ISSN 2201-2982
2017/5

IELTS Research Reports
Online Series

Candidates questioning examiners in the IELTS Speaking Test: An intervention study

Paul Seedhouse and Sandra Morales

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

Candidates questioning examiners in the IELTS Speaking Test: An intervention study

This study evaluated the effect of the addition of a fourth part into the structure of the IELTS Speaking Test (IST), intended as a two-minute section in which the candidate asked questions on a typical IST topic to the examiner, who then replied. The part adds value in a number of ways, creating more naturalistic, two-way interaction and useful extra information for rating purposes, while potential disadvantages are increased test duration and variation in amount and type of examiner talk.

Acknowledgements

The authors would like to thank Cambridge English Language Assessment for supplying relevant materials, the three IELTS examiners and 18 candidates who participated in the study, and CA Transcription Services for transcription work.

Funding

This research was funded by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Grant awarded 2015.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2017.

How to cite this article

Seedhouse P. and Morales S. 2017. Candidates questioning examiners in the IELTS Speaking Test: An intervention study. *IELTS Research Reports Online Series, No. 5*. British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Available at: https://www.ielts.org/teaching-and-research/research-reports

Introduction

This study by Paul Seedhouse and Sandra Morales of Newcastle University was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia, and Cambridge English Language Assessment) 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 English Language Assessment, 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 110 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 (http://www.cambridgeenglish.org/silt), and in *IELTS Research Reports*. Since 2012, in order 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 principal investigator in this study has completed a number of IELTS joint-funded research projects. On more than one occasion (Seedhouse and Egbert, 2006; Seedhouse and Harris, 2011), he has mooted the possibility of changes being made to the IELTS Speaking test so that a broader range of interaction types might be observed. In particular, the idea is for an interaction that is more candidate led rather than examiner led, and that also tests their ability to form questions.

A section like this was actually part of the IELTS Speaking Test prior to 2001, and at the time, the observation was that it "failed to elicit anything more than a perfunctory reverse question-answer scenario and thus did not provide the richer sample of candidate performance that was being sought" (Taylor, 2011, xii). That being said, IELTS has grown and changed quite a bit since then, so it might be opportune to revisit the question.

In the current study, the researchers consider two possible task types: one similar to that in the pre-2001 IELTS Speaking test, where the candidate and examiner work off of a cue card with bullet points to address, and another they call "examiner leading statement", where in response to the statement a candidate asks questions and leads the development of the conversation.

While the study involved only a small number of participants, the results are nonetheless promising. Whichever the task, it was shown that a broader range of discourse moves were in evidence, and that there were also distinct differences in the performance of stronger and weaker candidates. Thus, there is prima facie a case for further exploring this possibility.

References:

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*, pp. 69–124. IDP: IELTS Australia and British Council.

Taylor, L. (2011). Introduction. *IELTS Research Reports, Vol 12*. IDP: IELTS Australia and British Council.

The study raised a number of issues that require further consideration. Foremost among them is that more genuine interaction by definition means greater variation in talk, which may affect the amount of opportunity different candidates have to demonstrate their ability, and therefore affect the reliability of the test. How to balance the requirements of a good test when they compete with one another is the eternal question in assessment.

Another consideration is the criteria against which such performances should be marked. The study evaluated the tasks' ability to distinguish stronger and weaker candidates according to existing linguistic criteria. But to the extent that the tasks expand construct coverage, it would be for naught if these discourse and interaction management aspects of speaking ability were not ultimately captured in the score. On another note, taking the lead in the interaction was, for candidates from certain backgrounds, an alien and uncomfortable prospect, though it is argued that it is a skill they will need to develop anyway in the Western academic contexts they are going to.

The study limits itself to considering an additional section of the multi-componential speaking test, and where that additional section might best be placed. But to the extent that one is considering changes, one might decide to be more audacious. Why not go for a two candidate format to further extend the range of interaction types? Why not introduce a role play for greater verisimilitude? Why not have an online component, given that we nowadays increasingly interact through that medium? The possibilities are endless; this study points out some next steps.

Gad S. Lim
Principal Research Manager
Cambridge English Language Assessment

Candidates questioning examiners in the IELTS Speaking Test: An intervention study

Abstract

This study considers the possibility of introducing an element of more naturalistic, two-way interaction into the IELTS Speaking Test (IST). The research aimed to evaluate the effect of an intervention, namely the addition of a fourth part into the structure of the IST. This was intended as a two-minute section in which the candidate asked questions on a typical IST topic to the examiner, who then replied. Asking questions is a skill that university students have to develop, and such sequences could potentially provide useful rating data and a two-way interactional element.

This four-part test was trialled by 18 candidates and three (3) examiners under six (6) conditions which enabled evaluation of the best format and location for the new part. The study evaluated whether candidate-led question-answer sequences are actually produced and whether value is added to the test in any way. The tests were recorded, transcribed and analysed using a CA approach. Both candidates and examiners were interviewed about the intervention.

The new Candidate Question (CQ) part did generate candidate-led question-answer sequences as anticipated, even with weak candidates. The research suggests that the 'examiner leading statement' format after the existing part 2 would be optimal. The CQ part does add value in a number of ways, according to both examiners and candidates, creating a context for more naturalistic, two-way interaction. Higher-scoring candidates took a more active role, developing topic and making other kinds of speech moves outside the question-answer lockstep. Examiners felt that candidate questions provided useful extra information for rating purposes. Potential disadvantages are increased test duration and variation in amount and type of examiner talk.

Authors' biodata

Paul Seedhouse

Paul Seedhouse is Professor of Educational and Applied Linguistics at Newcastle University, UK. His monograph *The Interactional Architecture of the Language Classroom* was published by Blackwell in 2004 and won the Modern Languages Association of America Mildenberger Prize. Working with colleagues in computer science, he used two grants to build kitchens which use digital technology to teach users European languages and cuisines simultaneously www.europeandigitalkitchen.com. He has also had three grants to study interaction in the IELTS Speaking test; the *IELTS Research Reports* on these projects are available on the IELTS website.

Sandra Morales

Dr Sandra Morales took her PhD in Educational and Applied Linguistics at Newcastle University, UK. She is an experienced language teacher and teacher trainer and has worked with undergraduate and postgraduate TESOL students in her home country, Chile, and the UK. Her area of research is Computer Assisted Language Learning, mainly, teacher education and the use of online and blended learning resources for teaching and learning. Sandra has worked in a number of research projects sponsored by the European Union and has published her work in international journals and books. She has also presented in conferences such as, EuroCALL, WorldCALL and BAAL. Sandra is currently a lecturer in TESOL in the English Teacher Education program at Universidad Diego Portales in Santiago, Chile.

Table of contents

List of tables

1 Introduction

1.1 Background information on the IELTS Speaking Test

In this section, we provide information on how the IELTS Speaking Test (IST) is currently configured, as a baseline from which the research intervention was developed.

ISTs are encounters between one candidate and one examiner and are designed to take between 11 and 14 minutes. There are three main parts. Each part fulfils a specific function in terms of interaction pattern, task input and candidate output.

In Part 1 (Introduction), candidates answer general questions about themselves, their homes/families, their jobs/studies, their interests, and a range of familiar topic areas. The examiner introduces him/herself and confirms the candidate's identity. The examiner interviews candidate using verbal questions selected from familiar topic frames. This part lasts between four and five minutes.

In Part 2 (Individual long turn), the candidate is given a verbal prompt on a card and is asked to talk on a particular topic. The candidate has one minute to prepare before speaking at length, for between one and two minutes. The examiner then asks one or two rounding-off questions.

In Part 3 (Two-way discussion), the examiner and candidate engage in a discussion of more abstract issues and concepts which are thematically linked to the topic prompt in Part 2.

Examiners receive detailed directives in order to maximise test reliability and validity. The most relevant and important instructions to examiners are as follows: "Standardisation plays a crucial role in the successful management of the IELTS Speaking Test." (Instructions to IELTS Examiners, p. 11). "The IELTS Speaking Test involves the use of an examiner frame which is a script that must be followed…Stick to the rubrics – do not deviate in any way...If asked to repeat rubrics, do not rephrase in any way...Do not make any unsolicited comments or offer comments on performance." (IELTS Examiner Training Material 2001, p. 5).

The degree of control over the phrasing differs in the three parts of the test as follows: "The wording of the frame is carefully controlled in parts 1 and 2 of the Speaking Test to ensure that all candidates receive similar input delivered in the same manner. In part 3, the frame is less controlled so that the examiner's language can be accommodated to the level of the candidate being examined. In all parts of the test, examiners are asked to follow the frame in delivering the script. Examiners should refrain from making unscripted comments or asides." (Instructions to IELTS Examiners p. 5).

Research has shown that the speech functions which occur regularly in a candidate's output during the Speaking Test are:

- providing personal information
- providing non-personal information
- expressing opinions
- explaining
- suggesting
- justifying opinions
- speculating
- expressing a preference
- comparing
- summarising
- conversation repair
- contrasting
- narrating and paraphrasing
- analysing.

Other speech functions may emerge during the test, but they are not forced by the test structure.

Detailed performance descriptors have been developed (available on the IELTS website) which describe spoken performance at the nine IELTS bands, based on the following criteria:

> **Fluency and Coherence:** the ability to talk with normal levels of continuity, rate and effort and to link ideas and language together to form coherent, connected speech. The key indicators of fluency are speech rate and speech continuity. The key indicators of coherence are logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices (e.g. connectors, pronouns and conjunctions) within and between sentences.
>
> **Lexical Resource:** the range of vocabulary the candidate can use and the precision with which meanings and attitudes can be expressed. The key indicators are the variety of words used, the adequacy and appropriacy of the words used and the ability to circumlocute (get round a vocabulary gap by using other words) with or without noticeable hesitation.
>
> **Grammatical Range and Accuracy:** the range and the accurate and appropriate use of the candidate's grammatical resource. The key indicators of grammatical range are the length and complexity of the spoken sentences, the appropriate use of subordinate clauses, and variety of sentence structures, and the ability to move elements around for information focus. The key indicators of grammatical accuracy are the number of grammatical errors in a given amount of speech and the communicative effect of error.
>
> **Pronunciation:** the capacity to produce comprehensible speech in fulfilling the Speaking Test requirements. The key indicators will be the amount of strain caused to the listener, the amount of unintelligible speech and the noticeability of L1 influence (*IELTS Handbook 2005*, p. 11).

Equal weighting is given to each of the criteria. This is an analytic or profile approach (Taylor and Galaczi, 2011) in which several performance features are evaluated separately on their own subscale prior to combining sub-scores to produce an overall score.

1.2 Literature review

The rationale for this study is based on Seedhouse and Harris' (2010) suggestion of adding an additional fourth part to the IST. They argued that, although part 3 is termed 'two-way discussion', it is almost identical to part 1 interactionally, in that it consists of a series of topic-based question-answer adjacency pairs. There are hardly ever any opportunities for candidate to introduce or shift topic and they are generally closed down when they try to do so. They further claimed that, taking an overview of topic development in the Speaking Test as a whole, a problem is that it is almost entirely one-sided. Candidates currently have little or no opportunity to display their ability to introduce and manage topic development, ask questions or manage turn-taking. The clear empirical evidence is that part 3 currently does not generate two-way discussion as was originally envisaged. The authors' recommendation was to add a short fourth part, which might last for two minutes. This part would specifically avoid the examiner asking any questions at all. Rather, the candidate would have the opportunity to lead a discussion and to ask the examiner topic-related questions.

Part 4 could start in a number of ways. The examiner could introduce a topic by making a leading statement which the candidate can then follow up by asking a question. Alternatively, the candidate could be instructed to ask the examiner questions about topics previously discussed, or could be allowed to introduce a topic of their own choice. Such a part 4 would give candidates the chance to take a more active role and to develop topic in a different way. It would also allow a part of the IST to have a closer correspondence with interaction in university small group settings, in which students are encouraged to ask questions and develop topics. The current study is, therefore, an intervention based on Seedhouse and Harris's (2010) suggestions and an evaluation of their feasibility and of whether value is thereby added or not.

The relationship between examiner and candidate has been the subject of research interest, with variation in examiner behaviour being seen as a confounding variable (Fulcher 2003: 147). In relation to the IST, Taylor (2000) identifies the nature of the candidate's spoken discourse and the language and behaviour of the oral examiner as issues of current research interest. Wigglesworth (2001, p. 206) suggests that: "In oral assessments, close attention needs to be paid, not only to possible variables which can be incorporated or not into the task, but also to the role of the interlocutor…in ensuring that learners obtain similar input across similar tasks". Brown (2003) analyses two IELTS tests (old format) involving the same candidate taking the same test with two interviewers with different interactional styles. The candidate's communicative ability in the two interviews was rated differently by four raters. This study emphasised the need for interviewer training and standardisation of practices; this was subsequently implemented in the design of the current IST (Taylor, 2001).

Looking back at the history of the speaking component in IELTS, there is nothing new about candidates asking questions. Taylor (2011,vi) explains that:

> *Between 1989 and 2001, the original IELTS Speaking Test included a phase very similar to this in the middle of the test. Phase 3 (out of 5) was a 3 to 4 minute Elicitation task in which the candidate used a Candidate's Cue Card to question the examiner on a given topic, and the examiner responded by drawing on information contained in their Interviewer's Task Sheet (see examples of this task on pages 442–443 in Davies, 2008). Analyses of the operational test as part of the 1998–2001 IELTS Speaking Test Revision Project indicated that, although the candidate was ceded the floor and given the initiative to question the examiner and to develop the thread of discourse, Phases 3 and 4 often failed to elicit anything more than a perfunctory reverse question-answer scenario and thus did not provide the richer sample of candidate performance that was being sought. An additional risk was that the elicitation problems could lead to significant variations in amounts and type of examiner talk. As a result, the format was not reintroduced into the revised IELTS Speaking Test in 2001. It might be interesting, nevertheless, to undertake some small-scale experimental studies exploring alternative approaches that might successfully address the limitations in this area identified by the study.*

A universal question in language testing is the extent to which talk in one discourse setting can predict the ability to interact in another discourse setting. The IST and interaction in universities are related in terms of gatekeeping for entry into the next stage in an educational process Therefore, it is legitimate to examine the two varieties of talk in terms of whether the interactional experiences of students align or not in the different settings.

Interaction in universities is particularly relevant to IST design; as McNamara and Roever (2006: 16) suggest, in the case of admissions tests one needs to model the demands of the target setting and predict the standing of the individual in relation to this construct. In the case of small-group interaction in university seminars, workshops and tutorials, we know from the literature (e.g., Benwell and Stokoe, 2002) that students are expected to ask questions to tutors. It, therefore, follows that having a fourth part to the IST, in which candidates ask questions to the examiners, might facilitate a closer alignment between the two varieties of talk.

Although the literature on the IST and oral proficiency interviews (OPIs) in general contains a number of studies which focus on the questions which examiners ask candidates, there is a major gap in relation to research into questions which candidates ask to examiners. Although there have been a number of OPIs which involved candidates asking questions to examiners (for example the pre-2001 original IELTS Speaking Test), these do not appear to have resulted in published research studies of the specific phenomenon of questions asked by candidates. This is therefore the research gap addressed by the current study.

2 Research design

2.1.1 Research focus

This study aimed to evaluate the effect of a specific intervention which involved the insertion of an additional component into the 3-part structure of the Speaking Test as described above. The intervention was implemented as a section (intended to last two minutes) in which the candidate had to ask questions on a typical IST topic to the examiner, who then replied to these. This section was intended to generate more naturalistic, two-way interaction. Asking questions is a skill that university students have to develop. Such a sequence could potentially give raters very useful data to confirm decisions on grades.

This additional section was trialled by 18 candidates and three examiners considering the variables of format and location. This study enabled evaluation of the best variables for such an additional component. The intended outcomes are evaluations of whether candidate-led question-answer sequences are actually produced and whether value is added to the IST in any way.

2.1.2 Research questions

The main question is:
Does the new Candidate Question (CQ) section generate candidate-led question-answer sequences as anticipated, and if so, does this add value to the IST?

The sub-questions are:

1) Does the CQ section generate more 'naturalistic' and 'two-way interaction' than the existing 3 parts of the IST?
   This question will be answered by CA analysis of the interaction.

2) Which of the two possible CQ section formats (see Section 2.2.2 below) is most likely to be successful in generating candidate-led question-answer sequences? Which format seems to be a more 'authentic' task?

3) Which location of the CQ section is most likely to be successful in generating candidate-led question-answer sequences, namely after Part 1, 2 or 3?

Questions 2 and 3 will be answered by CA analysis of the interaction, by post-test interviews and by examiner focus group.

4) What is the relationship between candidate production of questions in the CQ section and their own allocated grade?

This question will be answered by analysis of the candidate questions compared with test results and by post-test rater reports.

5) Do the examiners believe that the CQ section adds any value to the IST? If so, in what way? If not, why not?

6) Do the candidates believe that the CQ section adds any value to the IST? If so, in what way? If not, why not?

Questions 5 and 6 will be answered by post-test interviews and examiner focus group.

2.2 Methodology

2.2.1 Intervention study

This was an intervention study, the aim of which was to evaluate the effect of an intervention in the form of an additional component in the IST. This was a short section (intended to last two minutes), in which the candidate had to ask questions on a typical IST topic to the examiner, who replied to these.

The fundamental aim of the current research project was to evaluate the feasibility and potential added value of such an addition (or CQ section) to the IST. An intervention study approach was therefore appropriate, since this was an addition to an existing system or structure. Taking IST materials for parts 1-3 which are no longer in use (supplied by Cambridge Assessment), CQ section frames for candidates and examiners were written to prompt candidates to ask a series of questions to which the examiner would reply, intended to last around two minutes. The topic of the questions was related to those developed for examiner use in parts 1-3. The frames were piloted and revised in cases where problems were found to occur.

2.2.2 Variables

There were two possible formats for the CQ frames, which were treated as variables in the research design and evaluated. It is important to consider the extent to which it is possible for candidates to prepare themselves in advance for spoken tasks. Therefore, the research was interested in whether one frame format might be more susceptible to preparation effects and to generating formulaic interaction than the other.

1) *Examiner Leading Statement (ELS):* The examiner introduced a topic by making a leading statement which the candidate then followed up by asking a question. The leading statement was related to topics previously discussed, e.g. "I saw a really good film recently" or "I like/don't like taking photographs". The candidate asked questions about this and took on the development of the topic. In this format, the candidate did not have prior notification of the topic, although it was related to a topic previously discussed during the same IST.

2) *Candidate Prompts (CP):* The candidate received a frame card with a series of bullet-pointed instructions to ask the examiner questions about topics previously discussed. For example, "ask the examiner about a good film s/he has seen" or "ask the examiner if s/he likes taking photographs". In this format, the candidate took the initial lead, had prior notification of the topic and prompts on how to develop the topic. The examiner also had bullet points so s/he knew what to expect.

It also needed to be established whether the CQ section might be best located after part 1, 2 or 3 of the IST and, therefore, these three locations were also treated as variables in the research design and evaluated. There were six variables in total:

Table 1: *Variables*

ELS after section 1	**ELS after section 2**	**ELS after section 3**
CP after section 1	**CP after section 2**	**CP after section 3**

2.2.3 Sampling and data collection procedures

Candidates

The intervention study took place in Newcastle University. The research used 18 students and three examiners from educational institutions in Newcastle and elsewhere in the North-East who volunteered for the study. Efforts were made to make the candidates as heterogeneous a group as possible within the group available. We tried for an even spread of candidates across bands, based on previous scores. We also tried to ensure a mix of other candidate features, such as country of origin and gender. All candidates had previously taken IST and were informed that they would be taking the standard IST with one additional section. Out of the 18 candidates, eight (8) were preparing for the IELTS test, seven (7) were MA students and three (3) were PhD students. Table 2 shows the candidates' background information.

Table 2: *IELTS candidates*

Candidate	Age	Gender	Country of origin	Time in the UK	Previous IELTS speaking score
1	33	Female	Ghana	10 months	8
2	36	Male	Colombia	5 years	8
3	31	Male	Saudi Arabia	Non specified	3
4	36	Male	Iraq	1 year	7.5
5	28	Male	Libya	Non specified	4
6	31	Female	Libya	7 months	6
7	20	Male	Angola	8 months	6
8	26	Female	Belarus	2 years	8.5
9	33	Female	China	10 months	6.5
10	26	Male	Libya	10 months	6.5
11	19	Male	Angola	8 months	6
12	32	Female	China	2 years	8.5
13	28	Female	China	10 months	6.5
14	37	Female	Iraq	6 months	6
15	29	Male	Libya	6 months	3.5
16	31	Female	China	1 year	6.5
17	21	Female	China	10 months	6.5
18	29	Female	China	10 months	7

Examiners

There were three experienced IELTS examiners, and each examiner covered each of the format and location variables once, making a total of 18 ISTs recorded and analysed. Each examiner, therefore, conducted six ISTs, rated them and took part in post-test interviews for each IST during a single day's work (six hours).

Table 3: *IELTS examiners*

Examiner	Gender	Nationality	Place of work	Years of experience in IELTS
1	Male	British	Further Education (College)	2007–present
2	Male	British	Further Education (College)	2008–present
3	Female	British	Higher Education (Centre for academic and language preparation for International students)	2008–present

2.2.4 Data collection procedures

This section outlines the various sources of data and how they were collected.

IELTS Speaking Tests

The CQ section was trialled by 18 candidates of varying levels of proficiency with trained IELTS examiners, and the interaction was recorded and transcribed. Test conditions were made as similar to real conditions as possible. However, examiners did not carry out the normal preliminary administrative procedures. The length of time actually taken by the 18 ISTs in this study were as follows:

Table 4: *CQ section duration*

Candidate/variable/score	P4 Test time/min, sec.	Total test time/min, sec.
1 (ELS after P1, score: 9.0)	2.17	12.17
2 (CP after P1, score: 8.0)	2.27	12.16
3 (ELS after P2, score: 5.0)	2.12	12.34
4 (CP after P2, score: 7.5)	2.28	12.53
5 (ELS after P3, score: 5.0)	2.23	12.50
6 (CP after P3, score: 6.5)	2.03	13.28
7 (ELS after P1, score: 7.5)	2.05	16.10
8 (CP after P1, score: 8.0)	2.16	17.19
9 (ELS after P2, score: 7.5)	2.14	14.41
10 (CP after P2, score: 6.0)	2.46	16.08
11 (ELS after P3, score: 6.5)	3.47	17.34
12 (CP after P3, score: 7.5)	3.02	16.49
13 (ELS after P1, score: 6.5)	2.00	12.54
14 (CP after P1, score: 5.5)	2.00	13.25
15 (ELS after P2, score: 4.5)	2.39	15.21
16 (CP after P2, score: 6.0)	2.55	15.38
17 (ELS after P3, score: 6.5)	2.26	15.10
18 (CP after P3, score: 7.0)	2.27	14.20
AVERAGE	**2.32**	**14.34**

IST scores

Each of the 18 new 4-part ISTs was rated following the usual procedures by three trained examiners and the resultant scores compared with their previous official ISTs.

Candidate question and score comparison

After the tests, data from the candidates' scores were processed and compared in order to answer sub-question 4 (later in this report).

Interviews

Interviews were conducted with both examiners and candidates immediately after the tests.

The 18 candidates were interviewed to see whether the CQ section altered their experience of IST in any way. The main question they were asked was: *Did you find the CQ section more or less challenging than the rest of the IST?*

The three examiners were interviewed to establish their opinion as to:

1. Whether the CQ section adds any value to the IST from their perspective
2. Which of the two CQ section formats they prefer and why
3. Where the CQ section should be located in the IST
4. Whether they think candidates might prepare themselves for the CQ section.

They were also asked to evaluate each of their candidates' individual performances on the CQ section as a rater report.

Interaction

The ISTs were transcribed and interaction was analysed using a Conversation Analysis (CA) approach to see whether the CQ section does actually deliver candidate-led question-answer sequences as expected. The methodology employed is Conversation Analysis (CA) (Drew & Heritage, 1992; Lazaraton, 2002; Seedhouse, 2004). Studies of institutional interaction have focused on how the organisation of the interaction is related to the institutional aim and on the ways in which this organisation differs from the benchmark of free conversation. CA institutional discourse methodology attempts to relate not only the overall organisation of the interaction, but also individual interactional devices to the core institutional goal. CA attempts, then, to understand the organisation of the interaction as being rationally derived from the core institutional goal. This institutional discourse perspective was applied to the interaction organisation of the IELTS Speaking Test in Seedhouse & Egbert's (2006) study, the overall finding being that "The organisation of turn-taking, sequence and repair are tightly and rationally organized in relation to the institutional goal of ensuring valid assessment of English speaking proficiency" (p. 191). In this study, Richards and Seedhouse's (2005) model of "description leading to informed action" is employed in relation to applications of CA. The study will link the description of the interaction to the institutional goals and provide proposals for informed action based on analysis of the data.

Furthermore, the transcribed interaction was analysed for evidence in relation to the variables:

1. the two alternative CQ section formats
2. where the CQ section should be located in the IST, namely after part 1, 2 or 3.

In relation to 1) the analysis focused on the flow of the interaction, evidence of trouble and relative success in the production of questions and answers. In relation to 2) the analysis focused on the relative smoothness of transition from part 1, 2 or 3 to the CQ section. Of particular interest is the question of whether part 3s become more 'two-way' or 'naturalistic' in a CQ section.

Taylor's (2011, vi) reported in relation to a review of candidate questions in the original pre-2001 IST that "candidate questions often failed to elicit anything more than a perfunctory reverse question-answer scenario and thus did not provide the richer sample of candidate performance that was being sought". Therefore, the basic analytical interest in this regard was in whether speech moves occur in the CQ section which are neither questions nor answers and which, therefore, indicate that the participants are breaking out of the perfunctory Q-A lockstep. This is, in effect, speech act discourse analysis involving the identification of speech acts.

Examiners' focus group

Several months later, having completed the data analysis as described above, we invited the three examiners to a focus group at which we presented the results and asked for their views on the results and on the CQ section, including its potential advantages and disadvantages. The focus group's views on the variables was of particular interest and we wanted to see if it would be possible to achieve group consensus on the optimal configuration of the CQ section in terms of the format and location.

2.2.4.1 Limitations

The fundamental limitation of the study was the small sample size of three examiners and 18 students, which was proportional to the award received. This meant that many statistical treatments were not possible. However, the study was rich in terms of the variety of data sources which could be brought to bear on complex and multi-faceted issues arising from the intervention. Furthermore, the quality of CA analysis does not relate to sample size.

A further limitation is that the CQ section was novel for examiners and candidates, whereas they had all had training for the 3 parts of the existing IST. Therefore, it is not possible to establish whether or not the problems reported by both examiners and candidates would disappear if the CQ section were routinised in the IST and if all participants were well prepared for it.

3 Findings

The main research question in our study was:

> *Does the Candidate Question (CQ) section generate candidate-led question-answer sequences as anticipated, and if so, does it add value to the IELTS Speaking Test?*

In order to answer it, six sub-questions were produced, which are answered below by application of data:

3.1 Sub-question 1: Does the CQ section generate more naturalistic, two-way interaction than the existing 3 parts of the IST?

This question is answered by using interactional evidence from speaking tests.

3.1.1 Evidence of naturalistic, two-way interaction in the CQ section

In this section, we examine interaction from CQ sections to search for any evidence that the interaction has become more like naturalistic two-way interaction. As specified in the methodology section, we are interested in whether speech moves occur which are neither questions nor answers and which, therefore, indicate that the participants are breaking out of the Q-A lockstep.

Extract 1

Test examiner 1, candidate 1. Variables: ELS after P1, Score: 9.0

109 C: I see:: .hhh (.) so ↑how ↑do ↑you (.) ↑cos the problem with me is
109 that when I ↑go to the::se (.) restaurants
110 E: hm mm.
111 C: I'm not sure what to ↑choo::se, (0.4) you know cos (.) I'm not used
112 >to that kind of food< ↑how do you use your intuition ↑o:::r? (.) you
113 go with a ↑friend?
114 E: o:::h (.) no:: w- I: rather I I:: (.) I ask advice from someone who's
115 ↑tried it
116 C: oka::y,
117 E: u:::::m °e:::r° I wouldn't <try something withou:t> (0.6) knowing
118 anything ab[out it]
119 C: [°ha: ha] ha° .hh so ya::'r a bit of a ↑sceptic when it
120 comes to:::,
121 E: ↑e:rm (0.4) ↑ye:s I think so[:: I] don't want to c'b=
122 C: [e::r]
123 E: =order something I don't like (0.6) °yeah°
124 C: good (0.4) er I think I lear- learn from that I'll pick a cue from
125 that it ↑happened to me once in London ↑I[::,] (0.6)=
126 E: [mm:]
127 C: =have no I was hungry:: (0.4) and so I just walked into the first
128 restaurant I, sa::w,
129 E: ↑ah ↑ha
130 C: and it was a total flop,
131 E: okay
132 C: .tch ha [ha ha ha .hhh .hh I] had to pay for something
133 E [ha ha ha ha ha ha]
134 E: wha[t happened]
135 C: [I didn't] ea:t,
136 E: a::h what was it,
137 C: ↑it ↑was a ↑ty- I don't I don't remember what it was but I ↑think I
138 just saw something with ri::ce,
139 E: hm
140 C: so I a↑ssumed it was the rice dish I I was used to at home and I
141 ↑[orde]red it and it wa:::s, (0.8) I couldn't eat the food
142 E: [okay]
143 E: ↑o:↑ka::y (0.4) okay ((name omitted) [u::m] ↑thank you=

In line 108 we see C start a question (how do you) and then tells of his/her own experience (of going to restaurants and being unsure what to choose), before completing the question in lines 112 and 113. This is a new and different speech move (a pre-question) by C which is typical of ordinary conversation and appears 'naturalistic'. Again in line 121, C performs a new and different speech move by ascribing an identity to E as "a bit of a sceptic", rather than asking a question. In line 124, C resolves to learn from E's strategy and tells a narrative about a previous restaurant experience in London, rather than asking a question. In lines 134 and 136, E actually asks two questions to C, i.e. the opposite of what is supposed to happen. From the perspective of information exchange, it is necessary for E to ask the questions to find out the missing elements and C completes the anecdote in lines 140 and 141.

The overall flow of the interaction is very naturalistic here and closely resembles that of ordinary conversation; a variety of speech moves are introduced and the topic is allowed to flow, rather than being restricted by the need to follow a candidate question-examiner answer lockstep template. We see both parties asking questions (a key indicator of two-way interaction) and a variety of conversational moves apart from questions and answers.

Extract 2

Test examiner 2, candidate 10. Variables: CP after P2, Score: 6.5

165 C: .hh ↑okay ↑and, (0.4) >I mean< (.) ↑whe:re (0.4) e::r (.) ↑where
166 e::r (0.4) where are the be:st places that you fi:nd when you find
167 yourself a:nd (.) I [mean] (0.4) for ↑me my room is my wo:rld
168 E: [mm::]
169 E: [right]
170 C: [and,] and ↑you: whe[re]
171 E: [hm] mm
172 C: where do you ↑think it's the (.) best pla:ce [to::] (.) feel (.)=
173 E: [hm::]
174 C: =relaxed
175 E: usually in my flat (0.4) e::rm if i::t's (.) exer↑cising, o::r
176 something or maybe going for a ↑ru::n then it'll be near my flat or
177 in the ↑↑gym,
178 C: hm ↑mm
179 E: .hh u::m, but also so:me (.) u::m (.) local pubs are usually quite
180 relaxing as well
181 C: °mm ↑hm°
182 C: and ↑how ↑o:ften (.) do you do: (.) thi:s I mean is there any
183 specific e::r (.) circumstances that you be in o:r it's (.) er (.)
184 it's something you usually do o:r always do

It is evident in the data that some candidates feel able to add different speech moves to their questions. For example, in line 167 C adds personalisation to the question about where E relaxes, noting that "for me my room is my world". In lines 182–4, C adds different speech moves, namely clarification and a series of options, to the initial question. In these ways, C makes the scripted questions more 'conversational' and naturalistic, rather than delivering the bare prompts.

Extract 3

Test examiner 2, candidate 11. Variables: ELS after P3, Score: 6.5

285 E: happy? .hhh ↑u::::::m, (.) ↑so:: (.) I ↑often relax by doing some
286 exerci:se
287 (1.0)
288 C: °that's goo:d° (0.4) e:::r (.) ↑which kind of exercise tell me mo:re

In the above extract, we see how a different candidate responds to E's answer with an evaluation, another question and then the instruction "tell me more". The candidate, therefore, adds different speech acts both before and after the question.

Extract 4

Test examiner 3, candidate 18. Variables: CP after P3, Score: 7.0

239 E: u:::m, ↑mm:: I::::: (.) listen to mu↑↑sic, (0.4) or I:::
240 [w- wa-]
241 C: [same as] me
242 E: watch mo↑vie:: (0.4) u::m (.) I ↑like ↑gardening actually .hhh if
243 [if the if] the weather's nice

In the above extract, C adds a new move to confirm they share the same method of relaxation (line 241 "same as me"). This is typical of the empathetic, rapport-building moves we find in ordinary conversation, but which are normally lacking in the IST.

Extract 5

Test examiner 3, candidate 18. Variables: CP after P3, Score: 7.0

295 E: [yea::h I] ↑yes I ↑really like tha::t (.) we >↑have
296 ↑a< (0.4) ↑big tree at the end the garden which [I don't] like =
297 C: [↑o::::h]
298 E: =because it creates a lot of sha::de
299 C: .hhh hhh
300 E: so:: I do[n't like that so much]
301 C: [you can ↑sit under it]
302 E: it ↑we::ll (.) >it doesn't get< (.) if it's hot it's nice
303 C: .hhhh ↑ah ↑ha ↑ha ↑ha

In line 301, we see a different kind of interactional move from C. E has described a big tree in her garden, which creates a lot of shade and C comments "you can sit under it", thus offering a possible direction to develop the topic, rather than asking a question.

Extract 6

Test examiner 2, candidate 12. Variables: CP after P3, Score: 7.5

308 ↑could you please te:ll me tha:t, (0.4) ↑normally u::m, (0.4) ↑ho:w
309 (0.4) do you relax yourself
310 E: hm ↑mm .hhhh so >similar to< you:: I (will) watch tee vee or movies
311 in the home (0.4) a:nd sometimes with a glass of ↑wine (0.4) .hhh I
312 might go for a ↑walk (0.4) e:::r along the ↑river cos I lived quite
313 close to the ↑river (0.4) .hhh and I go to the gy:m quite a few times
314 a week (.) as well

In the above extract, we see a different kind of evidence of two-way interaction in that interactants are now able to refer to and build on each other's previous turns. So in line 310, E, when asked about how he relaxes, replies that he does so in a similar way to the candidate. In the current IST, such moves are not commonly encountered.

Extract 7

Test examiner 3, candidate 16. Variables: CP after P2, Score: 6.0

146 C: °okay° (0.4) so (.) (normally) (.) what kinds of wa:ys would you
147 ↑choo:se to relax your↑self?
148 E: .hhh well like ↑you::: I do:: I like listening to music as we:ll
149 (0.4) that's one of my favourite wa:ys (0.6) of relax↑ing,

In a similar way, we see in line 148 that E refers back to C's previous turn when providing an answer.

Extract 8

Test examiner 3, candidate 17. Variables: ELS after P3, Score: 6.5

240 E: ↑u::::m, (0.4) .hh but I also like fi::lms like u::m, (0.4) .tch
241 e:::r (0.4) ↑lord of the ↑↑ri::ngs (0.4) and you know with sort of
242 those kind of ↑storie::s I like ↑good stories I think,
243 C: oh y[ea::h]
244 E: [↑good] stori[es]
245 C: [me] too (0.4) I ↑always I also like them romantic
246 fi:lms (0.4) a::nd (.) u::m (0.4) and but ↑whe- the how >often< do
247 you choose to:: (0.4) er see a film

Similarly, candidates sometimes refer back to the examiner's turns in the CQ section. In line 245 we see the new speech move of C agreeing with E's taste in films, which is again a typical feature of two-way interaction.

Another new speech move introduced into candidate talk in the CQ section is that of evaluation of E's responses. In the standard IST, examiners are trained not to express evaluations of candidate turns and candidates do not have the opportunity to evaluate examiner turns.

Extract 9

Test examiner 3, candidate 17. Variables: ELS after P3, Score: 6.5

264 E: [and] then I just to esca:pe fo:r (0.8) an hour or
265 two
266 C: yea:[:h good]
267 E: [I like] that
268 C: °good°
269 C: ↑and e::r, (.) and for our international students we always see a
270 film we can learn english
271 E: ↑o::h
272 C: from the films

Here we see two evaluation moves by the candidate in 266 and 268. Instead of asking another question, C follows this with a new move of an additional statement of information, which develops the topic of films from the perspective of ELT students in lines 269–70. In line 271, we see E reacting to this as new information.

So the interactional evidence from these data is that the CQ section has the potential (with some candidates) to deliver the type of two-way interaction which was originally intended to occur in part 3 of the IST.

We sometimes see in the CQ section that the examiners sometimes ask questions to the candidates, although it is supposed to be the candidates who are asking the questions. In these cases, it does not appear to be the case that the examiners are trying to seize back control of the interaction, but rather that more naturalistic, two-way interaction is developing.

Extract 10

Test examiner 1, candidate 2. Variables: CP after P1, Score: 8.0

83 E: o::h ↑Greek foo:d, (0.4) um I like the ↑cheeses and the ↑salads but,
84 ↑no:: I don't like big chunks of meat (.) °no° ↑do ↑you?
85 (.)
85 C: .hh yea:h (.) for ↑me >↑I ↑like ↑it >↑to ↑be ↑honest< I just came
87 fro::m (.) from vacations in there (0.6) >it was< kind of ni:ce (0.4)
88 a:::nd (.) do you like ↑the: (0.4) ↑↑cooking by your↑self?

In this extract, we see E asking a question to C in line 84 as a naturalistic follow-up to E's answer, in the same way as often occurs in conversation. C answers E's question and then moves on to ask a further question.

Extract 11

Test examiner 1, candidate 6. Variables: CP after P3, Score: 6.5

163 E: hm mm,
164 C: yeah (0.4) I ↑think (.) ↑u:::m, (0.4) ↑I think tha:t er, (0.4) I
165 don't kno::w w::- ↑why but e::r (0.4) I su↑ppose ↑that er becau::se e:::rm (0.4) .tch e:::r (0.4)
166 be↑cause of e:::r, (0.4) the ↑no:w
167 there is a lot of er, (.) there is ↑many u:::m develop↑ment?
168 E: °mm°
169 C: i:::n, (.) in ↑everything (0.6) so I think e::::r, (0.4) with e::::r
170 each develop↑me:nt there is, (0.4) a problem
171 E: °mm°
172 C: ↑more tha:n benefits
173 E: .hhh (0.4) oka::y (.) ↑u:m, (0.4) ↑let's move on now to talk about

The examiner has previously said (in response to C's questions) that he likes to relax by going for walks in the countryside on his own and that he talks to himself in his mind. In line 267, E follows up by asking C for his/her view on people talking to themselves and C provides the view that it is natural for people to carry out internal dialogues; E professes himself to be relieved at C's response. We see here that the interactants do not feel constrained to follow the candidate question-examiner answer pattern and are able to introduce alternative speech moves in order to develop the topic.

So the clear interactional evidence from these data is that the CQ section has the potential (with some candidates) to deliver the type of more naturalistic, more two-way interaction which was originally intended to occur in part 3 of the IST. How many candidates produced these new, additional speech moves in the CQ section? The extracts in this section feature nine (9) out of the 18 candidates, so the moves are certainly widespread amongst the sample. We should note that all of the above extracts are from candidates with a minimum 6.5. This tentatively suggests that the ability to introduce new, additional speech moves in the CQ section might be an effective criterion to distinguish higher from lower level candidates.

This finding is at variance with Taylor's (2011, vi) report in relation to a review of candidate questions in the original pre-2001 IST that "candidate questions often failed to elicit anything more than a perfunctory reverse question-answer scenario and thus did not provide the richer sample of candidate performance that was being sought". It is not clear what the reason is for the different findings. It is, for example, possible that different candidate prompting systems were being employed in the two different studies.

3.1.2 Differentiation of higher and lower proficiency candidates

Part of the rationale for a CQ section was the hypothesis that it might provide additional evidence to differentiate higher and lower proficiency candidates, in that, it is challenging for candidates to verbally construct a sequence of grammatically correct questions in English. Here we examine the interactional evidence by comparing high and low-scoring candidates responding to the same examiner leading statement to see whether differentiation is facilitated.

Extract 12

Test examiner 2, candidate 9. Variables: ELS after P2, Score: 7.5

161 E: o↑kay are you ↑ready?
162 C: ↑°yes°
163 E: alright so I ↑often relax by doing some exercise
164 C: oka:y um ↑what kind of exer↑cise do you usually do:?
165 E: e:::r I go to the gy:m three times a wee:k (0.4) °um° °°yeah°° mm
166 C: o:kay and ↑how do you feel whe:n you are doing the exer↑cise in the
167 gym
168 E: .hhh (0.4) e:::r I:: (0.4) u::m (.) fee::l (0.6) it's a sort o:f
169 (0.4) >a r- a r-< u:::m hhhh (.) an escape so I ↑focus everything on
170 what I'm doing in the gym (.) [and I] don't think about (0.4)=
171 C: [↑mm::]
172 E: =whatever else is happening outside the gym

In the above extract, the candidate (score 7.5) is able to construct grammatically correct questions in lines 164 and 166 without any pauses being recorded. There are some slight hesitation markers in the lengthening of 'okay' in both lines and 'um' in line 164. However, it should be noted that the native speaker examiner uses considerably more pauses and hesitation markers in lines 168–170. The question in line 166 is impressive from the grammatical viewpoint in that it contains a relative clause and combines present simple with present continuous. It is also impressive from the perspective of fluency in that C did not know what answer E was going to give in relation to the question about kind of exercise; C and E did not know each other. However, C is able to integrate into the follow-up question in line 111 the new sub-topic of 'gym' introduced by E, without any pause.

Extract 13

Test examiner 1, candidate 5. Variables: ELS after P3, Score: 5.0

255 E: now (0.4) ((name omitted)) I::: (.) ↑I:::: (.) u::m ↑I:: (.) ↑often
256 relax, (0.4) by doing some exercise
257 C: °yea::h°
258 E: mm
259 (0.6)
260 C: ↑why the: (.) [execi:fe] (.) exercise give you: e:::r re↑lax?
261 E: why (.) ↑o::h >I I I I< (0.4) I think (.) I think be↑cau:se (0.4)
262 when I ↑do it, (0.4) I don't think,
263 C: yeah .hh (.) ↑a::nd ↑which ↑ki:nd of e::::r exercises, (.) give you
264 :: e::::r (.) re↑lax?
265 E: oh ther- (.) different thi:ngs er (0.4) I think riding my bicycle to wo:rk that's ↑transport

The above extract features a candidate scored at 5.0 who receives the same examiner leading statement as in the previous extract, with the difference that this is after part 3. From the grammatical perspective, C's questions in lines 260 and 263 contain grammatical errors and also a lexical error, namely 'give someone relax' rather than 'make someone relax'. There are noticeably more hesitation phenomena than in the previous extract, including pauses, three elongated 'e:::r's and a self-corrected pronunciation in line 260. In terms of topic development, we should note that this candidate does not take-up and develop the sub-topic of 'relaxation stops me thinking' introduced by E in line 262. By contrast, the higher-scoring candidate in the previous extract incorporated the examiner's response in his/her follow-up question. Nonetheless, C's questions are comprehensible to the examiner.

The analysis suggests that the CQ section provides candidates with the additional challenge of producing grammatically correct, sequentially and topically appropriate questions in response to examiner leading statements, prompts or answers. The best candidates (as in Extract 12) appear to be equal to this challenge, whereas weaker students (as in Extract 13) display weaknesses in each of the four scoring bands. The CQ section is, therefore, potentially capable of providing extra rating data to examiners which can differentiate between stronger and weaker students and generate a cluster of assessable features (Seedhouse and Harris 2011).

3.1.3 How does the CQ section compare with the three existing parts?

We have seen in section 3.1 above that the CQ section has featured examples of more naturalistic, two-way interaction. However, in order to answer the research question of whether this interaction is more naturalistic and two-way than the existing three parts, it is necessary to see how the interaction compares to that in other parts of the same IST with the same participants. Also, if participants have engaged in more naturalistic, two-way interaction in a CQ section after parts 1 or 2, does this mean that part 3 features more naturalistic, two-way interaction as a result? The data suggest that part 3 interaction remains dominated by the examiner-led, topic-scripted QA adjacency pair.

In order to illustrate this, we examine here the part 3 interaction of the same participants who featured in the naturalistic interaction already analysed in Extract 1 above. Part 3 of this IST was lengthy at 128 lines. E asked 8 questions as well as 2 pre-questions. However, E did not make any other kinds of speech moves of the kind one might expect in naturalistic, two-way interaction. E provided a great deal of back-channelling, including 13 examples of 'mm' or 'hm mm', 4 of 'yeah', 6 of 'okay'.

Extract 14

Test examiner 1, candidate 1. Variables: EP after P1, Score: 9.0

E: thank you, (6.8) ↑so: ((name omitted)) we've e:::r, (0.4) ↑we've been talking abou:t (.) something you do (0.4) ↓e:::r to relax and I'd ↑li:ke to: discuss with you .hhh one or two more general question[ns]
C: [mm]::
E: related to that o↑kay? (0.6) ↑u:m, (0.4) let's talk about causes of stress
C: mm:::
E: right? .hh ↑u:::m, (0.8) ↑what are the main causes of stress that (.) people (0.4) suffer fro::m (0.4) or experience where you live,
C: well ↑I I'll use ↑myself e:r as an example
E: mm
C: ↑back ho:me I used to work in the ↑bank, (0.6) and so my work, (0.4) number o:ne was (.) the ↑biggest cau:se of my stress anytime I was

239 stressed it was work (0.4) .hhhh one becau:se working in the ba:nk
240 (.) takes all my ↑ti::me, (0.4) the ↑targets were qui:te, (.) ↑hu-
241 targe:ts
242 E: mm
243 C: and so ↑a:ll the time I was ↑constantly under pressure,
244 E: hm mm,
245 C: to deliver, and ↑any ↑time I was <fa:lling behi::nd,> (0.4) i- it it
246 ↑took a (0.4) terrific (drain) on me
247 E: mm
248 C: so I think work (0.4) number o:ne,
249 E: okay
250 C: and ↑two:: financially as well, (.) when you are not that, (0.4)
251 financially buoya:nt to do what you want to do
252 E: mm
253 C: I think [it's also]
254 E: [d'you thi]nk d'you think people u::m (.) suffer from that
255 kind of stress in ↑a:ll occupations or is it just specific ones
256 C: ↑well ↑I ↑think, (0.4) every kind of occupation has its own stre:ss
257 E: mm
258 C: because they are ↑different, (0.4) ↑hi:ghs and lows in every
259 occupation but, (0.4) .hhhh from my experiences working in the
260 ba::nk, (.) a ↑teacher would probably, (.) go through a different
261 stress >but I think< ↑most people >suffer from< work related stress

In the above part 3 interaction, E certainly develops a series of questions which build on C's answers, in typical 'interview' style. But the interaction in this and the other part 3s in the data remain firmly rooted in the examiner-led, topic-scripted QA adjacency pair archetypal sequence, even when a much more varied and naturalistic kind of interaction has occurred earlier in the very same IST. The reason for this may be that which has already been suggested, namely that the topic-scripted QA adjacency pair is a very efficient and economical mechanism for delivering the institutional business, in a similar way to the three part sequence generally known as IRF (Teacher Initiation, Learner Response and Teacher Follow Up or Feedback) in the L2 classroom (Sinclair and Coulthard, 1975). In the language of Complexity Theory, both structures are powerful 'attractors'. Larsen-Freeman and Cameron (2008: 235), in their discussion of complexity theory in relation to discourse, see the IRF pattern as an attractor on the classroom discourse landscape that shows variability around a very stable form, which has arisen through adaptation to particular classroom contingencies. The discourse system will tend to return to the IRF attractor because it is a pattern that works, or a preferred behaviour of the system. So, because in the IST, part 3 prompts are similar to part 1 prompts, part 3 interaction appears to be attracted to be similar to the topic-scripted QA adjacency pair in part 1 interaction and it is difficult for both participants to break away from this to produce the two-way interaction originally envisaged. This implies that if one wanted to generate two-way interaction in the IST, a totally different way of starting and maintaining the interaction would be necessary, which was the original rationale for the CQ section.

3.1.4 Variation in amount and type of examiner talk

Taylor (2011, vi) reported in relation to candidate questions (in the original pre-2001 IST) a concern that these would result in significant variations in amounts and type of examiner talk. The extracts above do indeed show variations in amounts and type of examiner talk in the CQ section. If, however, the aim is to have a section with more two-way, naturalistic interaction, then this does imply relaxing controls and de-standardising the interaction in order to escape the question–answer lockstep, which in turn, implies that there will be variation and heterogeneity in talk. Nonetheless, this represents a challenge in terms of maintaining the validity of the IST. The design of the IST is very much based on standardisation of examiner behaviour in order to ensure all candidates receive input and instructions which are as similar as possible, which in turn ensures valid assessment of English speaking proficiency. Taylor (2000: 8–9) reports on research which highlighted the problems of variation in examiner talk across different candidates and the extent to which this can affect the opportunity candidates are given to perform, the language sample they produce and the score they receive. The studies confirmed the value of using a highly specified interlocutor frame in OPIs which acts as a guide to assessors and provides candidates with the same amount of input and support.

If we examine Extracts 1 to 11 above, we notice a variety of speech moves being performed by examiners. On the one hand, this can be welcomed as evidence of more naturalistic, two-way interaction, but on the other hand, this can be seen as a problem for the maintenance of validity, in that candidates are receiving variable levels of input and support. We argue that this variation is positive. In the other parts of the IST, there is indeed standardisation of examiner behaviour and input, so the CQ section offers a different type of interaction for assessment, one which may be rather closer in nature to the target type of university small-group interaction. Furthermore, part 3 of the IST was intended to produce two-way interaction, but this has not actually happened; the reasons for this are discussed elsewhere. Therefore, the CQ section could deliver the two-way interaction originally envisaged for part 3, and some accompanying variation must also have been originally envisaged. The introduction of a CQ section would have to involve examiner training, and it should be possible to deal with concerns about excessive variation in examiner talk by means of training and task design.

There seems to be something of a paradox at work in relation to discourse in OPIs. In order to have an authentic task which generates naturalistic, two-way interaction, it appears that it is necessary to use a less scripted format and to allow a variety of speech moves by both examiner and candidate. However, this in turn means that examiner behaviour will be less predictable and less standardised, making it more difficult to ensure the validity of assessment and a level playing-field.

3.2 Sub-question 2: Which of the two possible CQ section formats is most likely to be successful in generating candidate-led question-answer sequences? Which format seems to be a more 'authentic' task?

3.2.1. Which of the two possible CQ section formats is most likely to be successful in generating candidate-led question-answer sequences?

In terms of the interactional evidence, the transcripts show that all of the 18 candidates are able to construct questions as required, whatever their level and whatever the format. The data show that the two formats are equally successful in generating candidate-led QA sequences, although in Section 3.2.2 below, it is shown that candidate prompt-based questions tend to be more standardised and predictable.

When we examined the data in relation to the number of words produced by the candidates, it was found that in the 'candidate prompt' (CP) format, the number of words was higher than in the 'examiner leading statement format' (ELS) (CP= 669, ELS=549). However, the sample size is too low to conduct reliable statistical analysis. Furthermore, there is no evident added value to a higher number of words.

The **candidates** stated in interviews that **both ELS and CP formats** were useful for them to ask questions and show their language proficiency to the examiners. It was not hard to form the questions and they felt relaxed. Nevertheless, they explained that in the ELS format they were not always able to ask a question about a topic they wanted and that they struggled with asking questions in relation to the statement produced by the examiner. In this sense, if they did not know much about the topic (e.g. walking, swimming), it was more challenging to generate the questions.

> Candidate 1 (score 9.0): 'I would have loved that the question is something you want to ask'.
>
> Candidate 7 (score 7.5): 'The question just come to your mind and you're gonna ask'.
>
> Candidate 13 (score 6.5): 'Once you can just feel a little bit relaxed, because the examiner will answer your questions and you will feel a little bit more confident about that'.
>
> Candidate 17 (score 6.5): '...I think it's depend on the topic, so, but um, um, for the examiner is the, give me topic about see a film, so and, and the examiner, um from the examiner's answers, she says she likes love stories. That kind of film. I also like that kind of film, so I continue asking questions.'

According to the examiners, however, in the **ELS format**, there is a certain freedom for the candidates, and themselves, to ask and answer questions. They also explained that this format fostered spontaneous communication between candidate and examiner, which gave the candidates an additional opportunity to use their own ideas. On the other hand, although the **CP format** was helpful for students as they had a base to start asking the questions, this seemed to undermine their ability to apply their own language. Some of the candidates even paraphrased the prompts into questions or just followed the order on the card without trying to include any self-generated ideas.

The **examiners' comments** in interviews about the **ELS format** were as follows.

> Examiner 1: 'It replicates more a normal kind of conversation or an attempt at conversation.'
>
> 'It's a conversational gambit isn't it? The kind of thing we do to hopefully get um get a conversation going.'
>
> '[in] The Examiner Leading Statement, they've got freedom. They've got freedom in their follow-up question or reaction.'
>
> Examiner 2: 'It seemed to produce more natural um questioning.'

3.2.2 Which format seems to be a more 'authentic' task?

We now examine the question of which format seems to be a more 'authentic' task. The overall impression from the interactional evidence is that the extracts involving the **CP format** tend to be slightly more standardised and predictable than those involving the **ELS format**. This is to be expected in that candidates have prompts written on a cue card for the **CP format**. Candidates are, therefore, likely to form questions based on the prompts. So in the following extract the candidate has received a card with the following prompts.

> "You should ask the Examiner questions about his/her views on food, using the prompts below. You will lead the conversation. Find out about:
>
> What kinds of food he/she likes and why
>
> What kinds of food he/she dislikes and why
>
> If he/she likes cooking and why/why not
>
> The relationship between food and culture."

The interaction is as follows.

Extract 15

Test examiner 3, candidate 14. Variables: CP after P1, Score: 5.5

56 E: ↑y- you ↑have to ask me some questio:ns
57 C: oka::y er ↑what kind of ↑foo- foo::d e::r you li::ke a:nd ↑why,
58 E: .tch ↑u:::m (.) I li::ke (0.4) e::::rm (.) .tch ↑lots of different
59 types of food I ↑like fish (0.4) a ↑lo::t,
60 C: °yeah°
61 E: and I like vegeta↑bles (0.4) .hhh u:::m I ↑try to eat healthy food
62 C: °oka::y°
63 E: mostly,
64 C: yea::h, (.) s'okay thank ↑you (.) .hhh u::::m, (0.6) er ↑if ↑you:
65 like cooking (0.4) e::r ↑why:? (.) or why not?
66 E: .tch (.) u:::m (0.4) I ↑do like cook↑i::ng
67 C: °°okay°°
68 E: u::::m (.) I: it helps me to re↑la:x, (0.6) u::m, so bou- I've
69 been working all da::y, (0.4) .hh er ↑teachi:ng,
70 C: ↑°°yea::h°°
71 E: ↑it [just helps] me it's [nice to::]
72 C: [↑yea:::::h] [yeah yeah]
73 E: ↑just, (0.4) to rela:x (.) u:m (.) .hhh I like making ↑foo:d for
74 my fami↑ly, (0.4) and for ↑frie::nds, (0.4) I enjoy ↑that,
75 C: °okay°
76 E: u::m, (0.4) .tch e:::r (0.6) ↑↑yeah I j- j- just (.) ↑mostly cos I
77 li- I like to kno:w where the food ↑comes from so I like to make
78 [it fro:]:m
79 C: [yea::h,]
80 C: ↑yeah (.) yeah original
81 E: all the original ingredients [ye:s yes]
82 C: [°mm:::::°] oka:y °it's perfect°.hhhhh

83 (.) e::::r (0.4) ↑can you ↑speak about the relationship between
84 ↑food and culture do you think ↑that's (.) e:r a relationship
85 E: .hhhhh (.) I ↑think I ↑think ↑some cultures it's ↑↑very importa:nt
86 um I: I: um (0.4) working with internationa:l students ↑I know
87 that ↑so:me (.) hhh some ↑cultu::res (.) um (.) they have many
88 different festi↑va:ls ↑based around ↑↑foo:::d, (0.4) .hh u:::m, and so it's s-
89 very im↑po:rtant (0.4) .hh I do- I don't think it's ↑so im↑porta:nt
90 i:n in ↑my culture in [in eng]la:nd (.) .hhh except perhaps fo:::r,=
91 C: [°yeah°]
92 E: =(.) um ↑christmas,
93 C: ↑yeah
94 E: christmas da:::y u:::m (.) is is invol- is bo- >you know< is,
95 C: ↑°mm°
96 E: based aroun[d ↑foo:d]
97 C: [↑°yea:h°]
98 C: °yeah°
99 E: an- we have ↑family (.) m- ↑family time togethe:r
100 (0.6)
101 C: °I see° (0.4) okay ↑thank ↑you

This particular candidate (score 5.5) follows the question prompts very closely in a formulaic way, rather than converting the prompts into more fully formed questions. For example, in line 65 the question asked is "er ↑if ↑you: like cooking (0.4) e::r ↑why:? (.) or why not?". This is a very close rendition of the prompt "If he/she likes cooking and why/why not". This suggests limited proficiency and lack of awareness that this is not a grammatical way of forming a question in English.

In the interviews, the **candidates** claimed that with the **CP format,** the communication felt authentic. However, they would have liked to use their own ideas, instead of having prompts. One issue with this format was that candidates tended to paraphrase the information in the prompts, rather than asking direct questions.

> Candidate 10 (score 6.0): 'Well, to to ask, I mean, I couldn't make a question as a question, I just paraphrased it.'

According to the **examiners**, the **ELS format** seems to trigger more genuine interaction, as the task is less prescriptive (there is no prompt) which tends to promote a more two-way, naturalistic conversation. The three examiners agreed that both formats could work but, on balance, the **ELS format** promoted a more authentic task.

> Examiner 3: 'I think the ones that that manage managed it, got, got a lot of satisfaction from it, and they react reacted more natur-, naturally with me.'

With regard to the CP format, on the other hand, the examiner commented as follows:

> Examiner 1: 'Candidate Prompt is just more of that same, it's more of long turn, it's more of that, it's nothing new. It is something new, but it's of the same format and it's, it's very prescriptive.'

> Examiner 2: 'I thought that, the fact that the, the Candidate Prompt had something which the students could read, erm to, to start helped, erm but that might be to do with the, the, the, the scripting of the instructions and if that was tightened up then maybe that would be not necessary.'

> Examiner 3: 'Because some of them with the prompts, particularly the weaker ones were using the prompts, but just sort of going through them automatically and perhaps not interacting as naturally with what I was saying.'

The concept of the 'authentic task' is a contested one and depends crucially on this issue: with what is the interaction being compared? If the comparison is with ordinary conversation, then it seems fairly clear that the **ELS format** produces interaction which is rather freer and more naturalistic than the **CP format**. However, the comparison could be with small-group interaction in universities, given that the majority of IST candidates are taking the IELTS test in order to enter university programs. From what is known of the literature on small-group interaction in universities (Seedhouse, 2013), there are no reports of students being given prompts for what questions they should ask, although it is clear they are expected to participate by asking questions in social sciences and humanities subject areas, in particular. It is, by contrast, possible to find sequences in which university tutors make observations or leading statements and students ask questions in response, as in the following sequence between a university tutor (T) and two students (S1 and S2).

Extract 16

1 T: year edition uh row-it's a Routledge book (2.0) it this version
2 is two thousand and five. Uh costs about fifteen quid on
3 Amazon, (1.5) I think there is a few copies in the library if
4 you can be bothered to go off up the road.
5 S1: is that the one that you would recommend?
6 T: this one, ((sighs)) to be honest (2.0) I think it's all wrong.
7 (1.0) I could probably write a better one myself, but I haven't
8 got around to it yet,
9 S2: is that the 2005 edition?

(Source: Newcastle University NUCASE data: NC058)

In the above extract, the university tutor makes observations or statements in lines 1–4 and 6–8. Two different students then ask questions relating to those observations/ statements in lines 5 and 9. So we can see that the pattern of tutor observation/ statement – student question is one which does actually occur in university small-group interaction. This means that the examiner-leading statement format does have some degree of authenticity in relation to the target interaction of university small-group interaction. So the examiner-leading statement format does seem to be a slightly more authentic task than the candidate prompt format, and this is the case whether the 'authenticity' is related to ordinary conversation or to university small-group interaction.

To sum up the answer to this question, **both ELS and CP formats** were successful in generating candidate-led question-answer sequences. **ELS** seems to trigger more genuine interaction, as the task is less prescriptive (there is no prompt) which tends to promote a more 'two-way' natural conversation. The three examiners agreed that both formats could work but, on balance, **the ELS format** promoted a more authentic task. The candidate interview data did not result in a clear preference for either format being displayed. There was no conclusive interactional evidence for or against either of the formats. The **CP format** generated more candidate talk, but there is no evident added value to this.

3.3 Sub-question 3: Which location of the CQ section format is more likely to be successful in generating candidate-led question-answer sequences, namely after part 1, 2 or 3?

Firstly, we consider the interactional evidence as to which location is more likely to be successful in generating candidate-led QA sequences. The overall picture is that all candidates were able to produce such sequences in all locations. In order to detect variation in this picture, it was decided to look for any instances of interactional trouble with starting the CQ section in each position. The post-part 2 position proved relatively smooth, with little major interactional trouble. The post-part 1 and 3 positions had slightly more noticeable trouble in establishing the CQ section in 4 out of the 6 extracts in both cases. The trouble tended to involve their doubt about how to participate or to commence their participation, as in the following examples.

Extract 17

Test examiner 1, candidate 1. Variables: ELS after P1, Score 9.0

64 E: thank you: (0.4) okay (.) .hhh ↑u::m, (0.4) ↑so::, (0.6) we've
65 bee:n (.) we've bee:n (.) discussing (0.4) u:::m (0.4) foo:d,(0.4)
66 o↑ka::y? (0.4) ↑u::::m, (.) and in this ↑part of the test(.)
67 u:::m, (0.4) y:::- I would ↑like you to ask ↑me some questions,
68 C: oka:[:y,]
69 E: [o↑k]ay
70 C: °okay°
71 E: u::::m, (0.4) ↑I've (0.6) I've always loved food,
72 C: mm:::.
73 0.8)
74 E: I've always loved food.
75 C: you've always loved food ↑have you ever ↑tried ↑any:: foo:d
76 outside the traditional food (.) like ↑british, (.) you're
77 ↑british I ↑guess

In the above extract, we see that C (score 9.0) does not immediately produce a question in response to E's leading statement in line 71, and so E repeats this in line 74; on the second occasion C does ask the question in response.

Extract 18

Test examiner 1, candidate 6. Variables: CP after P3, Score 6.5

225 E: ↑o::kay, (0.4) .hh okay ((name omitted)) now (0.4) ↑u:m (0.4)
226 be↑fo:re we were talking about what ↑you do:: (0.4) to relax (0.4)
227 ↑yes? (.) oka:y and now I'd like you to a:sk ↑me: (.) some
228 questions about what ↑I do (0.4) to relax (0.6) al↑right?
229 C: okay
230 E: u::m,
231 (3.8)
232 C: ask you:: (0.4) ↑no?
233 E: ask me what I: do to relax,
234 C: ↑o:kay (0.4) .hh u:::::m, (0.6) ↑what ↑do ↑you ↑do (0.4) e:::r to
235 re↑lax?

In the above extract, there is some evidence of hesitation by the candidate in the pause following line 229 and in the checking in line 232 regarding procedure. So, the interactional evidence suggests (rather tentatively) that the position after part 2 is the most suitable in terms of minimising interactional trouble.

Data from the interviews with the examiners was inconclusive regarding the best location for CQ section. For instance, examiner 1 indicated that after part 1 was best in order to warm up the candidates and to build rapport with them for the other stages. In the case of examiner 2, he explained that after part 2 was more useful so they can achieve more in the task. Examiner 3 suggested that after part 3 the interaction with the candidates was more natural as they were more warmed up. The examiners gave comments on the combined format and location variables as follows.

The three examiners agreed that **ELS format after part 1** was coherent, but examiners 2 and 3 explained that in this location the communication was not natural and not logical. They suggested that the candidates needed to be more 'warmed up' to produce their own questions. On the other hand, examiner 1 indicated that the ELS after part 1 was useful to get the students warmed up for the following stages.

The ELS format after part 2 was perceived positively by the examiners. They all agreed that the format after this part worked well and the task was better connected to the previous section than after part 1. They explained that ELS after part 2 promoted natural interaction and flow, and it was a good transition for the next part of the test. The candidates seemed to be more relaxed and in-tune with the examiner.

The ELS format after part 3 produced different opinions among the examiners. Examiner 1 believed that the format did not work well in this location as the flow broke and the task was not linked to the section. In the case of examiner 2, he expressed that the information seemed redundant after part 3. Nevertheless, for examiner 3, the format seemed to have worked better in this location than in the previous ones.

Regarding **the CP format**, the examiners thought that it was more controlled and tended to limit the conversation. It, therefore, did not provide much added value for the final evaluation.

The CP format location after part 1 was appropriate according to the examiners.

The examiners suggested that in **the CP format after part 2**, some candidates struggled with the prompts and ended up changing them into questions without adding new language. The location seemed to have worked properly. For instance, examiner 3 indicated that the after part 2, the candidates were warmed up, which made the transition to the next section easier.

In relation to **the CP format after part 3**, examiners 1 and 2 agreed that the format was prescriptive, restrictive and that the candidates were confused with the questions in this location. Examiner 3, however, thought that the format and location worked well and that interaction was produced naturally at the end of the test.

Considering the six variables overall, the examiner interviews provided no clear-cut winner as to which combination was the best. There was some evidence that the examiners thought **the ELS format** was best for the students to show their language proficiency, and that the best location for it would be after part 2. They explained that the task was well suited to this position and that the candidates seemed more focused and prepared to ask their questions. No clear evidence was found that one frame format might be more susceptible to preparation effects than the other. Some examiners felt that the CP format generated more formulaic interaction than the ELS format. There was no conclusive interactional evidence that any of the three locations was any more or less successful than another in generating candidate-led question-answer sequences.

To sum up the answer to this question, neither the initial interview nor interactional evidence provided a clear-cut favourite combination of format and location. However, the examiner focus group did reach unanimous subsequent agreement that ELS format after part 2 would on balance be best.

3.4 Sub-question 4: What is the relationship between candidate production of questions in the CQ section and their own allocated grade?

Firstly, we consider the interactional evidence of the relationship between grade and candidate question production. We compare only the questions generated by the highest-scoring candidate (9.0) with those of the lowest-scoring candidate (4.5), in order to see whether differences are evident or not.

Table 5: *Example CQ questions band 9*
BAND 9: 1 Candidate / Candidate 1 (ELS after part 1)

CANDIDATE'S QUESTIONS	EXAMINER'S ANSWERS
'Have you ever tried any food outside the traditional food? Like British, you're British, I guess…	'I am yes.'
'Have you eaten anything aside your normal British?'	'Yes, yes, I like, I like Indian food. Arabic food, and food from different countries.'
'And how would you, do you prefer them?'	'Yes, to British food…'
'You do, why's that?'	'I do, yes I do. Well, I think British food is quite bland, there's not much flavour, actually. So, so, I like I like to t- taste some spices an– and different flavours from all around the world.'
When do you get to eat this food here or when you travel?	Oh, here I think. These days in Britain you can buy, er food from all over the world. So yeah and when I travel, but, um yeah, I like t- I like to cook it myself.
Really?	As well yes.
How do you use your intuition? You go with a friend?	Oh, no, I rather I, I, I ask advice from someone who's tried it. ehh, I wouldn't try something without knowing anything about it.

Table 6: *Example CQ questions band 4*
BAND 4: 1 Candidate / Candidate 15 (ELS after part 2)

CANDIDATE'S QUESTIONS	EXAMINER'S ANSWERS
'what's the type music you like?	I like, different types of music, sometimes if I'm very tired, I like to listen to classical music. But then sometimes if I need some kind of energy or, I like listening to more modern music, I like singing, so…
You have a sound, beautiful sound or?	Oh I don't know.
where are you usually, listen, to music?	Often in my kitchen. When, I when I go home I I'm usually in the kitchen I'm cooking or doing something with my sons or, so often, often in the kitchen I have the radio on, but then sometimes, if I'm in the living room and I'm just sitting, I, I put on a I listen to my ipod yeah? and sometimes I listen to my ipod on the, on the metro.
how often you listening to music?	Every day really, yeah every day I think yeah I like it very much.
what do you feeling when you're listening to music?	well it can sometimes make me feel sad you know it depends, but generally it makes me feel it lifts me it makes me feel happy, so sometimes if I have problems I listen to music and I can, try to forget my problems.

Secondly, we consider how candidate scores in the mock 4-part IST compared with their previous genuine IST scores. In total, eight were allocated a higher score that in their previous test, five maintained their scores and five received a lower score than before. Table 7 shows that of the nine candidates who used the ELS format, seven increased their score and two maintained their score. None of the candidates who were given the ELS format got a lower score. On the other hand, of those candidates who were given the CP format (n=9), only one increased their score. Five were assigned a lower score and three kept their original score. This is useful information, nevertheless, it cannot be considered as evidence as the sample size is too small to make generalisations.

Table 7: *Candidates' previous and current scores*

Candidate	Age	Gender	Country of origin	Time in the UK	Previous speaking score	Current speaking score	Format	Location
1	33	Female	Ghana	10 months	8	9	ELS	After P1
2	36	Male	Colombia	5 years	8	8	CP	After P2
3	31	Male	Saudi Arabia	Non specified	3	5	ELS	After P3
4	36	Male	Iraq	1 year	7.5	7.5	CP	After P1
5	28	Male	Libya	Non specified	4	5	ELS	After P2
6	31	Female	Libya	7 months	6	6.5	CP	After P3
7	20	Male	Angola	8 months	6	7.5	ELS	After P1
8	26	Female	Belarus	2 years	8.5	8	CP	After P2
9	33	Female	China	10 months	6.5	7.5	ELS	After P3
10	26	Male	Libya	10 months	6.5	6	CP	After P1
11	19	Male	Angola	8 months	6	6.5	ELS	After P2
12	32	Female	China	2 years	8.5	7.5	CP	After P3
13	28	Female	China	10 months	6.5	6.5	ELS	After P1
14	37	Female	Iraq	6 months	6	5.5	CP	After P2
15	29	Male	Libya	6 months	3.5	4.5	ELS	After P3
16	31	Female	China	1 year	6.5	6	CP	After P1
17	21	Female	China	10 months	6.5	6.5	ELS	After P2
18	29	Female	China	10 months	7	7	CP	After P3

All of the questions produced by the band 4 candidate contain grammatical errors in relation to question formation and this provides a clear distinction to the questions produced by the band 9 candidate. We should note that the examiner is, nonetheless, able to understand and respond to all of the questions produced by the band 4 candidate. The questions produced by the band 9 candidate are not all formed using classical question structures, but rather have quite an idiomatic and conversational quality to them.

Looking at the lists of candidate-generated questions in relation to band scores, it is difficult to pick out definite differences in question formation when comparing candidates from adjacent bands. However, when comparing the questions of the band 4 and band 9 candidates above, we can conclude that the characteristics of the candidate questions correspond well to the anticipated features for grammatical range and accuracy in the band descriptors. There is, therefore, some tentative initial evidence that candidate question formation in the CQ section may vary in relation to allocated grade in the expected manner.

If the CQ section were to be added to the IST in future, some consideration would need to be given as to how exactly candidate questions would be evaluated, and whether this might involve amendments or additions to the band descriptors. To what extent is grammatical correctness important in question formation? Given that many native speakers do not form questions in the traditional grammar textbook formats, but rather use a range of more conversational questioning formats, what should examiner expectations of candidates be?

3.5 Sub-question 5: Do the examiners believe that the CQ section adds any value to the IELTS Speaking Test? If so, in what way? If not, why not?

The three examiners agreed in their individual interviews that the CQ section adds value to the IST. One advantage of having an additional part to the IST is that the candidates have more opportunities to show what they know. In this part, they can use a variety of language that perhaps they would not be able to use in the current test format. Also, they felt that the fact that candidates can ask questions to the examiners 'humanises' the test and makes it more natural and authentic. In this sense, having 'two-way' interaction replicates the type of communication that candidates will have to deal with in an academic context. Also, it was useful for the examiners to be able to take some distance to evaluate the candidate's speaking abilities more accurately.

In the focus group, the examiners agreed that **the ELS format** fostered more naturalistic responses from the candidates whereas **the CP format** seemed to be more restrictive of natural communication, as prompts were provided. They thought that **the ELS format after part 2** was empowering for the students as they were given the opportunity to interact more with the examiner and show their ability to communicate in a conversation. They also indicated that they felt odd saying any sentence that came to mind. This, they explained, was because they did not have any clear guidelines on what they should or should not say. Therefore, clarification of the implementation of the ELS should be considered in IELTS training and for examiners, if the format were to be implemented in the test. They also recommended that examiners are trained regarding how long they should talk for in this format and how to deal with possible questions from candidates (e.g. inappropriate queries).

The CQ section is also helpful to promote the candidate's confidence, as they seemed to have a positive reaction to it. They were more relaxed and motivated to ask the examiners questions. The examiners said that the personality and background culture of the candidate might influence how comfortable they feel questioning the examiners. Nevertheless, they explained that regardless of their proficiency level, the candidates were able to produce questions. This provided the examiners with more evidence to better assess the candidates and assign them the proper score. In the case of low scores, the CQ section gave the examiners the opportunity to identify the ability of the candidate to form questions. On the other hand, with the high scores, it was a good way to confirm the score.

Examiners' comments on the advantages of the CQ section were as follows:

> Examiner 1: 'Advantages are, it, the learner can have a further opportunity to, to show what language they have, um you know there's four parts as opposed to three, also it gives, some people speak in different ways, some people are very good at um monologue, some people are very good at talking about abstract questions. Some people are much better at personal communication and are genuinely interested in other people so, I dunno, the term escapes me at the moment but I think it would offer something to those people you know? Which the present test doesn't. It would make my, on a selfish level, it would make the speaking test a lot more interesting.'

> Examiner 2: 'It seemed the, the students seemed to respond to it quite well. Erm, er it gives them confidence, erm, it does in some cases, particularly I thought in fluency and coherence give you ability to assess more.'

> Examiner 3: 'Advantages definitely to hear, to, to have more natural interaction than is currently, because most of the, the, um test as it stands, is not natural. Um it's, you know part one, asking questions of the candidate, you don't react at all, you're not allowed to react in any way. Erm, a long turn, where again you can't interact in any way, um, so there's only the part three at the moment where, where there is that, but even so that's examiner-led, it's not really, where the candidate's waiting ...'

Regarding disadvantages of the CQ section, the examiners mainly suggested that it would be helpful to have clearer instructions to guide the candidates across the task. Also, it would make the test longer which might, for example, cause difficulties for novice examiners.

Examiners' comments on disadvantages:

> Examiner 1: 'Maybe um a kind of logistical thing.'

> Examiner 2: 'So if you're adding two minutes onto it that takes it to 16 minutes. You've usually got a 20 minute window...so for experienced examiners, I don't see that being an issue, but for new examiners it might be problematic because it would reduce the number of, the amount of time you have to listen back, um to the recording, which you can do. Erm, but largely other than, other than the time, um I can't see any particular disadvantages.'

> Examiner 3: 'If the instructions were clear that wouldn't be a disadvantage. Um it's, obviously it adds on time to the erm, but, so that could be a disadvantage.'

The examiners also made some recommendations which may help to improve the CQ section. For example, for the CP format, they said that images might work better than prompts, as the candidates would be less tempted to paraphrase the statements. Also, the candidates should be well prepared to follow-up the examiners' responses, which are unpredictable. Also, the instructions in the cards should be clearer so the candidates are not confused with what they have to do. Training should be provided to the examiners in order to know how to respond (e.g. in the case of personal questions) and for how long.

Examiners' comments on recommendations for the CQ section were as follows:

> Examiner 1: '[CP] It's just too, it's text, it's textual, it's they have to read it too carefully, I think, it's something more that just hits you, it could visual maybe, I can't, I'm just, I can't think of any practical solution right now.'

> Examiner 2: 'It's really important to get that smooth transition so that students really are confident in what they're doing.'

Examiner 3: 'The rubrics weren't clear and the, you know, so there's no scripted instructions, so there was some kind of hesitancy at times where I thought oh, I've got to explain that more clearly or the candidate wasn't sure when to start, that kind of thing, but that, so it, that's to do with just scripted instructions.'

In summary, the examiners felt that the CQ section did indeed produce more two-way, naturalistic interaction. They all felt that the rubrics or frames provided were not clear enough, and development work is certainly needed.

3.6 Sub-question 6: Do the candidates believe that the CQ section adds any value to the IELTS Speaking Test? If so, in what way? If not, why not?

The 18 candidates who took the test in this study suggested that the additional part was useful. Some of the advantages they stated had to do with, for example, being in control or having a certain power in a test that is usually very restrictive. The flow of interaction was more natural, as they felt it was a more authentic task and there was less distance between them and the examiner. This made them feel more relaxed and confident to ask the questions. The additional section was also an opportunity for them to display their knowledge of the language and improve their score. Most of the candidates indicated that they felt relaxed in this new part of the test. They also highlighted the fact that the roles were 'reversed' so it felt they were 'in charge' of the evaluation. In addition, candidates felt that there was less 'distance' between them and the examiners, usually seen as an 'authoritative figure', and that this helped them to be more confident during the evaluation. The candidates also revealed some challenges during the tests. For example, they felt out of their comfort zone as they did not have to ask questions in the original test. Also, they did not know how to react to some of the examiners' responses or how to make follow-up questions when their answers were rather short. (See candidates' comments below.)

The candidates also made reference to their own cultural background, as sometimes it felt strange for some of them to ask questions to the examiner, usually seen as the authority in the test. Also, they refer to the type of questions they should ask, as in, for example, what kind of questions they are allowed to ask (e.g. personal life, academic life). They also stated that it was difficult for them to know if their question was correct or not, if they should ask more or less. For this, the candidates suggested they should be prepared for the additional section, but do not foresee any difficulty for it as long they are told what is expected from them.

The candidates did not make clear statements about which location was better, however, some of them indicated that they felt quite comfortable with the section after part 3. They explained that the location had to do with how warmed-up they are, so probably they felt more relaxed to ask the questions at the end of the test.

Candidates' comments on the CQ section were as follows:

Candidate 1 (score: 9.0): 'Well I think at least, erm, it sort of gives you some power, you're not, you're not only there as erm, answering questions, and then you get to ask the some, the person who's asking you.'

Candidate 15 (score 4.5): 'don't difficult. But er need some practice and er need how you can ask this er examiner, but this very [diffi] er very very good for er for study and for er for me.'

Candidate 3 (score: 5.0): 'er, I don't know for it's er for me. It's the answer's correct or no.'

Candidate 6 (score 6.5): 'Yeah I, I I think it's OK. It's not so difficult, it's not to easy, um it's OK.'

Candidate 9 (score 7.5): 'I felt more stressed than when talking on my own, I think.'

Candidate 12 (score 7.5): 'Because when when I was told to, ask the questions, ask the same questions, based on the same given topic [so, or], it's, I repeated it's as I, I need to be clear whether it's just we ask something like the same thing, [they] said yeah, it's just your turn.'

Candidate 16 (score 6.0): 'When I change the roles, is kind of ooh how could I could ask the question, and the, no, er they ask the question should be reasonable.'

On the whole, the candidate interviews revealed a generally positive view that the CQ section changed the balance of power in the IST, with a more natural flow of interaction and a more authentic task. However, some did not feel properly prepared, some felt challenged and some felt the task to be culturally alien to them.

4 Conclusions

4.1 Answer to the main question

We are now in a position to answer the main question:
Does the new Candidate Question (CQ) section generate candidate-led question-answer sequences as anticipated, and if so, does this add value to the IST?

The clear answer is that it does generate such sequences successfully, even with weak candidates. CA analysis of extracts shows that it is also able to generate a variety of other speech moves as well. The CQ section does add value in a number of ways, according to both examiners and candidates. According to both groups, it creates a context for naturalistic, two-way communication. Candidates explained in their interviews that this new part was an extra opportunity for them to show their knowledge of the language. For the examiners, it allowed them to have an extra perspective on how the candidates used the language and this helped them in their ratings. It clearly makes the IST, as a whole, less one-sided and allows all candidates the opportunity to ask questions. CA analysis showed that higher-scoring candidates also take the opportunity to assume a more active role to develop topic and to make other kinds of speech moves, thereby escaping the question-answer lockstep and becoming more like the kind of two-way discussion originally intended to occur in part 3. In such cases, interaction bears a closer resemblance to small-group interaction in universities.

The CQ section is, therefore, potentially capable of providing extra rating data to examiners which can differentiate between stronger and weaker students and generate a cluster of assessable features. Examiners felt that candidate questions provided useful extra information for rating purposes. It also has some potential disadvantages, as detailed below.

4.2 What are the potential advantages of an additional CQ section?

The potential advantages are that a CQ section would:

1. allow a part of the IST to have a closer correspondence with interaction in university small group settings, in which students are encouraged to ask questions and develop topics.

2. provide examiners with more and different evidence to better evaluate the candidates. In the case of low scores, the CQ section gave the examiners the opportunity to identify the ability of the candidate to form questions. On the other hand, with the higher bands, it was a good way to confirm the score.

3. change the social dynamics of the IST for the better, with candidates reporting less distance from the examiners, who reported that it seemed to 'humanise' the test.

4. generate examples of more two-way, naturalistic interaction which would give candidates the chance to take a more active role and to develop topic in a different way. Originally in the IST, part 3 was intended to generate 'two-way interaction', but no evidence was found in the corpora of the current or any of the previous studies (Seedhouse & Egbert, 2006; Seedhouse & Harris, 2011; Seedhouse et al., 2014) that this was achieved in a regular, widespread way. The reasons for this are discussed in Seedhouse & Harris (2011). Although part 3 is termed 'two-way discussion', it is almost identical to part 1 interactionally, in that it consists of a series of topic-scripted question-answer adjacency pairs dominated by the examiner. There are hardly ever any opportunities for candidates to introduce or shift topic and they are generally closed down when they try to do so. The topic-scripted question-answer adjacency pair seems to be such a strong attractor in this setting that this would be difficult to achieve. Like the ubiquitous IRF/ IRE pattern in classroom discourse, the topic-scripted question-answer sequence is the most economical method of carrying out a single cycle of institutional business. This means that any examiner question sequence would be likely to end up reverting to the archetype, no matter how much one tried to make it resemble two-way discussion. It was, therefore, argued that a more feasible way of ensuring two-way interaction in the IST would be by having the candidate lead the interaction by asking questions to the examiner. The evidence in the current study supports this. Higher-scoring candidates took the opportunity to assume a more active role to develop topic and to make a variety of other kinds of speech moves, thereby escaping the question-answer lockstep and becoming more like the kind of two-way discussion originally envisaged.

4.3 What are the potential disadvantages of an additional CQ section?

If it were part of a 4-part IST to include the existing 3 parts, the CQ section would be likely to increase the duration of the IST by at least two minutes. As implemented in this study, the average length of the CQ section was 2.32 minutes; see Table 4. The average length of the revised 4-part IST in this study was 14.34 minutes, which is just beyond the normal 14 minute limit for ISTs. However, as these were mock ISTs, the initial administrative procedures which take place at the start of real ISTs were omitted, so their length cannot be directly compared to the length of genuine ISTs.

This study tends to suggest that the addition of an additional CQ section would lengthen the IST by 2½ minutes. However, since both examiners and students reported on the impact of the unfamiliarity of the new CQ section, it is possible that its duration would decrease as it became more familiar to both parties and normalised over time.

It is noticeable in the data that some candidates display a degree of hesitation and uncertainty in relation to what is expected of them and when to start in the CQ section. Some examples can be seen in the extracts above. However, it does need to be noted that, at the start of the CQ section, the examiner handed candidates cue cards to read which they have not seen before. Therefore, some degree of hesitation and some length of pause is to be expected before the first question is asked by the candidate.

Another possible disadvantage is that, in certain cultures, candidates may have certain reservations about the appropriateness of them asking questions to teachers/examiners, as this may be discouraged and indeed sanctionable in their own societies. In the candidate interviews, some participants reported feeling strange about questioning an authority figure and uncertainty about what it was permissible to ask. Piloting of CQs in a range of cultures and countries might, therefore, be advisable. The candidates taking part in this study do represent a range of cultures (see Table 2) and did not overtly display in the interaction much reluctance to asking questions to the examiners. So, there was no interactional evidence that reported reservations about asking questions actually resulted in candidates being unable to deliver questions in practice. However, they had all been in the UK studying for some time (see Table 2), so they may have become more acclimatised to British HE educational culture. We should also note that candidate questions did actually feature in the pre-2001 original IST without cultural concerns being reported as a serious issue. Furthermore, candidates entering Western universities would have to get used to asking questions to authority figures anyway, so it is arguably a relevant task from the perspective of cross-cultural integration.

Taylor (2011, vi) reported in relation to candidate questions (in the original pre-2001 IST) a concern that these would result in significant variations in amounts and type of examiner talk. The extracts above do indeed show variations in amounts and type of examiner talk in the CQ section, although it is not clear how significant these variations are. If, however, the aim is to have a section with more two-way, naturalistic interaction, then this does imply relaxing controls and de-standardising the interaction in order to escape the question–answer lockstep, which, in turn, implies that there will be variation and heterogeneity in talk. So there seems to be something of a paradox at work. In order to have an authentic task which generates naturalistic, two-way interaction, it appears that it is necessary to use a less scripted format. However, this means that the interaction will be less predictable and less standardised, making it more difficult to ensure the validity of assessment. Nevertheless, examiners did not report problems in grading the CQ section interaction, and analysis of the interaction shows that it enabled differentiation between proficiency levels. The CQ section would ensure there were three different varieties of interaction in the IST, whereas at present there are only two, in that both parts 1 and 3 are dominated by the topic-scripted QA adjacency pair sequence.

It is clear from the discussion in this section that it is feasible to add a CQ section to the IST. This offers the advantages of having more two-way and naturalistic talk, a different type of information for raters and closer resemblance to small-group university interaction. There would also be the disadvantage of increased test length. Variation in amounts and type of examiner talk would result, which could be seen as a disadvantage, although we would argue that this is necessary to achieve an element of two-way, naturalistic talk.

4.4 Recommendations

We recommend that consideration should be given, when reviewing the IST, to adding a fourth part in which candidates ask questions to examiners. The research suggests that the 'examiner leading statement' format after the existing part 2 would be optimal, although a variety of preferences were expressed by examiners. If the CQ component were to be adopted in future, we would recommend the following.

1. The interactional analysis showed that higher-level candidates are able to break out of the question-answer lockstep and produced new, additional speech moves in the CQ section. This might possibly be an effective criterion to distinguish higher from lower level candidates. If so, this might involve amendments or additions to the band descriptors.

2. Some consideration would need to be given as to how exactly candidate questions would be evaluated, and whether this might involve amendments or additions to the band descriptors. To what extent is grammatical correctness important in question formation? Given that many native speakers do not form questions in the traditional grammar textbook formats, but rather use a range of more conversational questioning formats, what should examiner expectations of candidates be?

3. The examiners all felt that the rubrics or frames provided were not clear enough, so development work by experts would certainly be needed.

4. It, therefore, follows that training and guidelines for the examiners in relation to how to respond to candidate questions would be necessary.

5. Candidates and their teachers would need preparation for the requirement for them to ask questions.

References

Benwell, B. and Stokoe, E. H. 2002. Constructing discussion tasks in university tutorials: shifting dynamics and identities, *Discourse Studies 4(4)*, 429–453.

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

Bryman, A. 2001. *Social Research Methods*. Oxford: Oxford University Press.

Drew, P. and Heritage, J. (Eds). 1992a. *Talk at Work: Interaction in Institutional Settings.* Cambridge: Cambridge University Press.

Fulcher, G. 2003. *Testing Second Language Speaking*. Harlow: Pearson Education Limited.

IELTS. 2011. *Information for Candidates Booklet.* Accessed 6 February 2011 at: www.ielts.org/pdf/information_for-Candidates_booklet.pdf

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

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

Richards, K. and Seedhouse, P. 2005. *Applying Conversation Analysis*. Basingstoke: Palgrave Macmillan.

Seedhouse, P. 2004. *The Interactional Architecture of the Language Classroom: A Conversation Analysis Perspective*. Malden, MA: Blackwell.

Seedhouse, P. 2005. 'Task' As Research Construct, *Language Learning 55 (3)*, 533–570.

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

Seedhouse, P. and Harris, A. 2011. Topic development in the IELTS Speaking Test, *IELTS Research Reports Vol 12*, 69–124. IDP: IELTS Australia and British Council.

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

Taylor, L. 2000. Issues in Speaking Assessment Research, *Research Notes 1*, 8–9. Cambridge English Language Assessment.

Taylor, L. 2001a. Revising the IELTS Speaking Test: retraining IELTS examiners worldwide, *Research Notes 6*, 9–11. Cambridge English Language Assessment.

Taylor, L. 2001b. The paired speaking test format: recent studies, *Research Notes 6,* 15–17. Cambridge English Language Assessment.

Taylor, L. 2011. Introduction to *IELTS Research Reports Vol 12*, i–xiv. IDP: IELTS Australia and British Council.

Taylor, L. (Ed). 2011. Examining Speaking: Research and Practice. In *Assessing Second Language Speaking*, Cambridge: Cambridge University Press.

Wigglesworth, G. 2001. Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan and M. Swain (Eds), *Researching pedagogic tasks: second language learning, teaching and testing*. Harlow: Pearson, 186–209.

Note: The following publications are not referenced as they are confidential and not publicly available:

- Instructions to IELTS Examiners
- IELTS Examiner Training Material, 2001
- Examiner script, January 2003
- *IELTS Handbook 2005*.