

For examiners, it's easier to listen for accuracy than to listen for complexity. (E22)

Sometimes you get candidates with lots of colloquial language and complex grammar, but they make more mistakes. So, how far do you penalise mistakes and how much do you credit for the good stuff? That is a very difficult judgment to make. And in training, there needs to be a lot more emphasis on what is a 6 and what is a 7. (E27)

The tension between accuracy and complexity was also problematised by a number of examiners in Galaczi et al.'s (2012) examiner survey. Furthermore, as noted in Section 4.1.3 earlier, spoken grammar does not necessarily involve complex grammatical structures compared to written grammar.

The above comments, again, highlight the need for ensuring that examiners are aware of the glossary in the *Instructions to Examiners* which include definitions of what is meant by 'frequent' and 'usually' when observing errors and rating jagged performances.

4.7.3 Lexical Resource (LR)

Of the examiners who responded to the questionnaire, 80.0% indicated that they find the descriptors for Lexical Resource easy to apply (Q37). E22 commented that the descriptors for LR seem to work the best and most of his examiner colleagues find them the easiest to use of all the rating criteria.

Other examiners commented on how other aspects of the candidates' performance (e.g. pronunciation, fluency, and familiarity with the task topic) may have an impact on examiners' evaluation of the candidates' lexical resource:

It's hard to rate candidates who seem to be extremely fluent but lower grammatical accuracy and high lexical resource – most challenging. They have so many accommodation and repair strategies, making it harder to notice their mistakes. (E32)

Pronunciation affects the evaluation on other criteria, with the [candidate's] L1 accent feature sometimes marring display of lexical resource. Examiners are not empowered by the band descriptors to do due diligence to identify the word production and not penalise on lexical resource because of issues with pronunciation. This should be made clearer in the band descriptors. (E23)

If the candidate is not familiar with the topic, there is no option to change the topic. This puts the candidate at a serious disadvantage. Some candidates, especially intermediate and undergraduate students, are extremely good at grammar and pronunciation but are not exposed to certain cultures or are not aware of certain lifestyles...It's difficult to apply the descriptors, namely Fluency and Coherence and Lexical Resource, for such candidates. (E34)

4.7.4 Pronunciation

Detailed descriptors for the odd bands

The absence of detailed descriptors for the odd bands in the pronunciation scale was negatively viewed by several examiners and perhaps best explains the disagreement rates in the survey results.

> *What would help are more detailed descriptors rather than just one single statement like all the positive features of band 6 some of the positive features of band 8.* (ET25)

> *You have descriptors for every other band and then all of band 5 and some of band 6. That is difficult to apply so descriptors need to be more fleshed out.* (ET08)

> *Why is it that the pronunciation scale has 'meets some of the positive [features] but not all of them? It takes me more time to identify the one above and the one below. They have a gradient of ability, but pronunciation is a catch all of some.* (E11)

> *What annoys me is only having the descriptors for 2, 4, 6, 8 and no intermediate ones. When I started doing it, we only gave these bands and now they have added things but they are not clear. I would like these descriptors to be spelled out more clearly.* (E16)

The need for delineating specific pronunciation features at Bands 3, 5, 7 has also been suggested in Yates, Zielinski and Pryor's (2011) IELTS examiner perception study, as well as in Issacs, Trofimovich, Yu and Chereau's (2015) IELTS examiner judgement study on different elements of features contributing to the IELTS pronunciation scores. Issacs et al.'s (2015) findings are of particular relevance for designing new descriptors, e.g., clear distinctions between Bands 6 and 7 for comprehensibility, vowel and consonant errors, word stress, intonation and speech chunking.

Distinguishing accent from clarity

Examiners pointed to the subjective nature of determining accent and clarity and the challenges of distinguishing between them.

> *Those are the major difficulties that you have at the high level because one examiner's lack of clarity may be another examiner's accent.* (E12)

Mispronunciations and impact on coherence

One of the examiners raised an issue regarding the impact of mispronunciations of key words on comprehension and the need to have further guidelines for such cases.

> *The emphasis is on prosody, rhythm, communicative ability as the big picture kind of things. But what do you do when candidates completely mispronounce one or two words and they are highly frequent which might detract from comprehension? They aren't contained in descriptors but it could be as part of additional information.* (E09)

4.7.5 Higher bands

Band 9

This was an issue raised by several examiners who described this band as 'too harsh' (E09), 'not always realistic' (E36), and 'discriminatory' (E03). They questioned the wording of the descriptors – including reference to L1 accent – and highlighted the need for more clarification and guidance particularly when assessing English L1 varieties.

> *I think there is a question about the top band. It's not necessarily the same as a native speaker – it 'suggests' native speaker, but a tiny bit of an accent is ok. A handful of errors are still allowed often. What does that actually mean? More clarification is necessary.* ***(E05)***

You're looking for sophisticated speakers, making few mistakes, who have a wider range of vocab, [linguistically] sophisticated vocab [to award a 9.0]. (E30)

Who do we want to have as a band 9? What is this elusive level of perfection searching for and how generous can we be with it? People's voices and accents… there are plenty of them, they are perfectly clear, as clear as the ice from the crystal-clear rivers. They have perfect grammar going through their late Victorian grammar. But because of accent they are not a 9. Again, I ask; do they have to be able to grace the stage of the West End or someone who can just go to the pub from the ship docks? (E05)

Indian English, Nigerian English. These candidates are speaking their L1, but could be difficult to understand for the examiners. They are native speakers and totally functional just difficult to judge in pronunciation or cohesion when you (the examiner) don't understand. It's a Global Englishes issue. (E13)

Relatedly, examiners raised issues about challenges of distinguishing between Bands 8 and 9, given the similarities in the descriptors and expressed a preference for more details.

Wording between [Bands] 8 and 9 is very similar, so difficult to distinguish. The descriptors don't give you a lot of support, it becomes very subjective. (E22)

The IELTS Partners, upon contact from the research team, responded that Band 9 is not looking for an 'ideal' or 'perfect' candidate or English variety, nor distinguishing native or non-native speakers. There are pools of benchmarked performances, some with specific L1s (i.e. south Indian candidates), that examiners can access at any time through test centres if more support is needed to define Band 9.

While it needs to be acknowledged that IELTS is developed to distinguish candidates around Bands 5–7 most accurately, which are critical bands that can inform the test users to make high-stakes decisions such as university admission or professional registration (Taylor, 2007), it is still problematic if examiners are not clearly instructed when to award the highest bands (and the lowest bands as will be discussed in Section 4.7.7. below).

4.7.6 Middle bands

More descriptors?

In line with the design of the IELTS Speaking Test, examiners also pointed to the middle bands as 'the most important bands to get right' (E27) given the consequences associated with these bands for high-stakes decision-making. They emphasised their view on the potential development of more descriptors:

Given that the vast majority of candidates fall into Bands 5–7, the descriptors do not adequately distinguish between these levels. This is especially true for Band 6 to 7. If an examiner gives a candidate a 6.5 instead of a 7 this can have enormous consequences for the candidate. (E27)

The Band 6 is really wide and difficult to reach a 7. Lots of candidates cluster around the same bands, between 6 and 7. There should be another level between 6 and 7. (E35)

Middle bands are not precise enough and it's this range which is important with consequences for getting into academic program or not. More precision in the descriptors would be very helpful (e.g. adding features to each band). (E26)

While having more descriptors in defining the bands and criteria may be helpful, in a large-scale standardised test like IELTS Speaking, there is a balance to be struck between over-description and conciseness of the rating scales. If bands and criteria were over-described, the scale would not be user-friendly and the descriptors would be too specific to be widely applicable, thus risking hindering the award of accurate scores.

4.7.7 Lower bands

Frequency of encounter and distinguishing between lower bands

Examiners commented on the infrequent occasions in which they have to award very low bands (below 5) and the challenges of reliably distinguishing between these bands.

> *When it's really low, there comes to the point where examiners need to think of some positive aspects of the candidates' performance. Thankfully, not having it very often, but it's difficult. Difficult to decide – you know it's a low grade, but how low?* (E30)

> *Lower bands have a different issue – there is very little difference between 1 and 2. Maybe they say 'my name is…', or answer a question, yes, or no, then it's 1.5, because they have said something. Band 2, maybe they form one other sort of utterance.* (E22)

> *I find that I have most problems in rating the rare candidates who clearly fall into Bands 2 or 3.* (E33)

The role of listening comprehension

The impact of listening on speaking performance of lower level candidates was highlighted in the interviews.

> *If they are weaker, most candidates may not have understood the question. So, it becomes a listening problem…the fact that they are not understanding you. (E07)*

> *The idea of IELTS is of a scale of 0 to 9, suggesting anybody at any level can take the exam. But if someone cannot even understand the question, then what's the difference between a 1 and a 3? They can't communicate, can't even answer questions in Part 1. Part 2 and Part 3 are completely out of their ability.* (E18)

The latter comment also challenges the appropriateness of taking candidates through the whole test when the more difficult tasks are considered clearly outside the ability level of candidates. To this issue, IELTS Partners responded that examiners are supposed to give every candidate every opportunity to demonstrate their ability, and therefore, they should administer all the parts of the test. IELTS Partners also highlighted that, although it is rare, there have been candidates who spoke very little in Parts 1 and 2 but were able to give more detailed answers in Part 3. Because of this, examiners are told explicitly to be wary of making the assumption that candidates will not be capable of producing (sufficient) rateable samples of language in the later part(s) of the test just because they did not do so in earlier part(s) of the test.

These examiner views on the role of listening in the IELTS Speaking Test are congruent with recent IELTS literature. Candidates' listening-related problems in relation to their need for repairs have been reported in IELTS Speaking studies such as Seedhouse and Egbert (2006) and O'Sullivan and Lu (2006). Following these studies, Nakatsuhara's (2012) mixed methods research on the role of listening in the IELTS Speaking Test also identified that those at Band 5.0 and below tend to encounter some difficulties in understanding the examiner. In contrast, candidates at Band 5.5 and above do not seem to have major listening-related problems, and even when they do, they are capable of repairing the problem naturally with an appropriate, specific request. For candidates who cannot understand the current Part 3 questions, such as those at Bands 3.0 to 4.0, Nakatsuhara (2012) recommends the preparation of conceptually easier questions which can be communicated in simpler language to considerably weaker candidates.

This is because, if candidates cannot understand the current Part 3 questions, which are conceptually more complex, paraphrasing would not help. Having a set of prepared questions for these candidates would allow them to follow the examiner, understand what is required and provide further speech sample to confirm their levels.

4.7.8 General comments

Including relevance in rating criteria

One aspect identified by examiners as both an issue within the Fluency and Coherence criterion and a criterion missing in the rating scales overall is *relevance*. Examiners pointed out how off-topic responses – 'a big sign of lack of coherence' (E22) – are not penalised according to the current rating scale descriptors, and this is particularly an issue when it comes to dealing with taught memorised responses to Part 2 questions.

> *Fluency and Coherence: Relevance is missing from these except at Band 8. Not talking on topic is a big sign of lack of coherence that is not in the descriptors for lower bands.* (E22)

> *Candidates may give memorised responses, and examiners do not and cannot penalise them for going off topic. For example, if the question was 'Describe the furniture in your home', the candidate may say something like 'The kind of furniture in my home is leather. Leather is one of my favourite materials…' Or, 'Tell me about a book you've read', and the candidate says 'I read so many books. Reading is very good. Normally we do reading at school. My school has many students…'* (E23)

The kind of memorised responses described by E23 is a string of speech that tangentially touches on several topics but which is marginally relevant to the given Part 2 question.

The issue of relevance does not only apply to the individual long turn (Part 2), but also Part 1 and Part 3.

> *For Fluency and Coherence, it needs something about the relevance of the answer. If someone doesn't understand the question, and they give you an off-topic answer, you're not supposed to penalise, but I feel that they should be. It should assess one's ability to answer the question. For example, for a question about plants – the person talked about their plans, but you can't stop them or clarify. Things like that, it should be somewhere in the rubric.* (E20)

This concerns whether the candidate is 'answering the question', something which also demonstrates the candidate's understanding of the interlocutor's prior talk or reveals non-comprehension or misunderstanding.

E22 talked about how examiners in his context have been instructed to deal with irrelevant responses, and suggested including relevance more explicitly in the rating scale descriptors:

> *What examiners are told to do is to rate 'relevance' on Lexical Resource; but the danger there is that it's not explicit. We talk about range, and whether the lexis is appropriate to the topic, but relevance to the question is a different thing. In IELTS, relevance is only explicitly included at Band 8 in Fluency and Coherence.* (E22)

From E22's comment, it can be seen that relevant, on-topic responses is viewed as a feature distinct from (or more than simply) using vocabulary appropriate to the topic. This lends support for more explicitly including relevance in the rating scale descriptors.

Double-rating

Examiners touched on the fairness and reliability problems associated with having a single examiner and expressed a preference for a 'second opinion' as illustrated below:

> *We always have different opinions. No matter how much work it goes in band descriptors or no matter how much training or experience you have as an examiner, the students deserve a second opinion. In TOEFL there are always two examiners, and if there is a discrepancy, there is a third and that is more fair.* (E01)

Having a second examiner would also help ease the mental effort necessary for managing the test and acting as interlocutor.

> *Having a second examiner would be great because you have to concentrate on so many things at the same time and 12 times in a row is exhausting; both would give a score but the other examiner could focus more on the language so I can focus on procedure.* (ET08)

We must note that the option of double rating was indeed considered during the development of the original IELTS test launched in 1989 and prior to the 2001 revision. However, it was thought sensible to adopt a single-rating system to prioritise the sustainability of the test given the scale of the IELTS test in those days. However, given the current financial stability owing to the exponential growth of candidate numbers during the past two decades (over 3 million candidates in 2017, as compared to 200,000 candidates in 2001), coupled with recent advances in computer technology, which have made the gathering and transmission of candidates' recorded performances much easier in a sound or video format, it is timely to consider the double-marking option once again.

While acknowledging that introducing a second examiner in the entire face-to-face IELTS Speaking Test might not be feasible, as it would change the test operation completely and drastically increase the cost of running the tests, we should also note that some IELTS studies reported examiner severity differences of over half a band (e.g., Khabbazbashi, 2013; Berry et al., 2018). Although a certain level of rater inconsistency is unavoidable (McNamara, 1996), efforts to minimise rater severity variation should be continuously considered to enhance scoring validity and to ensure fairness to candidates (e.g., AERA, APA and NCME, 1999). Berry et al.'s (2018) video-conferencing research, for example, recommended a part of the test (e.g. Part 2) to be double-rated using video recordings that would be available in video-conferencing tests.

4.8 Training and standardisation

In the online questionnaire, we asked the respondents' main roles concerning examiner training and standardisation. As a result, we identified 136 new Examiners, 876 experienced Examiners, 80 Examiner Trainers and four Examiner Support Coordinators. The questionnaire used Q47 to collect this information, and according to which option the respondents chose, either Q48a to Q51a (for new Examiners) or Q48b to Q51b (for all other options) were displayed next. While we are aware that the terms for examiner standardisation and certification are slightly different for new and experienced Examiners and the terms used in the questions were different (see Appendix 1, Q48a to Q51a and Q48b to Q51b), the results are put together as Q48 to Q51 in this report for easier comparisons.

4.8.1 Length and content of training

On Q48 of the questionnaire, over 60% of the examiners with different years of examining experience felt that the length of the Examiner Standardisation is *appropriate* (experienced Examiners: 70.7%, Examiner Trainers: 61.3%, Examiner Support Coordinators: 75.0%), except for new Examiners (47.1%).

Among the new Examiners, 35.3% felt that it was *a bit short* (Q48a), and so did 31.3% of the Examiner Trainers (Q48b). The follow-up interviews explored what would be desirable to be added to the current training.

Similar to Q48, the responses on Q49 showed that the majority of experienced Examiners (72.3%), Examiner Trainers (75.0%) and Examiner Support Coordinators (100%) felt that the number of samples used in the Examiner Standardisation is *appropriate*. However, only less than half (48.5%) of the new Examiners felt it was appropriate, and 37.5% found it was *a bit too small*, which was also explored further in the interview phase.

Regarding the training materials (Q50), across all the roles, nearly 70% or more respondents *agreed* or *strongly agreed* that the materials used in the (New) Examiner Standardisation are useful (new Examiners 69.1%; experienced Examiners 74.2%; Examiner Trainers 83.8%; Examiner Support Coordinators 75%).

In the follow-up interviews, we asked both new and experienced examiners to comment on different aspects of training and standardisation. The main themes are discussed below.

Need for variety and localisation of samples

Some examiners raised the issue of not getting an adequate number of samples in standardisation and re-certification from the candidature of their local examining context, and reflected on the implications for the utility of the re-certification process.

> *Videos usually feature candidates who are western European, Arabic, Indian, Chinese, Korean, Pakistani. We tend not to get videos of Japanese candidates, although those videos from the above backgrounds do sometimes feature in the test centre, but only few.* (ET28)

> *The experience in actual testing is different from the training or re-certification. The re-certification is useful in preparing you to examine around the world for all kinds of candidates, but it doesn't really prepare you for the work you're going to do where you are at.* (ET21)

> *The question then is, what are you being re-certified on? How similar is it to the actual candidates you encounter in the examining context? It is then not a fair procedure and does not add anything substantial to the process.* (E23)

> *The problem is variety not quantity. We tend to see a lot of – well too many – candidates from Asia (especially the re-certification set). We examiners here in the Middle East, would only get one Chinese candidate once every two to three years.* (ET25)

Accordingly, there are suggestions for tailoring the training and re-certification materials to the local examining context, with a higher proportion of test samples characteristic of the local candidature.

> *It's accent hindering communication. In Moscow you might get students from other places but 90% are Russian. And it's a challenge. We need sets more suitable to our context; an L1-specific set.* (E07)

> *Australian examiners need more samples of candidates from India.* (E19)

> *It would be good to have additional materials for rating Chinese and Indian candidates, for the local candidature.* (ET21)

It's good to have a mixture, but the majority of the samples should be from the local region of the test centre. Candidates from other backgrounds would not have direct relevance. For example, the probability of examining Chinese or Russian candidates [the types of samples available] in our local context is very low. I think it's also important for examiners coming from outside of the local region to have access to and familiarise themselves with local test samples. (E23)

However, examiners are also aware of the need for a balance between having a good variety of candidate samples and having more samples representative of the local candidature. The following comments from two examiner trainers are reflective of this view.

Examiners sometimes say that the samples don't reflect the candidature they encounter in their test centre. But they do actually need a variety, to be prepared for the odd candidate from a different background, so examiners actually do appreciate having those in the standardisation. (ET25)

Practically, there is value in having candidate samples from the local context, but it's also beneficial to see a variety of candidates that one doesn't see in their own context. Ultimately, the aim is to train examiners to be able to apply descriptors to performances, so it's useful to be exposed to a broad linguistic range. (ET28)

For a large-scale test like IELTS Speaking, which holds a global candidature, training and standardising examiners with a variety of performance samples is crucial; exposing examiners to performance samples that are not from their examining regions is vital in order to ensure test reliability and uniformity of test administration across different regions and L1s. Nevertheless, familiarising with local performance samples is equally important, so that examiners have more concrete points of reference which are closer to what they will encounter and assess on a regular basis. When the research team asked the IELTS Partners about the possibility of creating self-access pools of localised sample performances, they found that there already exist two localised pools of performance samples, one with Chinese L1 speakers and the other with L1 speakers of South Indian languages due to high demands. These pools are available for self-access, through test centres due to test security reasons, at any time that examiners wish to use them. The IELTS Partners also added that they would develop other localised pools if/where there are strong demands from examiners and test centres.

Effective training with more support materials

In the interviews, there were many positive comments on the training and standardisation procedures, such as that by E03 who praises the selection of good video- and audio-recordings for training and standardisation:

A very well-thought out set of interviews for training and standardisation. Some of them catch you out. Certainly in the sense that we have to think quite hard between the levels. (E03)

However, there were also a number of comments highlighting the areas that the examiners felt needed improvement. An examiner expressed a preference for fuller explanation and discussion than the current commentary accompanying (re-) standardisation performances:

It depends on the trainer. Some just show the video and ask us what we think and just read out what's in this script [commentary]. I need to understand why. (E15)

Similarly, other examiners hoped for more practice in the training – especially for new examiners.

If they allowed more hours for training all around it would be great. It shouldn't be rush, rush and abbreviated. This is a serious thing. It can afford a few more hours. (E10)

At the beginning it would have been nice to have more practice. In order to do organisation, administration, assessment, you need to simulate each person and we didn't have enough time. I didn't get to practice the whole organisation, getting materials together, etc. (E05)

In my first six or seven times [after being certified] I was taking time all through lunch period and listened to the candidates again. I just didn't want to wing it with the scores to 'let the gods of assessment decide'. (E10)

We only really did one full practice session with each other and maybe having a couple of practice sessions with real-life learners and then rotate with four English learners [would be great]. And see if they all come up with similar scores and then talk through all the scores. It'll be much more effective and useful rather than trainer just reading off the scores and commenting out loud. (E10)

All the comments above suggest useful changes that would provide more hands-on practice to the examiners and increase their confidence and perceived readiness in administering tests and rating candidates. However, increasing the length and amount of practice comes with cost implications, which might hinder the implementation of any changes. If increasing the amount and duration of training is not possible, developing and expanding self-access to a pool of performance samples may complement and enhance examiner training. The need for self-access materials was mentioned by different examiners at, as illustrated in E02's comment below:

Self-access materials should be available for everyone and not just for non-standard. They can be complementary for the rather short training. (E02)

By saying 'non-standard', E02 means those examiners who are evaluated as 'non-standard' through monitoring. If examiners get 'non-standard', they need to go through self-access training materials that are available via test centres as part of the standardisation process. When the research team asked the IELTS Partners whether there was a scope for making the self-access training materials available to all examiners, they were told that these materials are already available via test centres to any examiners wishing to access them and they are not just for examiners who are marked as 'non-standard'. E02's comment shows a lack of awareness regarding the availability and accessibility of these materials, which other examiners might also be unaware of or under the wrong impression that these are only for 'non-standard' examiners. Therefore, it is important to ensure that the availability of self-access training materials is known to examiners. In addition, test centres could schedule fewer tests per day in order to make time and allow access to the training materials for the examiners who want to self-train. Lightening the workload per day would also help with mental fatigue that some examiners can experience, as discussed earlier in Section 4.6.

Feedback on scores

Examiners, particularly new ones, highlighted the need for more regular feedback on their performances both as interlocutor and as rater.

As an examiner, and I'm not alone, we are all craving feedback and in the end, it's you in a room with your candidate and you're making the best evaluation. If it's only coming once a year on a few interviews, you wonder am I correct? Is the timing right? You have to wait for a year to get the answers. (E01)

This comment relates to the earlier discussion in Section 4.7.8 under 'Need for a second examiner'. As well as exploring viable ways to provide double rating, it also seems worth considering introducing a more regular feedback mechanism on examiner ratings.

4.8.2 Use of visual and audio recordings

For Q51, over 70% of the examiners *agreed* or *strongly agreed* that they were happy with the use of video recordings for training and audio recordings for certification or re-certification (new Examiners 78.7%; experienced Examiners 77.9%; Examiner Trainers 72.5%). The views of Examiner Support Coordinators varied, but there were only four of them, so we cannot generalise this result with confidence.

This question was developed based on the recent research finding on the double-rating methods of the IELTS Speaking Test that the ratings using video-recordings were comparable to those from live-rating (given in an experimental setting and not in operational tests), whereas ratings based on audio-recordings were consistently slightly lower (Nakatsuhara, Inoue and Taylor, 2017). Although the respondents in the current study seemed generally content with the current arrangements of IELTS Speaking, we explored in the interviews whether they have found it difficult to rate performances with the discrepant use of video- and audio-recordings concerning standardisation and certification. In fact, examiners expressed concerns towards the use of audio recordings due to the loss of visual information.

Views on the use of videos

Examiners generally did not raise any major concerns about the discrepant use of video- and audio-recordings concerning standardisation and certification; a point also verified by one of the examiner trainers who mentioned that this issue has not been brought up in his years of experience. Nevertheless, the use of video was positively viewed:

> *Videos are great, where you can see interaction with the candidate; facial expression, body language, use of materials; so, this is great for training and we need to keep that. But don't really see issues for using video or audio for re-certification.* (ET21)

> *Videos are helpful to use in re-certification as well because that mirrors the test.* (E32)

Views on the use of audio

In contrast, an examiner expressed a negative attitude towards the use of audio for marking given the absence of body language etc.

> *I don't like marking with audio…listening like that creates a distance, nuance is lost without body language or looking at their face so I prefer talking to them or marking videos.* (E36)

Here, it must be noted that there is no mention of the use of body language, eye contact etc. in the descriptors in the IELTS Speaking rating scale. In the communication with the research team, IELTS Partners said that the lack of mention of visual information is deliberate, and therefore, band scores should not differ whether or not examiners are rating with visual information. However, having visual information does complement and help contextualise what is being said, as illustrated by the above comments.

4.8.3 Balance of monitoring and authenticity of interaction

A recurrently mentioned theme that is not directly based on what we asked in the online questionnaire was the strictness of the monitoring in the IELTS Speaking Test and its potentially negative effects.

Some aspects of monitoring are too rigid and should be reviewed. (ET08)

The system of examiner monitoring is intended to ensure that examiners adhere to specified procedures and regulations in test delivery, and thus contributes to consistency across examiners and reliability of the test scores. While its rationale is well-understood and supported by examiners, several examiners expressed their frustrations and reported how they could face penalty for slight deviations from the time limit or interlocutor frame which have minimal consequences on candidates' performance. For example, in keeping to the time limit:

- In Part 2, the candidate is supposed to speak for 2 minutes. If the examiner allows the candidate to speak 1 minute 53 seconds, they get a cross on the box, and this could contribute to a 'non-standard' monitoring outcome.

Another example is in keeping to the exact wording in the interlocutor frame:

- End of Part 2: If the examiner says 'thanks' or 'thanks a lot', instead of 'thank you', they get told that's not standard.
- End of Part 3: Examiners are not supposed to say 'we're running out of time', even though it is only polite when you interrupt the candidate.

E34's experience echoes the above:

> *Monitoring by Examiner Trainer can be problematic – they just go by the procedure, very particular and rigid about timing and the words used in the booklet. For example, if the candidate finishes their long turn within one minute, examiners are to ask 'can you tell me anything more about that' – and it needs to be those exact words.* (E34)

Washback of examiner monitoring

For those who have passed the monitoring, there is still some negative washback on their examining practices, such as focusing their attention on procedure at the expense of listening to candidates' responses.

> *I have increasingly found that we are putting too much emphasis on delivery of the test at the expense of examiners actually assessing candidates. Many examiners focus 80% on procedure and 20% on listening to candidates out of fear of being non-standard examiners!* (ET29)

The interviews also revealed that concerns about getting penalised in monitoring seems to have made some examiners follow stipulated procedures at the cost of being interactionally appropriate – interrupting a fluent candidate's response to slot in a follow-up question:

> *According to examiner guidelines, we should ask follow-up questions [in Part 3], and we're monitored on that. I find asking follow-up questions extremely useful as a strategy when people are not forthcoming, when you want to draw them out. There're other candidates who just talk so fluently, but because I know I'm being monitored on this, I have to find a space in the conversation to ask follow-up questions, just to fill it, so to speak. I find that extremely stressful – why do I have to ask the follow-up questions if she's answered all the questions?* (E24)

Furthermore, the pressure faced by the examiners in managing the interview interaction, evaluating the candidate's performance and adhering to procedures seems to go beyond a matter of cognitive load. The examiners' interview responses revealed constant struggles in their moment-to-moment decision-making with conflicts between their sense of moral obligations on the one hand, and their professional self-interest at stake on the other.

This might not sound like such a big thing, but it is a big thing, as in terms of being a [Band] 6 or 7, it affects university entrance. In fact, from 5.5 to 7, they're the most difficult to grade, because I feel very high stake [for the candidates]. I don't even want a 0.5 up and down if possible, although I know examiners are given 0.5 leeway. If it's 5.5, they need to take a six-month remedial, to enter a program. (E24)

If you think about it from the candidate's point-of-view, the difference between a [Band] 6.5 and a 7 can mean the difference between going to university or not, or being able to emigrate or not. It's really life-changing for them. But for us, if it's a 0.5 difference, it does not really register as a problem in the monitoring system. You can be 0.5 out on every aspect of the test [rating criterion] and you can still be within an acceptable standard. So there's always been this kind of difference between the effect of a 0.5 difference in rating for the candidate and the effect it has on the examiner. So, that in itself is a temptation for the examiner to go somewhere in the middle if you're not sure what mark to give. (E27)

It would appear that the current system of monitoring may encourage examiners to focus on procedural adherence at the expense of rating precision. There is a struggle between fulfilling an examiner's moral obligations for fair and accurate rating decisions, which bear serious consequences on a candidate's future, on the one hand; and protecting their self-interest in face of procedural monitoring with real implications for their own professional career.

Having to consistently do the timing while rating takes your mental energy away. About 15% is gone to timing but I'd like to give all of that to asking the right questions and giving appropriate scores. (E11)

You can be a mark out with rating and be within standard but being over time even slightly means an examiner is marked down. (E03)

You are constantly aware of the importance of doing it right and the knowledge that if you get three black marks in one test, then you'll have to go through a really tedious procedure before you can start working again. So, the system does encourage examiners to re-focus very much on procedure. You also get your ratings monitored, but I think that the monitoring for ratings is not as demanding as the monitoring for procedure. (E27)

The intention of their remarks was not to suggest making the monitoring of scores stricter, but to make the monitoring of timekeeping less strict, so that examiners can focus more on considering and giving accurate scores.

The current system of monitoring procedure seems to not only present a dilemma for examiners to focus on procedure vs. rating decisions. Within the role of interlocutor, examiners at times also find themselves having to focus on procedure at the expense of providing an assessment environment that accommodates test-takers' affective characteristics. After giving an account of the dilemma between rating quality and procedure compliance as discussed above, E23 gave an example of how providing a helpful environment to a candidate (allowing them more time to talk) comes at the cost of facing penalty in monitoring.

For example, an examiner trying to encourage nervous candidates to talk would prolong the time and go beyond the time limit. Your examiner trainer would not be interested in the conducive environment you are creating for the candidate, but focus on the fact that you are seconds late or ahead. (E23)

Sometimes candidates cry for 'a good parent' and they talk about deceased fathers and mothers. As an examiner, every situation is different, but you have a stopwatch. I sometimes encourage with facial expressions or I have held their hands despite my better judgment. I have given them some time to collect themselves and breaking some rules while doing so. I can be penalised for monitoring but surely not as a human being? (E02)

The latter quote by E02 aligns with the theme of the human aspect of the interviews discussed earlier. Although E02 believed that she might have broken rules by giving some time to and comforting distressed candidates, that was not actually the case. As presented earlier in Section 4.2.1, according to the guidelines for special circumstances in the *Instructions to Examiners*, when a candidate breaks down, giving them some time to recollect themselves is exactly what examiners are supposed to do. Yet, the dilemma remains that in doing so, examiners are choosing to lose a certain amount of language production from the candidate because they then have to move on to the next part of the test once candidates recover.

Monitoring in 'the spirit of the test'

Drawing on his experience both as an examiner and an examiner trainer, ET29 suggested examiner trainers move away from the current 'punitive' monitoring practices, and called for a re-focus on 'the spirit of the test' when evaluating examiner practices and adherence to procedure. While acknowledging that this is one individual's views and suggestions, we believe that it is worth quoting his comments and taking into consideration in informing future monitoring practices.

Examiner Trainers should be able to assess whether an examiner is conducting a test in the spirit of the test and timing very well as opposed to penalising examiners for being a few seconds off target, which has probably made absolutely no difference to the candidate's performance. That degree of [strictness] is counter-productive and leads to cynicism and stress. (ET29)

In the interview, ET29 suggested that this could be a minor adjustment in monitoring practice, for instance, making the lower limit of the number of 'black marks' in the monitoring a bit more lenient He stressed that ETs should use their professional judgement, and more importantly, give examiners the benefit of the doubt.

This section has presented a number of examiner voices that advocate relaxing the test administration procedures and monitoring related to them. Echoing the examiner comments on the interlocutor frame described in Section 4.4, it seems that the issues with the very strict monitoring in the current test were perceived by examiners as over-standardisation. As discussed in Section 4.4, we value the examiners' voices, while keeping in mind that the current practices have been put in place in response to the need for standardisation in the pre-2001 version of the IELTS Speaking Test. We maintain the position that any actions taken in the future in this regard should aim to strike an optimum balance between the need for standardising examiner behaviours and the need for providing candidates with an authentic environment to display their speaking ability (O'Sullivan and Lu, 2006;).

4.9 Test and test use

The final section of the online questionnaire consisted of questions on the perceived construct and the use of the IELTS Speaking Test. While beyond their immediate examining experience, we thought it would be useful to explore examiners' views on test use in more depth as they form one of the key stakeholder groups for the test.

The results of the online questionnaire showed that a strong majority of examiners (89.9% [Q53]) believed the IELTS Speaking Test to be a suitable tool for measuring candidates' general English speaking proficiency with the agreement rates dropping

significantly as the statement became more specific, i.e. for academic English (66.6% [Q54]) and professional registration (50.6% [Q55]).

A similar trend has been found in the questions regarding the speaking skills assessed in the test. The percentages of the examiners who *agreed* or *strongly agreed* decreased on questions with more academically-focused contexts:

- communicating with teachers and classmates in English-medium universities (76.1% [Q56])
- making oral presentations in English-medium universities (53.7% [Q57])
- participating in academic seminars in English-medium universities (54.0% [Q58]).

In the follow-up interviews, examiners' main reservations about the use of IELTS for academic or professional purposes related to the speaking demands and situations in the target language use domain, which they believed (at times drawing from their professional experiences), were not necessarily represented or elicited in the IELTS Speaking Test. In other words, they referred to the lack of evidence from the speaking test to make inferences about a candidate's skills for a given profession (e.g. law or medicine). They commented that IELTS measures general English.

> *I work at a university and the kinds of presentations our students give are very specialised whereas the questions in IELTS are geared towards a discussion, but not a formal presentation.* (ET08)

> *I compared this to TOEFL which is more academic. In a university setting you will not have the same tasks as in the test. Some would be effective in a classroom but not necessarily in an academic context.* (E11)

> *The speaking test tasks don't really look like a seminar. The 2-min may reflect presentation, but not really. Just a topic to talk about in two minutes. What type of oral communication are they trying to emulate in this test?* (E17)

> *For professional situations, well, I teach business English and in business you have totally different situations to deal with. In this test you don't really test professional skills and same for academic skills; when you study abroad you have to talk about specialised subjects and not wild animals.* (ET08)

> *You need to have the language for professional purposes. Example of myself doing MA in anthropology and a band 6 is definitely not enough. You can still learn later on, but the test may not say much about ability to deal with that kind of academic language.* (E15)

> *While general proficiency is important, the test should be tailored to that profession... like the OET [Occupational English Test (i.e. an English language test for healthcare professionals].* (E06)

The examiners who responded in the interviews also had experience as stakeholders in another group (i.e. university teacher, student, etc.), which sheds more light on what the test measures or does not measure. This is very much in line with the discussion by Nakatsuhara, Inoue, Khabbazbashi and Lam (2018) that the IELTS Speaking Test has indeed been developed as a general speaking test, and while it serves well for a test for entry to the academic and professional disciplines, further language training is needed and should be provided within the disciplines.

5. Suggestions for test improvement

Following the presentation and discussion of quantitative figures from the large-scale survey with 1203 examiners and qualitative interview comments provided by 36 examiners as well as selected written comments in the online survey responses, this section will briefly summarise our suggestions for improving the IELTS Speaking Test based on the examiner voices gathered in this study.

5.1 More flexible interlocutor frames

From the online questionnaire responses, it was clear that a vast majority of the 1203 examiners agreed with the current one-to-one, face-to-face format, and that each test part elicits useful language samples from the candidates. However, the interview data analysis revealed that many examiners craved increased flexibility in the interlocutor frame, so that they can facilitate candidates more smoothly and provide more support to elicit more language. Moreover, probing candidates especially in Part 3 (as discussed in Sections 4.1, 4.3 and 4.4) would bring in further benefits that can only be achieved in a direct, interactive test of speaking.

Our suggestions for the interlocutor frames are listed below.

- Part 1: Allow more flexibility in timing and the sequence in which the questions are asked.
- Part 1: Raise examiners' awareness that a) questions can be skipped (under some circumstances) and b) changing the verb tense is allowed when asking questions.
- Part 1: Allow using: *When was that?* / *Why was that?* / *Tell me more* in addition to *why/why not?*
- Part 1: Revise the first frame to be inclusive of candidates not in work or study.
- Part 2: Emphasise that rounding-off questions are not compulsory and that the round-off questions is provided on the task rubric; allow examiners the flexibility to formulate their own short comments/questions.
- Part 2: Make the wording of the instructions to candidates clearer to indicate that the expectation is for them to speak for two minutes, rather than one minute.
- Part 2: Allow more flexibility in timing for those weaker candidates who are unable to fill the two minutes.
- Part 2: Raise awareness in examiner training that stronger candidates who can provide sufficient and appropriate language sample for rating in less than two minutes can move on to the next part early.
- Part 3: Allow and train examiners to form their own follow-up questions more flexibly.
- Across parts: Allow examiners to respond/make short comments based on what candidate has said (although some guidance on the phrases permitted to use is necessary).
- Across parts: Allow more opportunities for authentic interaction including short but relevant comments to indicate engagement with candidate (e.g., *Thank you for telling me about X*.)

As noted a few times earlier, we are aware that, historically, the IELTS Speaking Test had much more flexible examiner scripts (and less structured tasks), and the 2001 Test Revision project aimed at standardising the test much more strictly in order to increase the test's reliability. There is indeed a trade-off between standardising and naturalness. As one of our examiner interviewees (E18) put it, 'with this exam, you have two competing forces: consistent and reliable versus authentic.' However, since 2001,

a number of discourse-based studies on the IELTS Speaking Test have suggested that the test is now biased for over-standardisation, and that the test needs to be revisited to strike a better balance between the need for standardisation and the need for offering candidates a comfortable, interactive environment in which they can display their face-to-face spoken communication ability (e.g., O'Sullivan and Lu, 2006; Nakatsuhara 2012; Seedhouse and Egbert, 2006; Seedhouse and Harris, 2011; Seedhouse, Harris, Naeb, and Ustunel, 2014; Seedhouse and Morales, 2017; Seedhouse and Nakatsuhara, 2018). Therefore, putting back a certain degree of flexibility to the interlocutor frame would allow the test to be more authentic and less mechanical, as well as more capable of eliciting appropriate language samples from candidates, making fuller use of the advantages of a face-to-face speaking test.

5.2 Wider range and choice of topics

As discussed earlier in Section 4.2, we found that more than one in four examiners had doubts about the appropriateness of some of the topics in the test. While the expectation may be that examiners should be able to avoid unsuitable topics, we note the caveat that the requirement to 'vary the topics' specified in the *Instructions to Examiners* may obstruct examiners from doing so. The interviews identified the topics that may not be suitable for certain age, gender or background, based on which we put our suggestions together.

- Introduce a wider variety of topics and themes that are inclusive of candidates from different socio-economic backgrounds.
- Introduce choice of topics for candidates (for example in Part 2).
- Raise examiner awareness regarding choice of topics.
- Consider allowing the shifting of topics between Parts 2 and 3 where necessary.
- Make more use of the feedback form for test centres to communicate issues with the live test materials for speedy removal/modification of test materials as necessary.

Widening the topic pool for the test would require revisiting the current topics as well as careful development of potential new ones, in terms of not only the candidate backgrounds but also the capacity to challenge and probe candidates and elicit comparable language samples to other topics (through Parts 2 and 3).

5.3 Further guidelines and materials for examiners

5.3.1 IELTS Speaking Test: Instructions to Examiners

On the online questionnaire, a vast majority (85.4%) of the examiner respondents found the examiner handbook (*IELTS Speaking Test: Instructions to Examiners*) helpful, but almost one in three examiners felt that there could be other areas that it could provide further guidelines. Based on the findings in Section 4.5, our suggestions are:

- raise examiner awareness of the availability of self-access sample performances via test centres
- raise examiner awareness regarding the existence of guidelines for special circumstances.

5.3.2 Test administration

The majority of the examiners found the overall length of the test appropriate, and the general administration (i.e. delivery) of the tests manageable. In the interviews, it became clear that new examiners want more practice and training managing the dual role of being an interlocutor and assessor (Section 4.6). Related to that was the need for enhancing the reliability in rating, which was closely related to the rating (Section 4.7)

and training and standardisation (Section 4.8) of the test. Our suggestions are listed together in Sections 5.3.3 and 5.3.4.

5.3.3 Rating

In the interviews, it was suggested that it would be helpful for IELTS Partners to:

- consider developing descriptors regarding the relevance of responses.

Moreover, almost half of the examiner respondents found the Pronunciation scale difficult to apply due to the lack of unique descriptors in Bands 3, 5 and 7. It is assumed that developing fine-grained pronunciation descriptors was difficult due to the lack of research into pronunciation features when the decision was made not to provide any descriptors in those 'in-between' levels. However, recent advances in pronunciation research, particularly Issacs et al.'s (2015) findings from the discriminant analyses and ANOVAs of examiners' judgements of various pronunciation features, can offer a useful base to design level-specific descriptors in those 'in-between' bands (e.g., clear distinctions between Bands 6 and 7 for comprehensibility, vowel and consonant errors, word stress, intonation and speech chunking). There is a glossary in the *Instructions to Examiners* that define the terminology used in the scale descriptors, but we also suggest adding more illustrative audio/video samples to the examiner training resources in order to enhance examiners' understanding of different pronunciation and prosodic features.

The follow-up interviews also identified a number of issues with various other aspects of rating, most of which, we believe, could be better addressed with increasing the size and availability of benchmarked samples. The specific suggestions are listed in Section 5.3.4.

5.3.4 Training and standardisation

Although the majority of the examiners held positive views about the current training and standardisation of the IELTS Speaking Test, they also pointed out a number of areas that could enhance the test reliability and improve examiner performance. Below are our suggestions:

- raise examiners' awareness of the availability of self-access training materials
- collect and make available self-access materials with different more L1 varieties
- use video recordings for both certification and re-certification
- extend length of training time and provide more opportunities for practice both with mock candidates and with peers, especially for new examiners
- provide feedback on the scores more often
- review aspects of monitoring that are considered too rigid, particularly the timings
- introduce double-marking using video-recordings if video-conferencing mode of IELTS Speaking is introduced in the future

5.4 Test and test use

Examiners, on the questionnaire and in the interviews, echoed the common criticisms that the scores on the IELTS Speaking Test, which is a general speaking test, do not necessarily indicate that one can cope well with the linguistic demands of academic or professional disciplines (Murray, 2016). However, it should be noted that the IELTS Speaking Test has never claimed itself to be an 'academic' or 'professional' speaking test; it has always been a general English speaking test. Over the years, IELTS has come to be used for various purposes, including professional registration and immigration, which may not have been the primary purpose of the test when it was first developed. Some may argue that IELTS Speaking must be redesigned to claim its fitness for

particular purposes. However, according to Murray (2016: 106), despite it being a 'blunt' instrument due to the discrepancies between the test construct and the contexts of test use, 'generally speaking it does a reasonably good job under the circumstances'. Murray further emphasises that, although the idea of candidates taking English language tests based on and tailored to the discipline area in which they intend to operate might appear a logical option, in practice it makes little sense. This is because: a) we cannot assume that candidates will come equipped with adequate conversancy in the literacy practices of their future disciplines, as a result of diverse educational experiences; and b) candidates need to be trained in those literacy practices anyway, after entry to higher education or professional courses. The views of examiner interviewees in this study on the test use, particularly in the context of university entry (as discussed in Section 4.9), are indeed in line with the role that the IELTS Partners envisaged when designing the IELTS Test (Taylor, 2007). Taylor (2012, p. 383) points out that 'IELTS is designed principally to test readiness to enter the world of university-level study in the English language' and assumes that the skills and conventions necessary for the specific disciplines are something that candidates will learn during the course of their study.

Enhancing the understanding and appropriate use and interpretation of the test scores falls within the realm of enhancing language assessment literacy among stakeholders. The British Council, as communicated to the research team, has a dedicated team which conducts visits to various UK universities and presents to relevant personnel, including admission officers, what IELTS scores does and does not tell them. This is an extremely important area to invest in to ensure that score users, especially decision-makers, do not over-interpret test scores. Given that the IELTS Partners have already invested heavily in this area, it may perhaps be useful to look into the effectiveness of such assessment literacy enhancement that have been conducted. Existing data and records could be collated regarding the audience (i.e. stakeholder groups), as well as the types and amount of information presented. Furthermore, follow-up interviews could be conducted with the stakeholder groups in order to know whether the provided information has been understood, taken up and acted upon (e.g., enhancing the post-entry provision of support given the scope of IELTS test score interpretation).

Conducting this type of follow-up studies or audits would be beneficial in finding out what has or has not worked well, what factors might hinder the appropriate understanding and use of test scores, and what more could be done to improve the current practice.

6. Final remarks and acknowledgements

Gathering the voices of 1203 IELTS Speaking examiners on an online questionnaire and further exploring the voices of 36 selected examiners on individual interviews, this study has offered an in-depth analysis of examiners' perceptions and experiences of various aspects of the current IELTS Speaking Test and how the test could be improved. Examiners were generally positive about the current IELTS Speaking Test, but they also enthusiastically shared their views on various features of the test that can be improved in the future. We believe that the results and suggestions from this research will offer valuable insights into possible avenues that the IELTS Speaking Test can take to enhance its validity and accessibility in the coming years.

Finally, we would like to express our sincere gratitude to the following people.

- Ms Mina Patel (Assessment Research Manager, the British Council), who facilitated the execution of this project in every aspect, without whom it was not possible to complete this research.
- Professor Barry O'Sullivan (Head of Assessment Research & Development, the British Council), who reviewed our questionnaire and made valuable suggestions.
- Three IELTS Speaking examiners who generously shared their views in the focus group discussion prior to the development of the online questionnaire.
- The 1203 IELTS Speaking examiners who responded to our questionnaire and 36 examiners who further participated in telephone or video-conferencing interviews to elaborate on their views.

The process of gathering and analysing IELTS Speaking examiners' insights was truly valuable to us, not only as the researchers of this project, but as individual language testing researchers. Throughout all the stages of this project, we were overwhelmed by the enthusiasm of the IELTS Speaking examiners who genuinely wish to maintain and contribute to enhancing the quality of the IELTS Speaking Test and to offer a better examination experience for candidates. It is our sincere hope that this project has done justice to the IELTS Speaking examiners' hard work and has contributed to delivering their professional and committed voices to the IELTS Partners and IELTS test users all over the world.

References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council of Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. AERA, Washington: DC.

Berry, V., Nakatsuhara, F., Inoue, C., & Galaczi, E. (2018). Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery (Phase 3). *IELTS Partnership Research Papers, 2018/1*. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Retrieved from: https://www.ielts.org/teaching-and-research/research-reports

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

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey. (eds.). *IELTS collected papers: Research in speaking and writing assessment* (pp. 98–141). Cambridge: Cambridge University Press.

Brown, A., & Hill, K. (2007). Interviewer style and candidate performance in the IELTS oral interview. In L. Taylor & P. Falvey. (eds.). *IELTS collected papers: Research in speaking and writing assessment* (pp. 37–62). Cambridge: Cambridge University Press.

Davies, A. (2008). *Assessing academic English: Testing English proficiency 1950–1989: the IELTS solution.* Cambridge: Cambridge University Press.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), pp. 423–443.

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown, & K. Hill (Eds.), *Tasks and criteria in performance assessment* (pp. 43–74). Frankfurt: Peter Lang.

Galaczi, E., Lim, G., & Khabbazbashi, N. (2012). *Descriptor salience and clarity in rating scale development and evaluation.* Paper presented at the Language Testing Forum.

Harsch, C. (2019). English varieties and targets for L2 assessment. In C. Hall & R. Wicaksono (eds.) *Ontologies of English: Conceptualising the language for learning, teaching, and assessment.* Cambridge: Cambridge University Press.

Hughes, A., Porter, D., & Weir, C. J. (1998). *ELTS validation project: Proceeding of a conference held to consider the ELTS Project report*. British Council and UCLES.

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS Pronunciation scale, *IELTS Research Reports Online Series, 2015/4*, pp. 1–48. British Council, Cambridge Assessment English and IDP: IELTS Australia.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, 34(1), pp. 23–48.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral tests*. Cambridge: Cambridge University Press.

May, Lyn. (2011). Interaction in a paired speaking test: the rater's perspective. *Language Testing and Evaluation*, 24. Peter Lang, Frankfurt, Germany.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow, Essex: Longman.

Merrylees, B. & McDowell, C. (2007). A survey of examiner attitudes and behaviour in the IELTS oral interview. In L. Taylor & P. Falvey. (eds.), *IELTS collected papers, 2: Research in speaking and writing assessment* (pp. 142–184). Cambridge, UK: Cambridge University Press.

Murray, N. (2016). *Standards of English in higher education: Issues, challenges and strategies.* Cambridge: Cambridge University Press.

Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking Test. In L. Taylor, & C. J. Weir (eds.). *IELTS Collected Papers 2: Research in reading and listening assessment* (pp. 519–573). Studies in Language Testing 34. Cambridge: Cambridge University Press.

Nakatsuhara, F. (2018). Rational design: The development of the IELTS Speaking test. In P. Seedhouse, & F. Nakatsuhara (2018). *The discourse of the IELTS Speaking Test: The institutional design of spoken interaction for language assessment* (pp. 17–44). Cambridge: Cambridge University Press.

Nakatsuhara, F., Inoue, C. Berry, V. and Galaczi, E. (2016). Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery: A preliminary comparison of test-taker and examiner behaviour. *IELTS Partnership Research Papers*, 1, pp. 1–67. British Council, Cambridge Assessment English and IDP: IELTS Australia. Available online at: https://www.ielts.org/-/media/research-reports/ielts-partnership-research-paper-1.ashx

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017a). Exploring the Use of Video-Conferencing Technology in the Assessment of Spoken Language: A Mixed-Methods Study, *Language Assessment Quarterly,* 14(1), 1–18.

Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2017b). Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and video-conferencing delivery (Phase 2), *IELTS Partnership Research Papers, 3*, pp. 1–74. British Council, Cambridge Assessment English and IDP: IELTS Australia. Available online at: https://www.ielts.org/-/media/research-reports/ielts-research-partner-paper-3.ashx

Nakatsuhara, F., Inoue, C. & Taylor, L. (2017). An investigation into double-marking methods: Comparing live, audio and video rating of performance on the IELTS Speaking Test, I*ELTS Research Reports Online Series, 1,* pp. 1–49. British Council, Cambridge Assessment English and IDP: IELTS Australia. Available online at: https://www.ielts.org/-/media/research-reports/ielts_online_rr_2017-1.ashx

Nakatsuhara, F., Taylor, L., & Jaiyote, S. (2019). The role of the L1 in testing L2 English. In C. Hall & R. Wicaksono (eds.), *Ontologies of English: Conceptualising the language for learning, teaching, and assessment*. Cambridge: Cambridge University Press.

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

Sato, T. (2014). *Linguistic laypersons' perspective on second language oral communication ability.* Unpublished PhD thesis. University of Melbourne.

Seedhouse, P. (2018). The interactional organisation of the IELTS Speaking test. In P. Seedhouse, & F. Nakatsuhara (2018). *The discourse of the IELTS Speaking Test: The institutional design of spoken interaction for language assessment* (pp. 80IELTS Australia and British Council113). Cambridge: Cambridge University Press.

Seedhouse, P., & Egbert, M. (2006). The Interactional Organisation of the IELTS Speaking Test. *IELTS Research Reports, Vol 6,* pp. 161–206. IELTS Australia and British Council.

Seedhouse, P., & Harris, A. (2011). Topic Development in the IELTS Speaking Test. *IELTS Research Reports, Vol 12.* IDP: IELTS Australia and British Council

Seedhouse, P., & Morales, S. (2017). Candidates questioning examiners in the IELTS Speaking Test: An intervention study. *IELTS Research Reports Online Series, 5*. British Council, Cambridge Assessment English and IDP: IELTS Australia. Retrieved from: https://www.ielts.org/teaching-and-research/research-reports

Seedhouse, P., & Nakatsuhara, F. (2018). *The Discourse of the IELTS Speaking Test: The Institutional Design of Spoken Interaction for Language Assessment*. Cambridge: Cambridge University Press.

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports Online Series, 2,* pp. 1–30. British Council, Cambridge Assessment English and IDP: IELTS Australia.

Taylor, L. (2006). The changing landscape of English: implications for English language assessment. *ELT Journal,* 60(1), pp. 51–60.

Taylor, L. (2007). The impact of the joint-funded research studies on the IELTS speaking module. In L. Taylor & P. Falvey (eds.) *IELTS collected papers: research in speaking and writing assessment* (pp. 185–196). Cambridge: Cambridge University Press.

Taylor, L., & Falvey, P. (eds.) (2007). *IELTS collected papers: research in speaking and writing assessment*. Cambridge: Cambridge University Press.

Yates, L., Zielinski, B., & Pryor, E. (2011). The Assessment of Pronunciation and the New IELTS Pronunciation Scale. *IELTS Research Reports, Vol 12*. IDP: IELTS Australia and British Council.

Note: The *Instructions to IELTS Examiners* (2011) does not appear in the reference list here as it is confidential and not publicly available.

Appendix 1: Online questionnaire with descriptive statistics for closed questions

Note. Not all respondents answered all the questions. Unless specified, the percentages are calculated based on valid responses against (up to) the total of 1203 cases.

IELTS Speaking Examiner Survey

Thank you for agreeing to participate in this survey. The aim of this survey is to gather voices from IELTS Speaking examiners and examiner trainers on various aspects of the current test and what changes they would like to see. Your insights will offer the IELTS Partners a range of possibilities and recommendations for a potential revision of the IELTS Speaking Test to further enhance its validity and accessibility in the coming years.

Form of Consent

Principal investigator: Dr Chihiro Inoue (CRELLA, University of Bedfordshire)
chihiro.inoue@beds.ac.uk

Co-investigators: Dr Fumiyo Nakatsuhara and Dr Daniel Lam (CRELLA, University of Bedfordshire)

Please note:

- All personal data collected and processed for this research will be kept strictly confidential. We will not disclose any personal data to a third party nor make unauthorised copies.
- All citations from the data used in published works or presentations will be done so anonymously.
- Written comments may be used for any reasonable academic purposes including training, but with anonymity for all participants.

Declaration:

I grant to investigators of this project permission to record my responses.

I agree to my responses to be used for this research. I understand that anonymised extracts may be used in publications, and I give my consent to this use.

I understand that all data collected and processed for this project will be used for any reasonable academic purposes including training, and I give my consent to this use.

I declare that:

- I am 18 years of age or older;
- All information I provide will be full and correct; and
- I give this consent freely.

If you agree, please tick this box: ☐

1. BACKGROUND DATA

Years of experience as an EFL/ESL Teacher? M = 18.9 years; SD = 10.04 years

Years of experience as an IELTS Speaking Examiner? M = 7.49 years; SD = 5.68 years

Are you currently an IELTS Speaking Examiner Trainer? Yes/No

If yes, for how long? M = 6.3 years; SD= 5.1 years

Region where you currently examine/ train examiners as an IELTS Examiner/ Examiner Trainer?
Europe 35%; Middle East & North Africa 16%; East Asia 14%; Northern America 13%; South Asia 8%; Southeast Asia 6%; Africa 3%; Latin America 3%; Australia & New Zealand 1%; Russia & Central Asia 1%

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

1. Tasks

Part 1 – Interview						
Q1	I find the language sample elicited… to inform my rating decision.	1. Never useful 0.4%	2. Seldom useful 6.4%	3. Sometimes useful 32.2%	4. Often useful 42.2%	5. Always useful 18.8%
Q2	I find the length of Part 1…	1. Too short 0.5%	2. A bit too short 7.5%	3. Appropriate 80.6%	4. A bit too long 10.5%	5. Too long 0.9%
Part 2 – Individual long turn						
Q3	I find the language sample elicited… to inform my rating decision.	1. Never useful 0.3%	2. Seldom useful 2.7%	3. Sometimes useful 9.9%	4. Often useful 38.0%	5. Always useful 49.2%
Q4	I find the length of Part 2…	1. Too short 0.6%	2. A bit too short 6.7%	3. Appropriate 82.6%	4. A bit too long 9.7%	5. Too long 0.4%
Part 3 – Two-way discussion						
Q5	I find the language sample elicited… to inform my rating decision.	1. Never useful 0.3%	2. Seldom useful 0.9%	3. Sometimes useful 6.8%	4. Often useful 24.0%	5. Always useful 68.1%
Q6	I find the length of Part 3…	1. Too short 2.0%	2. A bit too short 19.6%	3. Appropriate 72.1%	4. A bit too long 6.1%	5. Too long 0.3%
Considering all three parts together…						
Q7	The number of test tasks is…	1. Too few 0.9%	2. 2.0%	3. Appropriate 91.4%	4. 3.8%	5. Too many 2.0%
Q8	The sequencing of the three parts is appropriate.	1. Strongly disagree 0.9%	2. Disagree 1.6%	3. Neutral 12.7%	4. Agree 53.5%	5. Strongly agree 31.2%
Q9	The range of task types in the current version of the IELTS Speaking Test is …	1. Too narrow 2.0%	2. A bit narrow 22.5%	3. Appropriate 71.9%	4. A bit wide 3.4%	5. Too wide 0.2%

Q9a	(If the answer to Q9 is 'too narrow' / 'a bit narrow') Which of the following new task type(s) would you like to be included in a revised version of the IELTS Speaking Test? • Picture description 9.8% • Asking questions to the examiner 8.6% • Role play 4.0% • Problem-solving 9.7% • Decision-making 10.0% • Information gap 2.8% • Presentation 3.6% • Free discussion 12.3% • Summarise a reading text 4.9% • Summarise a listening text 3.5% • Other (Please specify: _____) 3.5%
Q10	[optional] Please elaborate on any of your answers to Q1 to Q9.

2. Topics

Q	Statement	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
Q11	Overall, the topics in the test tasks are appropriate.	1.0%	14.7%	22.6%	54.5%	7.2%
Q12	The topics are appropriate for candidates of either gender.	0.8%	13.3%	18.2%	55.7%	12.0%
Q13	The topics are appropriate for candidates of different cultural backgrounds.	2.8%	25.6%	23.3%	40.6%	7.7%
Q14	The range of topics (task versions) which examiners can choose from in the Booklet is adequate.	3.7%	8.5%	13.3%	54.7%	19.8%
Q15	The connection in topic between Part 2 and Part 3 is a positive feature.	1.6%	5.2%	13.7%	48.8%	30.7%
Q16	In Part 3, examiners should be given the choice to change to another topic different from the one in Part 2.	10.9%	29.9%	22.7%	29.0%	7.5%
Q17	[Optional] Please elaborate on any of your answers to Q11 to Q16.					

3. Format

Q	Statement	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
Q18	The 1-to-1 interview format should be kept in the IELTS Speaking Test.	1.5%	3.1%	7.8%	28.6%	59.0%
	If the answer to Q18 is disagree / strongly disagree, please tick how you feel the format should change: • The IELTS Speaking Test should be in a paired format (2 candidates). 2.0% • The IELTS Speaking Test should be in a group format (e.g. 3 - 4 candidates). 0.2% • Other [please specify] 2.6%					
Q19	The face-to-face examiner-candidate interaction mode used in the current test is a suitable delivery for the test, as compared to a computer-delivered mode (speaking to a computer rather than a person (e.g. TOEFL iBT)).	1.0%	0.4%	3.6%	12.3%	82.7%

4. Interlocutor frame

Q	Flexibility/rigidity of interlocutor frame	1. too rigid	2. a bit too rigid	3. appropriate	4. a bit too flexible	5. too flexible
Q20	The interlocutor frame for Part 1 is…	15.0%	47.1%	37.6%	0.3%	0.0%
Q21	The interlocutor frame for Part 2 is…	5.5%	22.1%	72.1%	0.3%	0.0%
Q22	The interlocutor frame for Part 3 is…	2.5%	13.9%	80.3%	2.3%	1.0%
Q23	How often do you ask your own follow-up questions in Part 3?	1. Never 1.5%	2. Seldom 2.8%	3. Sometimes 15.9%	4. Frequently 34.0%	5. Always 45.7%

Q24	What potential changes to the interlocutor frame do you think might be beneficial? Please tick all that apply.	
	An optional extra question in Part 1 frames should be provided.	37.4%
	There should be an optional extra topic in Part 1 in case the candidate completes the first two topics quickly.	50.0%
	In Part 1 frames, there should be the option to ask the candidate 'tell me more' instead of 'why/why not'.	83.4%
	After the candidate finishes speaking in the individual long turn (Part 2), there should be no round-off questions.	34.2%
	After the candidate finishes speaking in the individual long turn (Part 2), there should be a third round-off question (in addition to the existing one to two round-off questions).	11.4%
	Other [please specify]	24.6%

5. IELTS Speaking Test: Instructions to Examiners

Q	Statement	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
Q25	The *Instructions to Examiners* are helpful for administering the test.	0.8%	2.4%	11.4%	56.0%	29.4%
Q26	The *Instructions to Examiners* cover all the necessary guidelines and questions I have about administering the test.	1.5%	13.1%	17.2%	50.4%	17.8%
Q27	[Optional] Please elaborate on any of your answers to Q25 to Q26.					

6. Administration of the test

Q	Statement	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
Q28	The overall length of the test is…	Too short 0.3%	A bit short 7.3%	Appropriate 86.8%	A bit long 5.6%	Too long 0.0%
Q29	The task of keeping time for each part of the test is manageable.	2.7%	13.0%	16.8%	51.2%	16.3%
Q30	The examiner's dual role of being the interviewer and the rater is easy to manage.	3.1%	14.2%	16.6%	46.6%	19.5%
Q31	It is easy to adhere to the guideline of administering test sessions for no more than 8 hours a day.	3.4%	9.3%	21.1%	42.6%	23.6%
Q32	It is easy to adhere to the guideline of taking a break at least once per 6 test sessions.	3.3%	10.2%	16.0%	46.7%	23.7%
Q33	It is easy to adhere to the guideline of conducting no more than 3 test sessions per hour.	3.9%	10.8%	16.0%	45.2%	24.1%
Q34	[Optional] Please elaborate on any of your answers to Q28 to Q33.					

7. Rating

Q	Statement	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
Q35	I find the descriptors in Fluency and Coherence easy to apply.	0.3%	10.3%	14.6%	57.8%	17.0%
Q36	I find the descriptors in Grammatical Range & Accuracy easy to apply.	0.4%	7.6%	13.5%	58.6%	19.9%
Q37	I find the descriptors in Lexical Resource easy to apply.	0.3%	6.5%	13.3%	59.7%	20.3%
Q38	I find the descriptors in Pronunciation easy to apply.	3.5%	20.5%	22.9%	42.1%	11.0%
Q39	I feel the number of bands (currently 9 bands) in the IELTS Speaking Test is adequate.	1.1%	4.0%	10.3%	52.5%	32.2%
Q40	The current IELTS Speaking Test measures higher band levels accurately (i.e. Bands 8.0–9.0)	1.4%	9.5%	19.5%	51.6%	18.0%
Q41	The current IELTS Speaking Test measures middle band levels accurately (i.e. Bands 5.0–7.5)	0.5%	5.4%	16.4%	56.8%	21.0%
Q42	The current IELTS Speaking Test measures lower band levels accurately (i.e. Bands 1.0–4.5)	2.0%	8.6%	22.1%	49.1%	18.2%
Q43	The use of audio recordings for second-marking is appropriate.	0.7%	2.3%	12.2%	53.3%	31.5%
Q44	[Examiner Trainers only] The use of audio recordings for monitoring is appropriate. (n=182)	0.0%	2.2%	13.2%	48.4%	36.3%
Q45	How often do you refer to the assessment criteria etc. in the *Instructions to Examiners* at the start of an examining day?	Never 2.5%	Seldom 11.7%	Some times 27.2%	Frequently 25.5%	Always 33.1%
Q46	[optional] Please elaborate on any of your answers to Q35 to Q45.					

8. Training and standardisation

Q47	Please indicate your main role concerning examiner training and standardisation.	A new Examiner 11.5%	An experienced Examiner 79.7%	An Examiner Trainer 7.4%	An Examiner Support Coordinator 0.4%

[New Examiners (n=136)]

Q48a	The length of the New Examiner Training is…	1.too short 12.5%	2. a bit too short 35.3%	3. appropriate 47.1%	4. a bit too long 4.4%	5. too long 0.7%
Q49a	The number of benchmark samples and standardisation samples covered in the New Examiner Training is…	1. too small 9.6%	2. a bit too small 37.5%	3. appropriate 48.5%	4. a bit too many 4.4%	5. too many 0.0%
Q50a	I find the materials used in the New Examiner Training useful.	1. Strongly disagree 2.2%	2. Disagree 8.8%	3. Neutral 19.9%	4. Agree 55.1%	5. Strongly agree 14.0%
Q51a	I am happy with the use of video recordings for training and audio recordings for certification.	1. Strongly disagree 0.7%	2. Disagree 5.9%	3. Neutral 14.7%	4. Agree 58.8%	5. Strongly agree 19.9%

[Experienced Examiners (n=876)]

Q48b	The length of the Examiner Standardisation is…	1.too short 2.1%	2. a bit too short 8.8%	3. appropriate 70.7%	4. a bit too long 14.3%	5. too long 4.2%
Q49b	The number of benchmark samples and standardisation samples covered in the Examiner Standardisation is…	1. too small 1.6%	2. a bit too small 14.2%	3. appropriate 72.4%	4. a bit too many 9.5%	5. too many 2.4%
Q50b	I find the materials used in the Examiner Standardisation useful.	1. Strongly disagree 0.9%	2. Disagree 4.5%	3. Neutral 20.4%	4. Agree 59.7%	5. Strongly agree 14.5%
Q51b	I am happy with the use of video recordings for training and audio recordings for re-certification.	1. Strongly disagree 1.4%	2. Disagree 5.3%	3. Neutral 15.5%	4. Agree 59.7%	5. Strongly agree 18.2%

[Examiner Trainers (n=80)]

Q48b	The length of the Examiner Standardisation is…	1.too short 2.5%	2. a bit too short 31.3%	3. appropriate 61.3%	4. a bit too long 5.0%	5. too long 0.0%
Q49b	The number of benchmark samples and standardisation samples covered in the Examiner Standardisation is…	1. too small 1.3%	2. a bit too small 20.0%	3. appropriate 75.0%	4. a bit too many 3.8%	5. too many 0.0%
Q50b	I find the materials used in the Examiner Standardisation useful.	1. Strongly disagree 0.0%	2. Disagree 6.3%	3. Neutral 10.0%	4. Agree 63.8%	5. Strongly agree 20.0%
Q51b	I am happy with the use of video recordings for training and audio recordings for re-certification.	1. Strongly disagree 1.3%	2. Disagree 8.8%	3. Neutral 17.5%	4. Agree 52.5%	5. Strongly agree 20.0%

[Examiner Support Coordinators (n=4)]

Q48b	The length of the Examiner Standardisation is…	1.too short 25.0%	2. a bit too short 0.0%	3. appropriate 75.0%	4. a bit too long 0.0%	5. too long 0.0%
Q49b	The number of benchmark samples and standardisation samples covered in the Examiner Standardisation is…	1. too small 0.0%	2. a bit too small 0.0%	3. appropriate 100.0%	4. a bit too many 0.0%	5. too many 0.0%
Q50b	I find the materials used in the Examiner Standardisation useful.	1. Strongly disagree 0.0%	2. Disagree 0.0%	3. Neutral 25.0%	4. Agree 75.0%	5. Strongly agree 0.0%
Q51b	I am happy with the use of video recordings for training and audio recordings for re-certification.	1. Strongly disagree 0.0%	2. Disagree 25.0%	3. Neutral 25.0%	4. Agree 25.0%	5. Strongly agree 25.0%
Q52	[Optional] Please elaborate on any of your answers to Q48 to Q51.					

9. Test and test use

Q	Statement	1. Strongly disagree	2. Disagree	3. Neutral	4. Agree	5. Strongly agree
Q53	The IELTS Speaking Test is a suitable tool for measuring candidates' general English speaking proficiency.	0.4%	2.7%	7.0%	52.4%	37.5%
Q54	The IELTS Speaking Test is a suitable tool for measuring candidates' Academic English speaking proficiency.	1.5%	12.4%	19.6%	45.1%	21.5%
Q55	The IELTS Speaking Test is a suitable tool for measuring candidates' English proficiency appropriate for professional registration (e.g., medical professionals; legal professionals).	2.6%	16.7%	30.1%	36.3%	14.3%
Q56	The IELTS Speaking Test assesses appropriate speaking skills necessary for communicating with teachers and classmates in English-medium universities.	1.2%	7.2%	15.6%	51.0%	25.1%
Q57	The IELTS Speaking Test assesses appropriate speaking skills necessary for making oral presentations in English-medium universities.	2.7%	17.7%	25.8%	40.9%	12.8%
Q58	The IELTS Speaking Test elicits appropriate speaking skills necessary for participating in academic seminars in English-medium universities.	3.1%	18.1%	25.8%	38.9%	14.1%
Q59	[Optional] Please elaborate on any of your answers to Q53 to Q58 here.					

Thank you very much for your responses.

[Optional] As a follow-up stage to this Survey, we are looking for Examiners and Examiner Trainers who are willing to share their views further via Skype or telephone. If you are happy to be contacted by us, please leave your name and contact details. Your identity and contact details will be known only to the three investigators of this research at the University of Bedfordshire, UK, and will NOT be shared with any of the IELTS Partners.

Name: ……..

Email address:……

Skype ID:……

Telephone no.(country code)……. (tel no.)……..

Appendix 2: Sample interview questions

1. In the survey, you mentioned finding the language samples elicited in the three parts often useful in circumstances when candidates have little to say. In what ways do you think these can be improved to be always useful?

2. You chose 'disagree' for the sequencing of the three parts being appropriate with Part 1 questions sometimes more abstract than Part 2; can you give me a few examples of this?

3. In the survey, you mentioned that the Interlocutor Frames for Parts 1 and 2 are 'a bit too rigid' but you were happy with Part 3; can you tell me in what ways you would like to see the frames for the first two parts improved?

4. You expressed a preference for varying the different round-off questions in Part 2; do you think these should be pre-scripted or would you like some flexibility in formulating these questions? Please expand.

5. You had selected 'disagree' in your responses regarding appropriateness of topics in particular in terms of cultural background and mentioned 'hats or boats' as not necessarily appropriate in the [examiner's area] context.
   a. Can you expand a bit on these examples?
   b. What typically happens in terms of candidate performance when facing these topics?
   c. And how do you deal with such problems as an examiner?
   d. In what ways do you think topic-related problems can be solved?
6. You believe that examiners should not be given a choice to switch topics from Part 2 to Part 3; can you elaborate on your reasons for this?

7. You mentioned wanting to see best practices from different centres; what sorts of areas in particular are you interested in? What would you like more guidance on?

8. You had selected 'disagree' for the descriptors related to Fluency and Coherence being easy to apply. You mentioned that the two relate to two very different criteria; can you elaborate a bit on this?

9. You said that the examiner standardisation is perhaps 'a bit too short' and samples too small. Is this about quantity or quality or both? In what ways can they be improved?

10. Lastly, in terms of test uses of IELTS for different purposes; you selected 'disagree' for the use of IELTS for academic purposes or professional registration. Can you elaborate on your views on this?

Appendix 3: Invitation to interview

Dear Colleague,

We are researchers from CRELLA (Centre for Research in English Language Learning and Assessment www.beds.ac.uk/crella) at the University of Bedfordshire and part of the team working with the IELTS Partners on the IELTS examiner survey that you kindly participated in recently. Thank you very much for sharing your valuable insights with us and for agreeing to be contacted for a follow-up interview.

If you are still happy and available to participate in an interview, please let us know by return email We will then arrange for an interview date/time that is convenient to you in March or April. We anticipate the interview to take approximately 30–40 minutes.

We are planning to use Skype, FaceTime, Google Hangout, or IMO for the interviews. The interviews will be, with your permission, audio-recorded for transcription and thematic analysis. Note that all responses will be anonymised and your details will not be shared with the IELTS partners. On completing the interview, there will be a small token of gratitude from us.

We look forward to hearing back from you.