IELTS Research Reports Online Series
ISSN 2201-2982
Reference: 2014/2

The relationship between speaking features and band descriptors: A mixed methods study

Authors: *Paul Seedhouse, Andrew Harris, Rola Naeb and Eda Üstünel, Newcastle University, United Kingdom*

Grant awarded: *2012–13*

Keywords: *"IELTS speaking test, assessable speaking features, discoursal features, conversation analysis, spoken interaction, second language acquisition"*

Abstract

This study looked at the relationship between how candidates speak in the IELTS speaking test and the scores they were given. We identified the features of their talk which were associated with high and low scores.

The research focus was on how features of candidate discourse relate to scores allocated to candidates, and the overall aim was to identify candidate speaking features that distinguish proficiency levels in the IELTS speaking test (IST). There were two research questions:

1. The first noted that grading criteria distinguish between levels 5, 6, 7 and 8 in the ways described in the IELTS speaking band descriptors and asked to what extent these differences are evident in ISTs at those levels. In order to answer this research question, quantitative measures of constructs in the grading criteria were operationalised and applied to the spoken data (fluency, grammatical complexity, range and accuracy).
2. The second question asked which speaking features distinguish tests rated at levels 5, 6, 7 and 8 from each other. This question was answered by working inductively from the spoken data, applying Conversation Analysis (CA) to transcripts of the speaking tests. The dataset for this study consisted of 60 audio recordings of IELTS speaking tests. These were transcribed, giving a total of 15 tests for each of the score bands (5, 6, 7, 8).

The quantitative measures showed that accuracy does increase in direct proportion to score. Grammatical range and complexity was lowest for band 5, but band 7 scored higher than band 8 candidates. The measure of fluency employed (pause length per 100 words) showed significant differences between score bands 5 and 8. The qualitative analysis did not identify any single speaking feature that distinguishes between the score bands, but suggests that in any given IELTS speaking test, a cluster of assessable speaking features can be seen to lead toward a given score.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2014.
This online series succeeds *IELTS Research Reports Volumes 1–13*, published 1998–2012 in print and on CD.

Web: www.ielts.org

AUTHOR BIODATA

Paul Seedhouse

Paul Seedhouse is Professor of Educational and Applied Linguistics in the School of Education, Communication and Language Sciences at Newcastle University, UK. His research is in spoken interaction in relation to language learning, teaching and assessment. He has published widely in journals of applied linguistics, language teaching and pragmatics. His book, *The Interactional Architecture of the Language Classroom: A Conversation Analysis Perspective,* was published by Blackwell in 2004 and won the 2005 Kenneth W Mildenberger Prize of the Modern Language Association of the USA.

Andrew Harris

Andrew Harris took a PhD at Newcastle University and is now a Lecturer in Applied Linguistics and TESOL in the Department of Languages, Information and Communications at Manchester Metropolitan University, UK. His primary research focus is on the micro-analysis of spoken interaction in institutional contexts, specifically in education, teacher education and assessment. He also has many years of experience as a language teacher, teacher trainer and school manager.

Rola Naeb

Rola Naeb took her PhD at Newcastle University and is now a Lecturer in Applied Linguistics and TESOL at Northumbria University, UK. Her main research interests lie in the fields of Applied and Educational Linguistics and Technology. She is particularly interested on the applicability of second language acquisition findings to technology-enhanced language learning environments. Her current work focuses on expanding models and creating tools to facilitate language learning in traditional and technology-enhanced environments.

Eda Üstünel

Eda Üstünel has been teaching at the Department of English Language Teacher Training, Faculty of Education at Muğla Sıtkı Koçman University (Turkey) since 2004. She received her MA degree (2001) in Language Studies at Lancaster University, UK, and her PhD degree (2004) in Educational Linguistics at Newcastle University, UK. Her research is in spoken interaction in relation to language learning and teaching at young learners' classroom. She has presented papers at international conferences and published her research at international journals. She was a Visiting Lecturer at Newcastle University from March to May 2013.

IELTS Research Program

The IELTS partners, British Council, Cambridge English Language Assessment and IDP: IELTS Australia, have a longstanding commitment to remain at the forefront of developments in English language testing.

The steady evolution of IELTS is in parallel with advances in applied linguistics, language pedagogy, language assessment and technology. This ensures the ongoing validity, reliability, positive impact and practicality of the test. Adherence to these four qualities is supported by two streams of research: internal and external.

Internal research activities are managed by Cambridge English Language Assessment's Research and Validation unit. The Research and Validation unit brings together specialists in testing and assessment, statistical analysis and item-banking, applied linguistics, corpus linguistics, and language learning/pedagogy, and provides rigorous quality assurance for the IELTS test at every stage of development.

External research is conducted by independent researchers via the joint research program, funded by IDP: IELTS Australia and British Council, and supported by Cambridge English Language Assessment.

Call for research proposals

The annual call for research proposals is widely publicised in March, with applications due by 30 June each year. A Joint Research Committee, comprising representatives of the IELTS partners, agrees on research priorities and oversees the allocations of research grants for external research.

Reports are peer reviewed

IELTS Research Reports submitted by external researchers are peer reviewed prior to publication.

All IELTS Research Reports available online

This extensive body of research is available for download from www.ielts.org/researchers.

INTRODUCTION FROM IELTS

This study by Paul Seedhouse and his colleagues at Newcastle University, UK 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 British Council and IDP: IELTS Australia under this programme complements those conducted or commissioned by Cambridge English Language Assessment, and together they 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 100 empirical studies having received 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*. To date, 13 volumes of *IELTS Research Reports* have been produced. But as compiling reports into volumes takes time, individual research reports are now made available on the IELTS website as soon as they are ready.

The IELTS speaking test has long been a distinctive aspect of the exam and the focus of much IELTS-funded research (e.g. Brown, 2003; Taylor and Falvey, 2007; Wigglesworth and Elder, 2010). The present study is the latest in a series by Seedhouse and his colleagues investigating and describing the speaking test using Conversation Analysis methodology. The first one (Seedhouse and Egbert, 2006) looked into the nature of interaction in the test, and the second one (Seedhouse and Harris, 2011) investigated the role played by topic in shaping that interaction. They now take that work one step further, using a mixed methods approach to compare observed interaction features with the scoring criteria for the test.

For this study, the researchers analysed 60 transcribed IELTS speaking tests, with an equal number of performances from each of bands 5, 6, 7 and 8. Findings from ANOVA were generally in the expected directions. The stronger the candidate, the more words they produced, the fewer grammatical errors they made, and the shorter their pauses. These reflect directly or indirectly the criteria in the IELTS speaking band descriptors.

On the other hand the Conversation Analysis, looking in greater detail at the data, not unexpectedly introduced some complexity into the picture. For example, pauses can indicate a lack of lexical resource on the one hand, but can be a resource for holding the floor on the other. That being the case, performance features tend not to have a straightforward one-to-one relationship with score outcomes. Also, the analysis identified performance features not in the scoring criteria but which nevertheless could conceivably impact on score outcomes, e.g. using one's responses to construct an identity as "hard-working cultured intellectuals and (future) high achievers", which appears to be associated with higher band scores. The researchers therefore conclude that no single speaking feature can distinguish candidates across band scores, but rather, that clusters of features predict score outcomes, which include features not mentioned in the scoring criteria.

Now this might, at first blush, appear to be problematic, as it seems to imply that candidates are not being scored according to the band descriptors. But this is actually as the literature predicts it would be (Lumley 2005). Examiners observe a large number of features about any given performance and, left unconstrained, would lead towards unreliable score outcomes. But band descriptors cannot describe every feature that an examiner might observe. (It would also be quite pointless if they did, because they would simply replicate examiners' observations.) It thus becomes apparent that band descriptors are necessarily selective in what they highlight, so that examiners' myriad observations can be channelled in order to produce the institutional goal of more reliable, if less detailed, summative outcomes.

In any case, while on the topic of examiners, the researchers identified quite a few features that they hypothesise could affect score outcomes, which can only be confirmed by conducting research with examiners, perhaps using think-aloud protocols, in order to determine the extent to which they notice the same features and how much these features impact upon their scoring decisions. That would be the logical next study in this series of research, which we look forward to seeing.

Dr Gad S Lim
Principal Research and Validation Manager
Cambridge English Language Assessment

References to the IELTS Introduction

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

Lumley, T, 2005, *Assessing second language writing: The rater's perspective*, Frankfurt am Main: Peter Lang

Seedhouse, P, and Egbert, M, 2006, The interactional organisation of the IELTS speaking test, *IELTS Research Reports Vol 6*, IELTS Australia and British Council, Canberra, pp 161-206

Seedhouse, P, and Harris, A, 2011, Topic development in the IELTS speaking test, *IELTS Research Reports Vol 12*, IDP: IELTS Australia and British Council, Melbourne, pp 69-124

Taylor, L, and Falvey, P (eds), 2007, *IELTS collected papers: Research in speaking and writing assessment,* Cambridge: Cambridge ESOL/Cambridge University Press

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

TABLE OF CONTENTS

List of tables and figures

1 RESEARCH DESIGN

1.1 Background information on the IELTS speaking test

IELTS speaking tests 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.

- **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 the candidate using verbal questions selected from familiar topic frames. This part lasts between four and five minutes.

- **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.

- **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 that standardisation plays a crucial role in the successful management of the test. "The IELTS speaking test involves the use of an examiner frame which is **a script that must be followed** (original emphasis)… 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 written out in Parts 1 and 2 of the test so that all candidates receive similar input phrased in the same manner. In Part 3, the examiner frame is less rigid so that the examiner has the freedom to adjust to the level of the candidate. Examiners should not make unscripted comments. Detailed performance descriptors have been developed which describe spoken performance at the nine IELTS bands, based on the criteria listed below (IELTS Handbook, 2005, p 11).

Fluency and Coherence refers to 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. For coherence, the key indicators are logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices (eg connectors, pronouns and conjunctions) within and between 'sentences'.

Lexical Resource refers to 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 refers to 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 refers to 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.

The IELTS speaking band descriptors are available in Appendix 4. In this project, only the constructs of Fluency, Grammatical Range and Accuracy were investigated.

1.2 Research focus and questions

The research focus is on how features of candidate discourse relate to scores allocated to candidates, and the overall aim is to identify candidate speaking features that distinguish IELTS proficiency levels in the IELTS speaking test (IST). There are two research questions:

1) The grading criteria distinguish between levels 5, 6, 7 and 8 in the ways described in the speaking band descriptors (see Appendix 4). To what extent are these differences evident in tests at those levels?

In order to answer this research question, quantitative measures of constructs (fluency, grammatical complexity, range and accuracy) in the band descriptors are applied to the spoken data.

2) Which speaking features distinguish tests rated at levels 5, 6, 7 and 8 from each other?

This question is answered by working inductively from the spoken data, applying Conversation Analysis (CA) to transcripts of the speaking tests.

1.3 Relationship to existing research literature

This study builds on existing research in two areas. Firstly, research which has been done specifically on the IST, as well as on oral proficiency interviews (OPIs) in general. Secondly, it builds on existing research into the specific issue of how features of candidate discourse relate to scores allocated to candidates. The first of these areas is historically represented by a broad range of research methodologies, approaches, and interests, from investigations into test taker characteristics to cognitive, scoring and criterion-related validity (Taylor, 2011).

However, the interest in the relationship between candidate speaking features and their scores did not came to the fore until the late 1980s, as researchers turned to the question of the authenticity of OPIs (Weir et al, 2013). This interest was initiated in part by van Lier's (1989) now seminal call to investigate the interaction which takes place in the OPI. Nonetheless, according to Lazaraton (2002, 161) there has still "been very little published work on the empirical relationship between candidate speech output and assigned ratings". It is important to know how candidate talk is related to scores for a number of reasons. Test developers may use discourse analysis of candidate data as an empirical basis to develop rating scales (Fulcher, 1996; 2003). Similarly, evidence of the relationship between candidate talk and grading criteria can provide valuable input for validation processes.

Douglas's (1994) study of the AGSPEAK test related candidate scores to the categories of grammar, vocabulary, fluency, content and rhetorical organisation and very little relationship was found between the scores, given and candidate discourse produced. Douglas suggests this may have been due to inconsistent rating or raters attending to aspects of discourse which were not on the rating scale. Brown (2006a) developed analytic categories for three out of the four rating categories employed in the IST and undertook quantitative analysis of 20 ISTs in relation to these analytic categories. While she found that, in general, features of test-takers' discourse varied according to their proficiency level, there was only one measure which exhibited significant differences across levels, which was the total amount of speech. Her overall finding (2006a, 71) was that "while all the measures relating to one scale contribute in some way to the assessment on that scale, no one measure drives the rating; rather a range of performance features contribute to the overall impression of the candidate's proficiency". Brown's study identified a number of discourse features in advance and then searched for these in the ISTs in her sample, using a quantitative approach. Young (1995) also took a quantitative approach to a comparison of different levels of candidates and their respective speaking features (in the First Certificate in English), and found that the high-level candidates produced more speech at a faster rate, and which was more elaborated, than those at the lower level.

Other researchers have applied qualitative methodologies to OPI talk. Lazaraton (2002) presents a CA approach to the validation of OPIs, suggesting that qualitative methods may illuminate the process of assessment, rather than just its outcomes. Lazaraton's (1998) study of the previous version of the IST examined 20 tests and compared the relationship between candidate talk and ratings. Findings were that: there are fewer instances of repair at higher levels; higher scoring candidates use a broader range of expressions to speculate; grammatical errors are more common in lower bands and complex structures in higher bands; and appropriate responses are more common in higher bands, as is conversational discourse.

Seedhouse and Harris's CA (2010) study of the IST found that the characteristics of high scoring and low scoring tests in relation to topic are as follows.

Candidates at the higher end of the scoring scale tend to have more instances of extended turns in which topic is developed in parts 1 and 3. There is some evidence that very weak candidates produce short turns with lengthy pauses in part 2. There does appear to be a correlation between test score and occurrence of trouble and repair: in interviews with high test scores, fewer examples of interactional trouble requiring repair are observable. This confirms Lazaraton's (1998) finding in relation to the previous version of the IST. Candidates gain high scores by engaging with the topic, by expanding beyond minimal information and by providing multiple examples, which enable the examiner to develop the topic further. Candidates with low scores sometimes struggle to construct an argument and a coherent answer. High-scoring candidates develop the topic coherently, using markers to connect clauses. Candidates with a high score may develop topic using lexical items which are less common and which portray them as having a higher level of education and social status. Candidates who achieved a very high score typically developed topics that constructed the identity of an intellectual and a (future) high-achiever on the international stage. Candidates with low scores, by contrast, developed topics in a way that portrayed them as somebody with modest and often localised aspirations. Examiners may take several features of monologic topic development into account in part 2.

Seedhouse and Harris (2010) suggest that in parts 1 and 3 of the IST, there is an archetypal organisation which combines turn-taking, adjacency pair and topic, as follows. Examiner questions contain two components: a) an adjacency pair component, which requires the candidate to provide an answer; and b) a topic component, which requires the candidate to develop a specific topic. This organisation may be called a 'topic-scripted question-answer (Q-A) adjacency pair'. So in the IST, unlike conversation, topic is always introduced by means of a question. In order to obtain a high score, candidates need to do the following: a) understand the question they have been asked; b) provide an answer to the question; c) identify the topic inherent in the question; and d) develop the topic inherent in the question. This core interactional structure therefore generates multiple means of differentiating high- and low-scoring responses. Whereas topic development is mentioned in the band descriptors (Fluency and Coherence), candidate ability to answer the question is not; we revisit this issue in section 2.1.1.

The overall picture from the research literature is that there is a great deal still to be learnt in respect of speaking features that distinguish IST proficiency levels. There is no simple relationship between the candidate's score and features of their interactions, since a multitude of factors affect the examiner's ratings (Brown, 2006a, 71; Douglas, 1994, 134). Some studies have pre-specified discourse features and searched for these in the data using quantitative techniques, whereas Seedhouse and Harris (2010) looked inductively in the data for differences using a qualitative approach. However, no studies have so far tried to combine both of these approaches using a mixed methods design.

1.4 Methodology

This study employs a mixed methods approach that "combines elements of qualitative and quantitative research approaches … for the broad purposes of breadth and depth of understanding and corroboration" (Johnson et al., 2007, 123). The benefit of this methodology is that it provides a two-pronged approach to the overall aim of identifying speaking features that distinguish IST proficiency levels. The two sets of analyses were carried out concurrently and independently of each other. For the first question, Rola Naeb carried out the quantitative analysis of the dataset. For the second, Andrew Harris carried out the qualitative (CA) part, and it was not until the final stage of the project that we merged the results of the two methodological strands. In doing so, we treated the two datasets and their merging as an opportunity to "explore the potential of different perspectives on the research process" (Richards et al., 2012). The mixed methods design also approaches the data from two different directions. The first starts with the grading criteria and operationalises the concepts of fluency, grammatical complexity, range and accuracy to permit coding of a corpus of transcripts at the four bands. The second starts from the data (audio recordings and transcripts) and attempts to distinguish in an inductive fashion any differences in speaking features in test performances at the four levels.

The first research question asked: the grading criteria distinguish between levels 5, 6, 7 and 8 in the ways described in the speaking band descriptors in terms of: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. The question is: *To what extent are these differences evident in tests at those levels*? A matching methodology was used to answer this research question. The descriptors (see Appendix 4) anticipate the differences which will emerge in ISTs at these different levels. The descriptors were operationalised and matched against the evidence in the recordings and transcripts.

Given the restricted scope and budget of the project, we investigated only the descriptors for Fluency, Grammatical Range and Accuracy by adapting standard tests for these constructs (Ellis and Barkhuizen, 2005). This approach was thought suitable for this research question because it employs standard measures which have previously been shown to provide valid measurement of the constructs targeted here. To assess accuracy, we used the number of errors per 100 words (Mehnert, 1998). To assess grammatical range, two different measures were adapted, as both grammatical range and complexity are constructs employed in the band descriptors. For grammatical complexity, we adapted Foster et al's (2000) measure of the amount of subordination. We adapted Yuan and Ellis's (2003) measure of the number of different verb forms used to access the range of structures employed. To assess fluency, Skehan and Foster's (1999) measurement of pause was employed, but adapted to become pause length per 100 words.

Like Brown (2006a, 74) and numerous other studies, the research team experienced great difficulty in adapting constructs and measures originally developed for L1 written texts to the analysis of L2 speaking data in a test setting. We now explain how we adapted and operationalised these measures. Grammatical range in terms of complexity was measured in terms of subordination, and Foster et al's (2000) concepts were adapted to code the transcripts. The total number of clauses and subordinate clauses was calculated based on Foster et al's (2000) operationalisation of AS units. However, they did not describe fully how they operationalised these in relation to unit boundaries, hesitation markers, etc. To ensure inter-rater reliability (IRR), three workshops were carried out where two raters coded a transcript independently and the numbers of AS units were compared. In the first two workshops, IRR was not satisfactory and therefore further sets of rules were developed to cover areas where divergence occurred. A full list of the rules produced is provided in Appendix 1. Complexity was therefore measured using two sub-measures: the ratio of A units (subordinate clauses) to AS units and the ratio of A units to total number of words. We adapted Foster et al's (2000) system because we noted when coding the transcripts that some candidates used many main clauses within an AS unit without any A units. In the final workshop, inter-rater reliability of 90% was achieved and considered satisfactory.

Grammatical range in terms of variety was measured by adapting Yuan and Ellis's (2003) measure of syntactic variety, in order to access the range of structures employed. Yuan and Ellis (2003,13) state that they measured the total number of different grammatical verb forms used, specifically tense, modality and voice. Since that study focused on planning in relation to oral narrative tasks, we adapted the measure for the IST by providing a list (Appendix 2) of all of the verb forms targeted by the IST, using as source, the IELTS grammar preparation book by Cambridge University Press. Hopkins and Cullen (2007, vii) state that "this book covers the grammar you will need to be successful in the test". We deployed this measure by counting the first time only that one of the verb forms was used accurately by the candidate. Two workshops were carried out to ensure IRR and in the second workshop, a score of 80% was achieved.

Defining and operationalising the concept of fluency is a thorny issue (Luoma, 2004, 88), not least because the host of definitions available across the literature refer to a plethora of aspects attributed to fluency: speech rate, flow, smoothness, absence of pausing and hesitation markers, connectedness and length of utterances (Koponen, 1995). Within this study, we adapted Skehan and Foster's (1999) measures of candidate pause length. Any intra-turn candidate pause beyond the threshold of 0.5 seconds was measured and collated, to give an overall score for the candidate's fluency. In the first workshop, the IRR rate was 98.9%. We finally measured fluency as pause length per 100 words after noting in the data that the total number of words produced increased in direct proportion to score.

To assess accuracy in this study, a combination of two measures was employed: a total word count produced by the candidate per test, and the total number of errors produced by the candidate during the test. Accuracy was

therefore calculated as a function of how many errors candidates produced per 100 words (Mehnert, 1998). Although candidate errors that were self-corrected were not included in the count, this does not remove the intrinsic issues (for the analyst) of determining what should be counted as an error (Ellis and Barkhuizen, 2005), particularly when these measures are applied to spoken interactional data. In the first workshop, IRR rates were as follows: Wordcount 98.4%: Errors 87.4%.

The second research question set out to identify the speaking features that distinguish tests rated at levels 5, 6, 7, and 8 from each other. To answer this, the methodology employed was Conversation Analysis (CA) (Lazaraton, 2002: Seedhouse, 2004; Young & He, 1998). This methodology is suitable for two reasons. Firstly, CA institutional discourse methodology attempts to relate the overall organisation of the interaction to the core institutional goal, so we need to focus on the rational design of the interaction in relation to language assessment. Secondly, analysis is bottom-up and data driven; we should not approach the data with any prior theoretical assumptions or assume that any background or contextual details are relevant.

The first stage of CA analysis has been described as *unmotivated looking* (Psathas, 1995) or being open to discovering patterns or phenomena, rather than searching the data with preconceptions or hypotheses as to what the speaking features are that distinguish different levels. After an inductive database search has been carried out, the next step is to establish regularities and patterns in relation to occurrences of the phenomenon and to show that these regularities are methodically produced and oriented to by the participants. After the *unmotivated looking* phase of the analysis, the focus turned to analysing the dataset, in order to answer the second question. A number of approaches to this task were taken that included treating the various score bands as individual collections, looking for patterns and trends of individual speaking features, their occurrence, and their distribution *within* bands. We also focused on particular speaking features, and analysed their occurrence *across* various speaking bands. The attempts to identify speaking features that distinguish between score bands relied, in part, on the employment of *informal quantification* (Schegloff, 1993, 100). Here, terms such as 'commonly', 'overwhelmingly' and 'ordinary' are employed to indicate the analyst's 'feel' for frequency and distribution. However, the employment of these terms within CA is *not* an attempt to formalise a quantitative analytic stance on the data. As Schegloff (1993, 118) has stated, CA and 'formal' quantification "are not simply weaker and stronger versions of the same undertaking; they represent different sorts of accounts", and in this study we employ them as such.

Much of the focus of the qualitative analysis within this project was on the ways in which candidate speaking features are incorporated into the design of their turns-at-talk. From the perspective of CA, turns-at-talk are constituted by one or more turn construction units (TCUs). TCUs can consist of a single embodied action, such as a head nod, or a stretch of talk that delivers a 'complete unit of meaning'. At the end of any given TCU is the potential for a change of speaker. These places in an unfolding turn are called transition relevance places (TRPs). At a TRP, a speaker can either select another speaker to take the floor, for example by asking a question; another speaker can self nominate and take the floor; or the current speaker can self-select and continue with their turn. The ways in which candidate turns are designed, through TCUs, will be a key element of the qualitative analysis in this study.

1.5 Data information

The dataset for this study consisted of 60 audio recordings of IELTS speaking tests. These tests include 26 that had previously been digitised and transcribed for our earlier project (Seedhouse and Harris, 2010), as well as 34 new tests, which were provided for this project, pre-digitised and edited. The new tests were selected by UCLES and send digitally to Newcastle University. The audio recordings were then transcribed, in accordance with CA's strict attention to detail and conventions, by Andrew Harris, an experienced CA transcriber and analyst. The combined dataset for the study then consisted of the audio recordings of 60 ISTs and their transcripts, giving a total of 15 transcribed tests for each of the score bands (5, 6, 7, 8+). The recordings are from the years 2004 and 2011. The transcripts were subject to quantitative measurements for the constructs of fluency, accuracy and grammatical complexity and range in relation to the first research question. The audio recordings, and a separate set of transcripts, were subject to the qualitative CA analysis in relation to the second research question. The sample consisted of 22 male and 38 female candidates. The candidates came from different L1 backgrounds as summarised in Table 1.

Language	Frequency	Language	Frequency
Tamil	1	Tagalog	12
Marathi	1	Chinese	12
Malayalam	1	Arabic	8
Bosnian	1	Thai	4
Ga	1	Spanish	4
Vietnamese	1	Kannada	2
Urdu	1	Farsi	2
Gujarati	1	English	2
Burmese	1	Korean	2
Luo	1	Other	2

Table 1: Candidates' L1 distribution

2 DATA ANALYSIS

The following sections present the analytic findings of this study. The first of these outlines the quantitative analysis (2.1). The second presents the findings of the qualitative analysis (2.2).

2.1 Quantitative analysis

2.1.1 Descriptive analysis

Table 2 shows the descriptive statistics for the four measures. Looking at the mean scores for each measure, it is evident that:

1. The total number of words per test increased in direct proportion to the scores, band by band.
2. The percentage of errors per 100 words decreased as the scores got higher, band by band. This suggests that accuracy increases in direct proportion to score.
3. The measure of pause length relates to the construct of fluency. Pause length is highest at level 5 and lowest at level 8, following the expectations set out in the IELTS descriptors. In the raw data, there is a higher level of pause at level 7 than at level 6. However, the measure of pause length per 100 words shows that fluency increased in direct proportion to the scores. Standard deviation measures show that variations within the same band decreased as score increased.
4. Both measures for grammatical complexity showed the same trend. While complexity is lowest for band 5, those at band 7 showed more complexity than those at band 8.
5. The same trend was seen in the grammatical range measure. While band 5 shows the lowest number of verb forms, those who have scored 7 used a wider range of verb forms than those at band 8.

		Accuracy		**Fluency**		**Complexity**		**Grammatical Range**
		Total no. of words	Errors per 100 words	Pause length	Pause length per 100 words	Ratio of A units to AS units	Ratio of A units to total no. of words	No. of verb forms
IELTS score 5	Mini	358.00	1.40	4.00	0.40	13.75	1.59	4.00
	Maxi	1064.00	6.65	115.30	18.96	59.65	9.50	13.00
	Mean	***762.67***	***4.05***	***28.51***	***4.21***	***29.77***	***3.42***	***7.67***
	Std. dev	227.80	1.26	29.82	4.80	12.21	1.97	2.92
IELTS score 6	Mini	654.00	1.55	0.70	0.06	22.95	2.47	5.00
	Maxi	1220.00	6.72	44.20	4.64	52.24	5.14	15.00
	Mean	***970.47***	***3.33***	***19.10***	***2.14***	***36.64***	***3.78***	***7.80***
	Std. dev	180.34	1.45	15.84	1.79	10.00	0.78	3.12
IELTS score 7	Mini	753.00	0.31	4.90	0.33	18.63	1.73	8.00
	Maxi	1591.00	2.84	70.80	5.51	152.94	9.43	20.00
	Mean	***1121.87***	***1.54***	***22.15***	***2.08***	***54.23***	***5.11***	***12.00***
	Std. dev	242.53	0.77	16.84	1.52	35.58	1.92	3.78
IELTS score 8	Mini	840.00	0.10	1.70	0.11	22.70	2.63	6.00
	Maxi	1608.00	2.19	53.80	4.61	65.35	6.76	18.00
	Mean	***1213.20***	***0.78***	***15.89***	***1.38***	***39.88***	***4.45***	***11.60***
	Std. dev	182.38	0.52	15.48	1.40	9.75	1.09	2.95

Table 2: Descriptive analysis across the four measures

2.1.2 Association between measures and score bands

In order to verify whether differences in mean scores across the four levels are statistically significant, inferential statistics were used.

2.1.2.1 Total number of words

To explore differences in relation to the amount of speech, measured as total number of words spoken by the candidate, across band scores, ANOVA was used.
It revealed that the differences were highly significant among the four groups with the amount of speech increasing with higher scores, F(56,3)= 13.18, p< 0.001

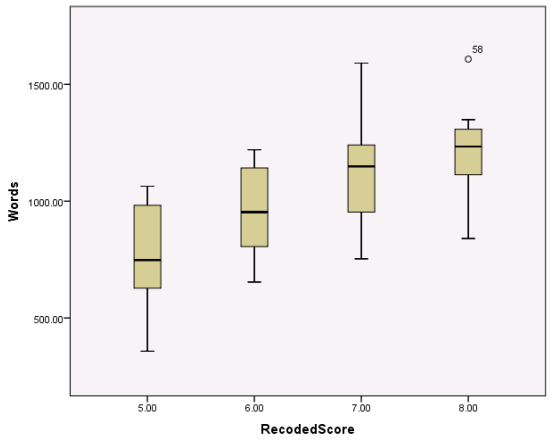

Figure 1: Total number of words ANOVA

It is obvious from the boxplot that candidates who scored 7 varied widely in the amount of speech produced with few of them producing more words than those who scored 8. However, when considering all candidates in the two groups, level 8 candidates produced significantly more words than level 7.

2.1.2.2 Accuracy

ANOVA test revealed that the difference among the four band scores were statistically significant F(56,3)= 30.6, p< 0.001.

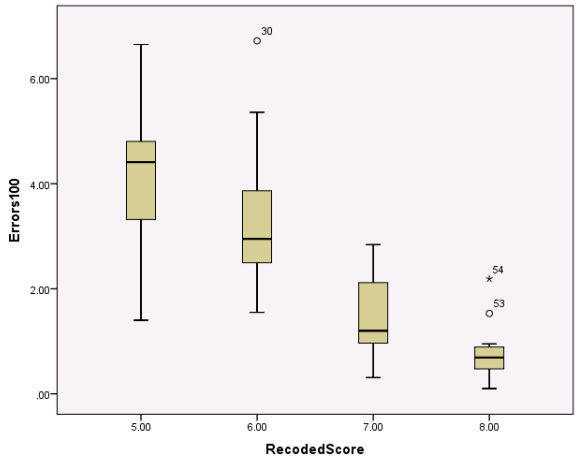

Figure 2: Accuracy ANOVA

2.1.2.3 Fluency

Looking firstly at the raw measure (pause length), the differences among the four band scores were not significant, F(56,3)= 10.4, p< 0.38

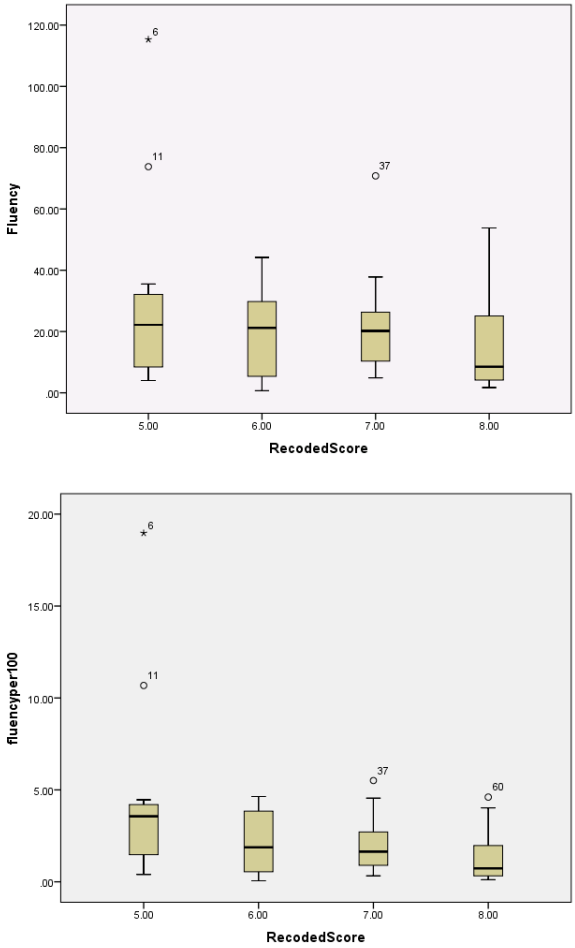

Figure 3: Pause length and pause length per 100 ANOVA

However, the measure of pause length per 100 words revealed that there are significant differences across the four band scores F(56,3)= 2.92, p< 0.04. Post hoc Tukey tests revealed that significant differences exist only between score bands 5 and 8 (p <0.03).

2.1.2.4 Complexity

Complexity was measured using two submeasures: the ratio of A units to AS units and the ratio of A units to total number of words. ANOVA showed that the differences among the four score bands for the ratio of A units to AS units were significant, F(56,3)= 3.95, p< 0.01.

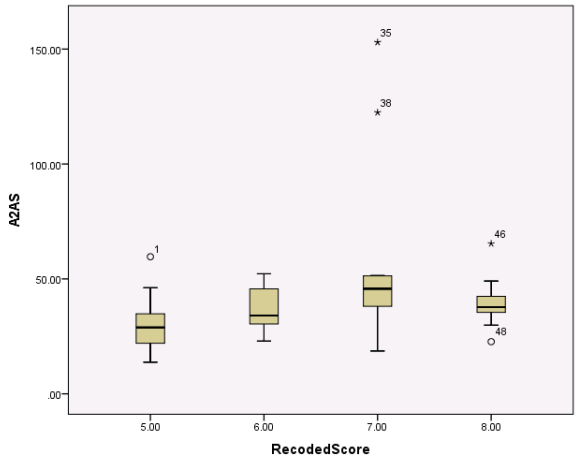

Figure 4: Complexity A to AS units ANOVA

ANOVA also showed that differences were significant for the ratio of A units to total number of words, F(56,3)= 3.58, p< 0.01

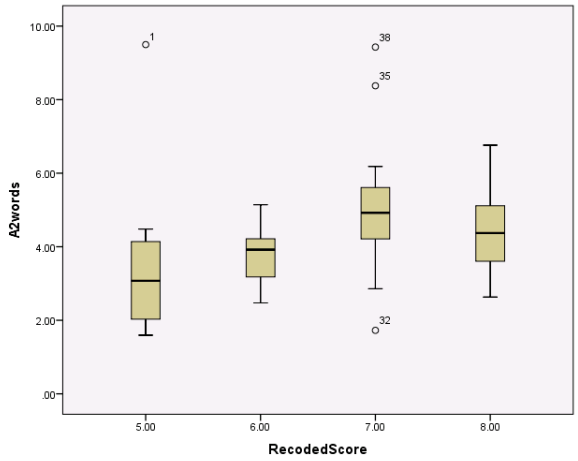

Figure 5: Complexity A to total number of words ANOVA

2.1.2.5 Grammatical range

ANOVA also revealed that the differences were significant among the four groups in the total number of verb forms used, F(56,3)= 8.06, p< 0.01

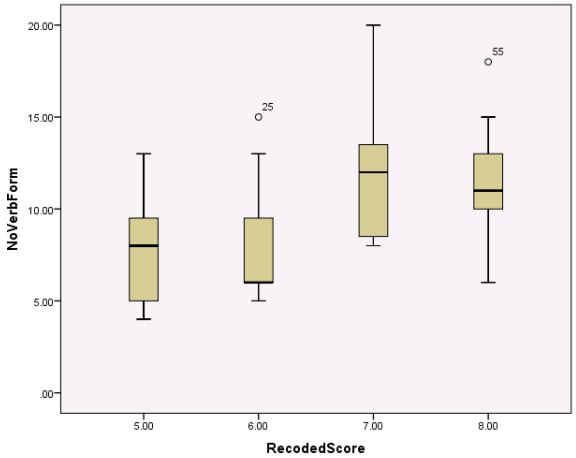

Figure 6: Grammatical range ANOVA

2.1.3 MANOVA

In order to avoid the possibility of overestimating the significance of the differences reported above by running more than one ANOVA test, a multivariate MANOVA test was used. The one-way multivariate analysis of variance (one-way MANOVA) is used to determine whether there are any differences between independent groups on more than one continuous dependent variable, which is the case in this study. In this regard, it differs from a one-way ANOVA, which only measures one dependent variable at a time. MANOVA showed that the differences in relation to the four measures/descriptors across the four IELTS band scores were statistically highly significant, F (3, 56) = 5.33, p < .001.

2.2 Qualitative analysis: Speaking features that have the potential to influence candidate scores

This section will outline the initial findings of the qualitative CA analysis. It will focus on describing and illustrating candidate speaking features that have the **potential** to impact upon candidates' scores, both 'positively' and 'negatively'. These candidate speaking features have been divided into two categories:
1) 'inter-turn' features (responses to examiner questions), and 2) 'intra-turn' features (the candidate's oral production within a turn) which have the potential to lead to an increase or decrease in a candidate's score.

For each of the speaking features identified, extracts will be presented to demonstrate how that particular feature occurs during ISTs. Where relevant, extracts from low and high score bands will be compared to illustrate the typical differences.

2.2.1 Answering the question: Inter-turn speaking features that can influence candidate scores

The interactional organisation of many institutional settings is dominated by question and answer sequences (Drew and Heritage, 1992, 39). The interactional organisation of parts 1 and 3 of ISTs is based on the topic-scripted QA adjacency pair (Seedhouse and Harris, 2010). To obtain a high score, candidates need to do the following: a) understand the question they have been asked; b) provide an answer to the question; c) identify the topic inherent in the question; and d) develop the topic inherent in the question. However, the band descriptors do not contain indicators relating specifically to a candidate's interactional ability to answer the questions that they have been asked. In spite of this, it seems likely that a candidate's interactional performance can contribute to the overall impression made on the examiner and, therefore, it has the potential to impact on their score. This section focuses on the variety of ways in which candidates can respond to a question and the potential impact of this on score.

2.2.1.1 Candidate requests repetition of the examiner's question

There are a number of ways in which candidates can mark trouble with an examiner's question. Extract 1 below illustrates a candidate (C) explicitly requesting repetition of the examiner's question (E).

Extract 1

```
46 E:    what traditions for na::ming babies are there (.) in you::r
47       culture
48       (1.4)
49 C:→   >er uh uh< hh (0.6) please HHH .HHH repeat the qu.hh.est.hh.ion
50       .hh
51       (.)
52 E:    <↑what traditions for naming babies (0.3) are there (.) in your
53       culture
54       (0.6)
55 C:    .hhh er:::m: hhh .hhh (0.7) ↑may↓be:: hh (0.5) ↓m:::. hh .hh
56       (1.3) .hhh ↑maybe some .hh (ing) or cul↓ture .h maybe (0.3)
57       er::m: (0.6) som::e (0.7) babies na::me (group) (.) represents
58       erm .hh happy meaning or ((inaudible)) .hh or .hh er (.) the
59       parents hope they have a .hh good future (0.2) (in ee:)
60       (1.3) in [their life]
017454T509 (5.0)
```

A lengthy pause follows the examiner's question, after which the candidate's turn opens with several markers of hesitation (line 49). After another pause, the candidate explicitly requests the repetition of the question. This request is formulated with embedded, soft laughter tokens, possibly marking an attempt to mitigate the potential threat to their assessment by this request. The examiner responds by repeating the question (though with slightly different intonation) and after a pause the candidate manages to provide an answer, although in quite a lengthy and hesitant manner. The kind of candidate trouble with answering questions described above does occur across all speaking bands, but in general there are more instances of trouble and repair in the lower bands than in the higher. In terms of the impact on candidate's scores, it is likely that these kinds of requests are assessed as problematic, with the potential to negatively impact on a candidate's score.

2.2.1.2 Candidate trouble with a question leads to a lack of an answer

Extract 2 below illustrates a different kind of trouble occurring with a candidate's attempts to answer a question, leading to the examiner abandoning the current question and moving onto the next.

Extract 2

```
77 E:    and how popular is <photography> (.) in your country? hh
78       (1.1)
79 C: →  m:↑::↓:(hh)m:: .hhh a person I $d(huh)on't kno::::w$=
80 E:    =how popular is (0.3) photography
81       (1.4)
82 C:    m::::. (0.4) I don't know:: (0.6) like a person?
83       (0.3)
84 E:    >it's okay< what kind of <photos d'you like looking at>
001638T521 (6.0)
```

The examiner's question (line 77) is followed by a lengthy pause, after which the candidate opens their turn with an extended floor holding token, followed by an in-breath. The candidate then utters a clarification request, the second part of which is marked by smile voice, an embedded laughter token, and sound stretching. The content of this clarification request focuses on a 'person they don't know', and it is unclear how this relates to the examiner's question. The examiner responds with a latched turn, repeating the first part of their initial question (line 80). A lengthy pause then follows. The candidate's response opens with a floor holding token ("`m:::::.`"), then after a pause, a reformulation of their clarification request. However, rather than reformulating the question again, the examiner's turn opens with a slowly delivered acknowledgment ("`>it's`

okay<"), followed by the next question in the sequence. The examiner's actions in the above extract demonstrate that they are not willing to continue engaging in attempts to elicit an answer from the candidate, but rather they move the candidate on to the next question. The examiner is following to the letter the guidance provided for examiners, namely to repeat the question once and then move on. This inability to answer the question is likely to be assessed as problematic by the examiner, and will have an impact on the candidate's score.

2.2.1.3 A candidate produces a problematic answer

Another area where candidates demonstrate trouble in answering examiner questions is illustrated below in extract 3. The candidate produces an answer to a question which is inappropriate in some way and the examiner moves onto the following question.

Extract 3

```
95 E:    .hh would you prefer to pick (0.2) to bu::y a picture postcard
96       or take a photo of a new place
97       (1.4)
98 C: →  ↓er:::. (0.8) YEs::
99       (0.9)
100 E:   now I'm going to give you a topic
101      (.)
102 C:   o[kay]
017454T509 (5.0)
```

The formulation of the examiner's question (in line 95) requires the candidate to choose between one of two possible answers (buy a postcard or take a picture). However, following a long pause, the candidate's response opens with a floor holding token ("↓er:::."), a further pause, and then "YEs::", delivered loudly with sound stretching. The candidate's inappropriate answer is followed by a long pause, after which the examiner initiates the transition to part 2 of the test.

Extract 4

```
146 E:   [d]'you have
147      many ↑neighbours
148      (1.3)
149 C:   ↑no:.=
150 E:   =thank you=
017454 509 (5.0)
```

In extract 4 above, the candidate (score 5.0) provides a direct answer to the question, but it is monosyllabic and does not develop the topic inherent in the question in any way.

Extract 5

```
144   E: .hhh thank you  (0.6) .hhh d'you have? ↑many ↑neighbours?
145      (0.5)
146   C: .hhh er I do (0.5) °(yeah ha)° [th]ere (has) abou:t er:: four=
147   E:                                [t-]
148   C: =to five of them? (0.2) but er ((name omitted))'s wife who is
149      of:: er our age yeah >they're they're< slightly elderly so::
150      (.)
151   E: thank you
022059 509 (8.0)
```

In extract 5 above, the candidate (score 8.0) is asked exactly the same question as in extract 4. Here, however, the candidate provides an additional item of information to develop the topic as well as answering the question. This difference in response to the topic element inherent in the question is likely to influence score.

Extract 6

```
7    E:  what subjects (0.2) are you studying
8        (0.5)
9    C:  er::m: (0.3) I study chemistry .hh eleven (0.3) erm:: (.) maths
10       eleven .hhh and english (.) an::d erm (0.4) ((inaudible)) study
11       .hh and drama
12       (0.6)
13   E:  ↑why did you decide to study these subjects
14       (1)
15   C:  uhm (.) because::e (.) I wi::ll go to- (0.2) I study .hhh er::
```

```
16      (.) canadian .hh er:: subjects I will go to:: .hh erm canada
17      (0.7)
017454 509 (5.0)
```

In extract 6 above, the candidate (score 5.0) struggles in lines 15 and 16 to provide a clear answer to the question of why they chose to study these subjects, or to develop the topic coherently.

Extract 7

```
238  E:  [so] >we have< a lot of international travel now: (0.5) <is
239      there any negative effects to that>
240      (0.5)
241  C:  m:: yes (0.9) er::m::: (0.7) m:: (0.3) sometimes  (0.2) er::
242      (1.1) it oo- (0.3) takes lon::g ti:me (0.7) er::: (0.4) and for
243      example if you travel to another country .hh er::: and it's a
244      takes  th::e (0.5) about er:: (0.5) (airbus) (0.3) to europe
245      (0.5) for about ten:: hours=
246  E:  =[uhu ]
247  C:   [time] .hhh and er the also (0.2)  the ticket is very
248      expensive (0.8) er::m: (0.7) °↓m:::.°
249      (.)
250  E:  so the negative what is the negative impact (0.5) <of
251      international travel>
252      (1.1)
253  C:  er:: (0.7) er::
254      (0.3)
255  E:  what's negative about it (0.3) besides (0.4) the ticket is
256      expensive (0.4) that's more like a difficulty
257      (0.7)
258  C:  °we::ll::°
259      (2.8)
260  E:  HH o[↑ka↓y::             ] thank you very much
054529 507 (5.0)
```

In extract 7 above, the candidate (score 5.0) provides an answer to the question in lines 241-245, but displays a limited, personal perspective on the issue. The examiner tries twice in lines 250 and 255 to get the candidate to engage with the broader, higher-level issues inherent in the topic, for example the negative impact on society or the environment. However, the candidate proves unable or unwilling to engage with these and the IST is terminated.

2.2.1.4 Features of answers by high-scoring candidates

What, by contrast, are the features of answers by high-scoring candidates? We look below at two candidates with a score of 8.0 in extracts 8 and 9 below.

Extract 8 (the question is: why would you recommend this area as a place to live?)

```
26   E:  why::.
27      (0.5)
28   C: well:: er:: first of all the erm (0.7) environment is: shall we
29      say:: (0.4) friendly? (0.5) and is it is pollution free .hh cos
30      there's not much transportation: er: around (.) so:: .hh the
31      pollution is at least at its lowest level
32      (1.9)
007673130 (8.0)
```

Extract 9

```
17   E: m↓hm (0.2) .hh so why did you decide to study dentistry
18      (0.4)
19   C: °oh° I was interested in life sciences right from the very
20      beginning [.hhh] an:d .hhh er: when I passed my (twelve)=
21   E:           [mhm ]
22   C: =standard? (.) we had to do an entrance examination? .hh and
23      the marks that I got (.) and the aptitude that I showed .hh
24      pointed more towards dentistry
25      (0.2)
26   E: m[↑hm]
015475 507 (8.0)
```

Both candidates answer the question directly and add at least one item of information to develop the topic inherent in the question. The answers are packaged in a linguistic format which displays some high-level features of fluency, accuracy, lexical choice and complexity. The two extracts are not presented as representative of band 8, but rather of successful answers which provide all of the required elements.

2.2.2 Speaking features that have the potential to influence candidate scores – 'intra-turn'

This section of the qualitative analysis focuses on individual speaking features that occur during the delivery of a candidate's turn. Although localised patterning was found in the distribution and frequency of these speaking features within particular candidates' ISTs, these localised patterns did not occur across score bands in a way that would allow us to claim that any particular speaking feature distinguishes between score bands. To claim that there is a simplistic relationship between any one of these speaking features and a candidate's overall score, from the perspective of the qualitative analysis in this study, would be going beyond the evidence provided by the data.

2.2.2.1 Functionless repetition

A speaking feature within turns which has the potential to negatively impact on their score in terms of fluency, is the presence of a high concentration of functionless repetitions. This feature is illustrated in extract 10 below.

Extract 10

```
201  E:    is it important (0.3) for teachers ts: (.) t'make students
202        (0.3) fee:l that they've done ↑well.
203        (1.7)
204  C:→   °°↓↓er:°° °it's° (1.5) I (0.4) you mean if I- (.) I was a teacher
205        .hh[h  I] (.) will no::- eh- (.) I will (.) I=
206  E:       [m↑hm]
207  C:→   =will- (0.2) will .h as:m- (.) will (.) I will (.) never .hh
208        (0.8) will:: $uhHH$ will (.) I will never let my students know
209        that HHH (0.5) who is the:: (0.2) goo:d (.) who is: ba::d
210        [.hh]
211  E:    [why] not
212        (0.3)
032622T521 (5.0)
```

The examiner's question is followed by a lengthy pause, after which the candidate's turn opens with a hesitation marker, then a 'false start' ("°it's°"). Both of these utterances are delivered quietly and indicate hesitation (also see extract 6) and they are then followed by another lengthy pause. The candidate restarts their turn on a different footing ("I"), continuing with a clarification request ("you mean if I- (.) I was a teacher"). The examiner provides a confirmation of this request (line 206) in overlap with the candidate's continuing turn. The candidate then produces multiple attempts to continue their turn (lines 205, 207-208), repeating various formulations of "I will". After five attempts, the candidate shifts to a second formulation "I will never", followed by several further repetitions of 'will', until they finally produce a (fairly) coherent unit of meaning at the end of line 208. Repetition is a multi-faceted interactional phenomenon, which has the potential to carry out a range of social actions. The kind of functionless repetitions illustrated above directly relate to the band descriptors. These state that repetitions, of a particular item (or items) within a turn, can be assessed as a mark of disfluency and therefore impact negatively upon the candidate's score.

2.2.2.2 Hesitation markers

Another candidate speaking feature that exhibits localised patterning within any given candidate's IST, though not in a generalisable way across score bands, is floor-holders, often described as hesitation markers in assessment literature (eg Ellis and Barkhuizen, 2005), such as "er::." and "erm::". Although a regular feature of ordinary conversation, high concentrations of hesitation markers disrupt the flow of a speaker's production and as such may be assessed as a mark of disfluency. Candidates who produce a high density of these interactional devices, as a regular and repeating feature of their turn design, could be assessed as disfluent within an IST; hesitation is included in the band descriptors within 'fluency and coherence'.

The candidate's answer to the examiner's question, in extract 11 below, is replete with examples of floor holders or hesitation markers. Although hesitation markers occur across the score bands, it is more common to find high concentrations of these features in some of the candidates in the lower score bands.

Extract 11

```
39    E:    ↑oh. (0.3) okay (0.6) let's move on to <talk about using
40          com↑pu↓ters> (0.5) <what do you generally use a computer for>
41          (0.9)
42    C:→   er::. we:ll: (0.3) er:m sometimes I:: (.) <chat with my> (0.2)
43          friends .hh er:: (.) er::m using ((inaudible)) computers .hh
44          such as (0.2) erm (pue pue) and em es en ((MSN)) .hh it's very
45          convenient .hh and to: my:: (.) er touch my friends (0.3) er::
46          (.) er ↑the:: (0.3) erm:: (0.3) er who ar::e (0.3) in a fu- who
47          ar:e $are from::$ (0.3) m:: (0.9) er far from::: (0.5) m:: $me
48          huh huh .HH$
49          (0.3)
054529T507 (5.0)
```

Extract 12 below (score 7.0) illustrates that although this candidate still utters floor holders or hesitation markers of the same types in the extract above, they are considerably less frequent. They are therefore less likely to interrupt the flow and disrupt the texture of the candidate's utterance.

Extract 12

```
37    E:    let's: (.) move on to talk about using <computers when do you
38          generally use a computer.>
39          (0.3)
40    C:→   erm (0.8) actually I generally use my computer (.) erm (0.2)
41          when I h:ave my leisure time (.) an and also:: (0.5) when
42          I want- want to watch (0.3) movies >free online< movies I-
43          (0.5) use it and (0.3) .hh also:: erm especially at night (0.6)
44          yeah
45          (0.3)
004017T507 (7.0)
```

2.2.2.3 Candidate's identity construction

As discussed in our previous report (Seedhouse and Harris, 2010), candidates display aspects of their identity within speaking tests which may impact upon their scores. Candidates at the higher scoring bands in this study almost exclusively present themselves as hard-working cultured intellectuals and (future) high achievers, with the exception of candidates still studying at high school.

Extract 13

```
42    E:    right (0.2) oka::y? (0.5) er what will be the subject or your
43          ma:jor study (.) for your erm: (0.4) future study.
44          (0.5)
45    C:→   erm right now i'm studying law?
46          (0.2)
47    E:    law (0.2) m hm (0.2) okay .hh (0.3) a::nd (0.2) what do you
48          like (0.2) about (0.2) studying (.) law (0.3) is there a
49          particular area that you:: (0.3) that appeals to you?
50          (0.6)
51    C:    i ↑think law: is very i:nteres↑ting:: erm particularly i
52      →   think I like the criminal? (0.3) [parts] in law
53    E:                                      [m hm ]
54          (0.2)
55    E:    okay (0.7) er:: and (0.8) what job would you like to do in the
56          future what area you'd like to specialize in criminal law?
57          (0.2)
58    C:→   er:::m well i'm thinking about (.) being a barrister?
59          (.)
60    E:    m hm
61          (.)
62    C:    but i haven't really decided if i wanna specialize in criminal
63          or simple.
64          (.)
005698T132 (8.0)
```

Extract 13 above illustrates a candidate (score 8.0), who constructs her identity as a (future) high achiever. In response to the examiner question about the candidate's studies, she describes herself as currently studying law (line 45), interested in criminal law (line 52) and that she is considering becoming a barrister in the future (line 58).

Extract 14

```
197   E:    y'know there are times in peoples lives (.) .hh often when they
198         want to be (.) the best number one at something .hh erm .hh
199         what are those (.) times (0.6) in people's lives

205   C:    .hhhh (.) I probably say education (0.5) when you're at that
206         point in high school and you're about to graduate you just- (.)
207         want to push yourself to get there to get to the best
208         university you wa::nt .hhh to:: (.) get into the field that
209         you've always wanted .hh and you j- (0.3) there's no boundary
210         as to how much you study there's no boundary as to how much
211         .hhh $coffee you're drinking just to stay up$ and get to w-
212         (0.5) get to su↑cceed y'know get to wherever you want .hhh
```

300643 521 (8.5)

In extract 14 above, the candidate (score 8.5) presents herself as a very hard-working high achiever.

Extract 15

```
118   E:    here's your to↑pi:c (0.4) I'd like you to des↑cribe .h
119         something you would like to succeed in doing

130   C:    .hhh (0.6) er::m (0.3) I've always wanted to create a vinci
131         (0.3) vinci was the city where da vinci was born .hh and
132         vinci turned out to be:: (0.2) like a cultural art hub (.)
```

030595 521 (8.0)

In extract 15 above, the candidate (score 8.0) portrays him/herself as a highly ambitious and cultured person who would like to recreate today a Renaissance-style centre of culture and art.

Extract 16

```
53          er::: (1.3) let's talk about what you do during your holidays
54          (0.3)
55    C:    £okay huh£ [duri]ng my holidays? well .hh erm:: during the=
56    E:               [yeah]
57    C:    =la:st erm two years I actually had no holiday[s .h]h I s::=
58    E:                                                  [m:: ]
59    C:    =I I stayed in surrey y'know .hh and erm I studied er the whole
60          time be[cause it]'s really hard to:: .hh erm .h study
61    E:           [m::     ]
62          (0.3)
63    C:    ((inaudible)) two two [diff]erent really different faculties=
64    E:                          [m::.]
65    C:    =[.hhh   ] erm so so I: I sp- (0.6) usually spend a: a: a:ll=
66    E:     [m::↑:↓:]
67    C:    =my free time .hh erm erm (0.3) er studying or:: (0.3) rather
68          .hh erm: going further into the subject [I like]
```

000053 132 (8.0)

In extract 16 above, the candidate (score 8.0) portrays herself as somebody who is so extremely hard working that she takes no holidays as she is studying in two different faculties at university – we learn earlier in the test she is doing degrees in both literature and law.

2.2.2.4 Candidate's lexical choice

Another speaking feature more commonly found in the talk of candidates at high score bands is the employment of less common lexical items, as anticipated in the descriptors.

Extract 17

```
412   E:    =what do you think about the future d'you think our lives will
413         be more stressful or less stressful=
414   C:    =£hm hm well .[hh]£ I I think er::m .hh er actually I had a=
415   E:                  [uh]
416   C:→   =course in methodolog[y e]r:m I was the teacher .[hh]h er:=
417   E:                         [m: ]                      [m:]
418   C:    =the mock teacher it was a mock class and er:m it was about
```

```
419         married people with [tex]t and the [les]son it it dealt with=
420   E:                        [m::]          [m::]
421   C:→   =it .hhh er::m in er::m (0.8) they predicated that er:m the
422         distinction between men and women would be completely (0.3)
423         erased that erm I dunno .hh identity crime can be used to
424         (enter a park) that er:m .hh r- r- really different things but
425         erm the- this whole machinery .hh er: is is indeed (.) erm a
426         very complex (.) er::m .hh issue the er: erm (0.3) scientific
427         advance is very closely connected to it and [if] you .h=
428   E:                                                    [m:]
429   C:    =do no:t .h er:m m= er:m (0.3) if if you don't .hh (0.2) give
430         (0.4) the right orienta:tion to it
431         (0.3)
432   E:    m[::]
433   C:     [th]en it turns (.) to the opposite
434         (0.3)
435   E:    m:[:]
000139T134 (8.5)
```

In extract 17 above, the candidate demonstrates a number of speaking features that may have a beneficial impact on their score. The candidate describes a previous experience during which they were the (mock) teacher of a methodology course (lines 416 and 418), and in doing so positions themselves as an intellectual high achiever (see section 2.1.3.1). The candidate then goes on to employ a number of less common vocabulary items in their extended turn. These include the use of 'distinction', 'scientific advance', 'machinery' (in an metaphorical, abstract sense), and 'orientation'. Although the employment of these lexical items is not always accurate, they nonetheless present the candidate as intellectually capable, and this may have a beneficial impact on their score.

By contrast, the following extract demonstrates a low scoring candidate's response to the same question.

Extract 18

```
184   E:    okay .hhh can you speculate on whether our lives will be more
185         (0.3) or less stressful in the future
186         (1.2)
187   C:→   i think it will be (0.5) more stressful (.) than now
188         (1.6)
189   E:    okay okay alright well we'll finish there thank you very much
000134T134 (5.0)
```

In extract 18 above, the candidate's turn does not orient to the examiner's request to "`speculate`". Instead, they produce a direct answer to the part of the question that asks "`whether our lives will be more (0.3) or less stressful in the future`", using similar lexical choice to the examiner. The candidate had the 'interactional space' to expand or extend their initial answer, but did not do so. In this case, the examiner orients to this lack of development and subsequently closes the test.

Extract 19

```
132   E:    m::. (0.5) m .hh and what kind of clothes do you like.
133         (0.3)
134   C:    .h (0.2) er::m (0.3) well I like erm (0.7) feminine clothes
135         er::m I: and I like clothing that er:m .hh underlines my
136         femininity (0.2) but that- does not exploit it in a:: (0.2)
137         dangerous w[ay: ]:
000053 132 (8.0)
```

In extract 19 above, the candidate (score 8.0) moves beyond description of clothes to relate clothing to more intellectual concepts such as exploitation of sexuality. The lexical choice includes less frequent items like 'exploit' and 'femininity'.

Extract 20

```
89    E:    .hh do birds have any special meanings in your culture
90          (1.6)
91    C:    ((tuts)) (1.7) yes:: (.) there are >certain birds that do have
92          special meaning for instance the> crow:: (0.5) is erm ((clears
93          throat)) (0.8) .hh the crow: is considered to be a bad omen (.)
94          .hh in most cases .hhh bu:t (1) sometimes it's also::  er:m:
95          (1.6) revered in the sense that er: (0.2) they believe that our
96          (0.7) there's some sort of ancestral connection with the bird
97          and the spirit and .hhh (0.3) yeah (0.9) so that's one of the
```

98 (0.6) examples
030595 521 (8.0)

In extract 20 above, the candidate (score 8.0) successfully develops the topic inherent in the question, managing the intellectual feat of conveying the dual significance of the crow in his/her own culture in a clearly structured fashion. The concepts are packaged in infrequent lexis (omen, revere, ancestral).

2.2.2.5 Candidate's 'colloquial delivery'

There are a number of speaking features that can give a given candidate's delivery the 'feel' of a 'colloquial' L1 user, and therefore have the potential to positively impact upon a candidate's score. These features are more commonly found in particular candidates at the higher score bands, but also occur in some lower band candidates.

Extract 21

```
23   E:   .hhh what will you do er::m (0.2) when you er::m (0.7) when you
24        complete your course (1) sorry what- (0.2) what will you do:
25        (.) now that you've completed your course.
26        (.)
27   C:   like hopefully I will start (.) like y'kno::::w working? (0.7)
28        I have a couple of businesses in my mind that I y'know I wanna
29        work (.) for: (0.7) an::d like you know (0.2) I can (0.3) I
30        think I can y'know achieve my goa:ls (0.5) °but° (0.9) °there°
31        (0.3)
```
300245T507 (8.0)

In extract 21 above, the candidate's answer employs various formulations of lexical items that give the sense of 'native-speaker' delivery, such as "like", "y'know" and "wanna". These aspects of 'colloquial delivery', when employed appropriately by a candidate, have an immediate impact upon the listener, presenting the speaker as a 'fluent user of English'. As such, they have the potential to positively impact upon the examiner's holistic impression of a candidate and achieve a 'halo effect'. These features are mentioned in the descriptors as 'spoken discourse markers', 'chunking' and 'elision'.

2.2.3 How clusters of speaking features distinguish tests rated at different levels from each other

In section 2.2 above, we introduced a range of speaking features which have the potential to influence candidate score. However, at this point we need to make an important caveat. Overall, we feel that attempting to focus on discrete individual features gives a misleading impression of the data. Rather, we feel that there are no individual speaking features that can be said to robustly distinguish between tests at the various bands. Rather, *clusters* of speaking features can be seen to distinguish candidates in various bands. We illustrate this point by examining the following extract.

Extract 22

```
52   E:   is unhappiness:: (.) always a bad thing?
53        (2.2)
54   C:   ↑not ↑necessarily (0.7) bu:t (.) you have to limit it (0.7)
55        like you can be: unhappy like on::e (0.8) a dear frie::nd or
56        someone that you know have passed away (.) you can you know (1)
57        have some grief (0.3) it's something you know healthy for you
58        to grieve (1.2) but like it's y'know it's just a process and
59        then you have to go y'know get back (.) to life (.) and you
60        know (0.2) start finding your happiness again
61        (1.3)
```
300245 507 (8.0)

The above extract 22 (score 8.0) demonstrates the dangers of trying to identify individual speaking features which can differentiate between scores. Hesitation and repetition are phenomena which are mentioned in the band descriptors as decreasing as scores increase. In the short extract above, we note six instances of hesitation of more than 0.5 seconds and repetitious use of "you know". Nonetheless, the question is answered, the topic is developed coherently and accurately with a range of structures and vocabulary. In the case of this specific topic and candidate, the use of pauses and "you know" may actually give a positive impression of authentic native-speaker-like philosophical musing with a friend about unhappiness and life, as opposed to non-native speaker lack of competence. The point to be made here is that it is not possible to isolate any single speaking feature which can unambiguously be related to a high or a low score. Furthermore, it is useful to employ a mixed methods approach to investigate this area as qualitative approaches can be employed to understand the significance of how features are employed in interaction by particular candidates in response to specific questions.

The qualitative analysis in this study is able to offer a way of partially understanding the complexities intrinsic in the relationship between speaking features and candidate scores. The following detailed analysis will illustrate how *clusters* of speaking features, rather that individual ones, can be seen to distinguish between candidates at the high and low ends of the range of score bands. The analysis will focus primarily on the speaking features that relate to the constructs of fluency,

grammatical range, complexity and accuracy analysed in the quantitative strand of this study. It will focus on the formulation and turn design of candidate answers to the first questions in the first part of the IST. Here we will focus on the work-related questions: "let's talk about what you do, do you work or are you a student?" and "do you enjoy the work?" There are three parts to this question on the examiners' script. Extract 23 will analyse an example of a candidate in score band 5. This will be followed by extract 24, which focuses on a candidate in score band 8.

Extract 23

```
4     E:    =d'you work or are you a student
5           (0.2)
6     C:    er: actually I'm both I'm er:: (0.3) I study and I: er: work
7           .hh=
8     E:    =.hh alright .h so what work do you do:
9           (.)
10    C:    .hh er I'm avi- an aviation engineer I just graduate
11          (0.3)
12    E:    .hh (0.3) hh .hh and d'you enjoy the work? hh
13          (0.2)
14    C:    yeah (1) I enjoy it er well- (.) .HHH
15          (0.3)
16    E:    why:.
17          (.)
18    C:    .hh because I:: er:: (0.3) studied f:- about er fixing
19          aeroplanes .hh and now I'm doing that
20          (0.2)
300169T507 (5.0)
```

In extract 23 above, the candidate's response to the first question (line 6) opens with a floor holder or hesitation marker ("er:"), which is followed by an answer to the question. The formulation of the examiner's question projects a response from the candidate of one of two options: work or study. However, the candidate's response does not meet this expectation: "actually I'm both", and as such can be described interactionally as a dispreferred response, in the sense that it does not fit with the normatively expected response: a second pair-part indicating either work or study. This may account for the candidate's employment of a hesitation marker in turn-initial position. The candidate continues by reformulating their answer, initially performing a self-initiated self-repair, "I'm er:: (0.3) I study", which indicates the candidate has identified trouble in their own utterance and carried out a 'grammatical' repair. The formulation of this repair includes another hesitation marker and a pause, common features in self-initiated self-repairs. If we relate the interactional features to the examiner's rubrics, we can see that this turn has the potential to lower the candidate's score. Although the speech rate of the turn is "not too slow" (FC), the candidate has uttered a number of hesitation markers and a self-initiated self-repair, which potentially represent problems with the candidate's "speech continuity" (FC).

In the next turn, the examiner asks the next question in the sequence. The candidate's in-breath, which follows the final utterance in their turn, is latched with the examiner's in-breath that opens this turn. In-breaths can be employed as an interactional device by which speakers indicate their intention to take the floor, and can perform the social action of taking the floor. And in this case, the examiner's in-breath carries out the social action of taking the floor. Here the length of the candidate's turn is interactionally restricted by the action of the examiner, who then goes on to ask the second question. The candidate's answer opens with an in-breath followed by a hesitation marker, and then another self-initiated self-repair, "I'm avi- an aviation engineer" (line 10). Here the candidate constructs the identity of a (future) high achiever, possibly impacting positively on their score. The candidate concludes their turn by further specifying that they have recently graduated, however, the grammatical formulation is problematic, as the verb is conjugated as "graduate". In this second answer, the candidate also employs a number of speaking features that could negatively impact on the score: a hesitation marker, a self-repair, and a grammatical error. Like the previous turn, the candidate does not elaborate their response.

The examiner then asks the third question and in the candidate's answer (line 14), there is a cluster of features that could negatively impact on their score. The candidate opens with a confirmation followed by a lengthy pause, potentially assessable as a marker of disfluency (FC). Their expansion of the answer is delivered with a hesitation marker and a 'non-standard' collocation, "I enjoy it er well-". Both of these features have the potential to lower the score, and furthermore, the candidate does not continue to elaborate. At which point, the examiner asks for a reason. The candidate's answer is again marked by features that could be detrimental to their score, there are two hesitation markers, the second of which occurs during a 'word search': "studied f:- about er fixing", which contains a "false start" (FC) and a 'non-standard' grammatical construction ("studied about fixing"). The candidate does, however, employ appropriate tense structures in the delivery of this turn.

The analysis of the above extract, from a candidate who scored 5.0, highlights a cluster of speaking features that are likely to have negatively impacted on their score. It demonstrates a number of features that could be assessed as problematic in terms of fluency, and grammatical range and accuracy. These included a high concentration of hesitation markers, self-initiated self-repairs, a false start, a word search, and a grammatical error. Furthermore, the candidate's turns are short and do not develop the topics inherent in the questions, even with the examiner's prompting.

The following extract, from a candidate who scored 8.5, illustrates how a radically different combination of speaking features can cluster to place the candidate in a high score band.

Extract 24

```
1     E:    so in this first part of the test I'd just like to ask you
2           some questions about yourself [.hhh] erm let's talk about=
3     C:                                  [okay]
4     E:    =what you do .hhh do you work or are you a student:?=
5     C:    =I'm a student in university? er::.=
6     E:    =and what subject are you studying.
7           (.)
8     C:    .hh I'm studying business human resources
9           (.)
10    E:    .H ↑ah. and why did you decide to study this subject.
11          (0.3)
12    C:    I've always loved business it's something I've always wanted to
13          do::. since I was a little gir::l I used to pretend like I was a
14          business woman [.hHHH] $A.HH.nd huh huh .HH sit around with=
15    E:                   [°mhm°]
16    C:    =a sui:::t n:: wear some glasses: n: pre[tend] like I'm doing=
17    E:                                            [°mhm°]
18    C:    =statistics: so yea:h$ it's something:: I've always wanted to
19          do >as my dream<
20          (0.5)
```

300643T521 (8.5)

During the delivery of the examiner's first question in extract 24 above, the candidate orients to the potential TRP left by the examiner's in-breath (line 2), by uttering an acknowledgment token ('okay') in overlap. This could potentially be assessed as an indicator of interactional fluency; the candidate has identified where the next potential change in speakers occurs, in the examiner's turn, and oriented to it with an appropriate utterance. This can be seen as back-channelling, as opposed to the minimal turns frequent in low-scoring turns (see extract 4). The examiner, however, continues holding the floor and asks the first question in this sequence. Again, the candidate's answer is skilfully coordinated, latching to the close of the examiner's question; the answer is well formed grammatically and delivered at a 'native-speaker like' rate. The first two of these speaking features have the potential to be assessed as positive markers of fluency (FC).

The candidate then utters a floor-holder or hesitation marker. As in extract 12, the examiner overlaps and takes the floor, asking the next question in the sequence (line 6). The candidate's answer is once again grammatically accurate, direct, appropriate, fluent, and constructs their identity as a (future) high achiever. This turn demonstrates speaking features which relate to the higher score bands of the IST, in the Fluency and Grammatical Accuracy categories. Although there has been little elaboration from the candidate in terms of grammatical range, the examiner has successfully taken the floor from the candidate twice, moving on to the next question without allowing the candidate interactional space to elaborate. The first few turns of this IST have illustrated a cluster of speaking features that place the candidate in the higher bands of the IST. However, the following candidate answer clearly demonstrates why they are in the highest band investigated in this study.

The candidate's answer in line 12 opens with two accurate formulations in present perfect tense with appropriate adjectives, demonstrating grammatical range and accuracy. They are delivered without pause or hesitation, at a 'native-like' rate of speech. The second of these TCUs is closed with a sound stretch and a slight drop in intonation, indicating the closing of a turn. However, the candidate does not leave a pause, which could allow the examiner to take the floor, but rather moves seamlessly into the next TCU. Again, this turn demonstrates a high level of skill and accuracy in the employment of grammatical forms, the use of two different constructions to build the time frame around her narrative ("since I was" and "I used to") (GR: "range of sentence structures, especially to move elements around for information focus"). She also employs "like" (line 13), which in terms of written discourse might be deemed grammatically inaccurate, but in this context it lends her delivery a colloquial tone. This further adds to the combination of speaking features that she has demonstrated, which relate to high scoring candidates.

At the end of the candidate's first TCU in line 14, her loud in-breath is overlapped by the examiner with a continuer ("[°mhm°]"), which is delivered quietly and does not lead to a change of speakership. The candidate then delivers a connective with an embedded in-breath followed by laughter tokens ("$A.HH.nd huh huh .HH"). The laughter is very 'natural', confident, and interactionally effective as a pre-cursor for the upcoming 'humorous' part of the narrative. This is formulated, in smile voice, as a list of things 'she used to wear' and 'used to pretend' to do. Each of the item in the lists is connected by the colloquial pronunciation of 'and', as a sound stretched "n::", which further strengthens the projection of a 'native-like speaker' in her delivery. The candidate's list is closed with "so yea:h", another 'native-speaker like' feature; before she continues to move towards closing her turn. She opens her closing, by repeating a formulation from the opening of the narrative ("it's something:: I've always wanted to do") then closes with ">as my dream<". This final TCU adds further evidence to the candidate's demonstration that they are

skilled at constructing an effective narrative. Not only do they link the closing of the narrative to its opening, but they also close with a phrase that emphasises the 'lifelong' importance of their interest in this area, and invokes the emotional aspect of childhood dreams.

The micro-analysis of these candidate turns has demonstrated how clusters of speaking features, some of which relate directly to the examiner's rubric, can distinguish this candidate, in score band 8, from those of the lower score bands. In terms of fluency, the candidate has produced a very low frequency of hesitation markers; there are no false starts, no functionless repetitions and no word searches, being features which have the potential to negatively impact on a candidate's score. Furthermore, the candidate does not generate *any* intra-turn pauses throughout these turns (though unsurprisingly, this does not continue throughout the test), a feature which could be assessed as demonstrating fluency. They also demonstrate other features relating to fluency that could improve their score, such as their interactional coordination with the examiner, including the timing of their talk (lines 1-6), their 'native-like' rate of speech, and the use of 'colloquial delivery'. In terms of grammatical range and accuracy, the candidate's turns analysed above also demonstrate a number of speaking features that relate to the high score bands. The candidate produces a number of 'difficult' grammatical structures within a single turn. As well as being grammatically well formed, these TCUs are also sequenced in a highly appropriate and sophisticated way for the construction of a narrative about childhood dreams.

The uncovering of particular *clusters* of candidate speaking features, described in the analyses above, demonstrates how CA can explicate the complexities and subtleties of candidates' interactional 'performance' during an IST. Furthermore, comparing candidates from the highest and lowest score bands investigated allows us to see how these combinations of speaking features can be seen to distinguish between candidates at those score bands. Although the analysis has not identified any **individual** speaking features that can be seen to distinguish between the score bands, it has demonstrated that even within a few turns of any given IST, a cluster of assessable speaking features may impact a candidate's score. The qualitative analysis of this dataset also suggests (though this cannot be analytically demonstrated with this methodology) that the examiners are highly attuned to identifying clusters of speaking features within candidate's turns-at-talk in a given IST, and drawing upon their 'noticings' of these clusters of features to effectively assess a candidate performance, and in doing so, place them into a particular score band.

3 ANSWERS TO RESEARCH QUESTIONS

3.1 Research question 1

The quantitative strand of this study focused on the first research question: *the grading criteria distinguish between levels 5, 6, 7 and 8 in the ways described in the speaking band descriptors – To what extent are these differences evident in tests at those levels?*

Table 2 showed the descriptive statistics for the four measures. Looking at the mean scores for each measure, it is evident that:

1. The total number of words per test increased in direct proportion to the scores, band by band, at a significant level. This confirms Brown's (2006a, 84) finding that the total amount of speech exhibited significant differences across levels.

2. The percentage of errors per 100 words decreased as the scores got higher, band by band, at a significant level. This suggests that accuracy increases in direct proportion to score, which confirms Brown's (2006a, 82) finding.

3. The measure of pause length relates to the construct of fluency and we adapted this to be pause length per 100 words in view of point 1 above. This showed that fluency increased significantly in direct proportion to the scores across the four bands, although post hoc Tukey tests revealed that the significant differences exist only between score bands 5 and 8. This confirms Brown's (2006a, 80) finding that the ratio of pause decreased as score increased, although differences were not significant across levels.

4. Both measures for grammatical complexity showed the same trend. While complexity is lowest for band 5, those at band 7 showed more complexity than those at band 8

5. The same trend was seen in the grammatical range measure. While band 5 shows the lowest number of verb forms, those who have scored 7 used a wider range of verb forms than those at band 8. So all of the measures for grammatical range and complexity confirm the same picture at a significant level. This does confirm Brown's (2006a, 82) finding that "Band 8 utterances were on average less complex than those of Band 7".

6. In the case of all four measures, we see variation in the expected directions when we compare bands 5 and 8, and the results therefore provide validity evidence. However, the measures did not all show linear progression throughout the four bands.

3.2 Research question 2

The qualitative strand of this study set out to answer the second research question: *which speaking features distinguish tests rated at levels 5, 6, 7 and 8 from each other?*

3.2.1 Speaking features which have the potential to impact upon candidate scores

The features which were identified in section 2.2 as having the potential to affect scores were: how well a candidate answers a question; hesitation markers; functionless repetition; identity construction; lexical choice; colloquial delivery; and incidence of trouble and repair. Seedhouse and Harris (2010) also reported: engaging with and developing a topic; constructing an argument; and turn length in part 2.

Although the analysis uncovered localised trends and patterns, particularly within individual candidates' ISTs, these localised patterns did not occur across score bands, and therefore could not be said to distinguish between score bands. None of the speaking features 'tracked' across the dataset, including those described in section 2.2 above, demonstrated simplistic, generalisable patterning *across* the score bands investigated, which would allow a robust analytic finding to be drawn from the data. For every localised pattern that was identified, a significant number of counter cases where found that contradicted this localised patterning, when viewed across the whole corpus.

There are therefore no individual speaking features that can be said to robustly distinguish between tests at the various bands. Rather, *clusters* of speaking features can be seen to distinguish candidates in various bands. An atomistic approach to relating individual speaking features directly to bands was not successful with this dataset. Rather, a qualitative approach which identifies how aggregates of speaking features cluster at different band levels has proved to be more successful, as illustrated in section 2.2.3. Band ratings relate more clearly to clusters of features than to individual features.

4 CONCLUSIONS

4.1 Combining the answers to the research questions: Findings

The features which were identified in section 2.2 as having the potential to affect scores were: how well a candidate answers a question; hesitation markers; functionless repetitions; identity construction; lexical choice; colloquial delivery; and incidence of trouble and repair. Seedhouse and Harris (2010) also reported: engaging with and developing a topic; constructing an argument; and turn length in part 2. The qualitative analysis also suggested that it is extremely difficult to establish any clear-cut, one-to-one correspondence between the features described in the band descriptors and the interaction produced in the IST.

The analysis has not identified any individual speaking feature that can be seen to distinguish between the score bands. Rather, it has proposed that in any given IST, a cluster of assessable speaking features can be seen to lead toward a given score. The overall picture is that the speaking features show variation in the anticipated directions from bands 5 to 8, and this provides validation evidence for the IST.

This overall picture is compatible with the quantitative evidence. In the case of all four measures, we see variation in the expected directions when we compare bands 5 and 8, and the results therefore provide validity evidence. However, the measures did not all show linear progression throughout the four bands. Accuracy and fluency do increase in direct proportion to score. However, both measures employed for grammatical range and complexity showed the same trend. While it is lowest for band 5, those at band 7 showed more range and complexity than those at band 8. So for 2 of the 4 measures used, there has not been clear linear progression throughout each of the 4 bands. Similar findings are reported by Brown (2006a, 83), namely "While all measures broadly exhibited changes in the expected direction across the levels, for some, the difference between two adjacent levels were not always as expected".

From both the qualitative and quantitative perspectives, a similar picture emerges of the relationship between speaking features and band descriptors. On the whole, the relationship varies in the expected way. However, when we examine individual performances, there may be considerable variation in relation to the four band descriptors. Furthermore, there may be anomalies in relation to adjacent bands; it has now been reported in two studies (Brown 2006a and this study) that band 8 utterances are less complex than those of band 7. So we cannot expect the speaking features of a band 8 performance to match exactly the features of band 8 in the band descriptors. The quantitative and qualitative studies have approached the data from different directions, but paint a similar picture, namely that band ratings relate more clearly to clusters of features than to individual features. The overall pattern of results is similar to that of Brown's (2006a) quantitative study. Brown (2006a, 71) concluded that "Overall, the findings indicate that while all the measures relating to one scale contribute in some way to the assessment on that scale, no one measure drives the rating; rather a range of performance features contribute to the overall impression of the candidate's proficiency". This study adopted a mixed methods approach and has reached a similar conclusion. An atomistic approach to identifying which discrete individual components of a candidate's performance determine their score is unlikely to be successful.

4.2 Discussion, implications and recommendations

Why might it be that there is no straightforward correspondence between individual band descriptors and individual features of candidate talk such that we can find all of the features of a band 6 descriptor in a 6-rated IST performance? To suggest a possible explanation, we make some observations on the discoursal organisation of the IST. When candidates speak in parts 1 and 3 of the IST, the institutional aim is for their talk to be evaluated as a speech sample by the examiner in relation to the band descriptors. However, as shown in the CA analyses of data above, candidates have to attend to demanding discoursal requirements on a moment-to-moment basis, as imposed by the task structure and specifically the topic-scripted QA adjacency pair. As Seedhouse and Harris (2010) put it, in order to obtain a high score, candidates need to provide an answer to the question and develop the topic inherent in the question. The 'discourse involvement hypothesis' which we propose states that the candidate may orient during parts 1 and 3 primarily to the discoursal demands of the topic-based QA adjacency pair rather than to ratings scales.

From the point of view of the institutional aim of matching talk to descriptors, the examiner's prompts might ideally be a neutral platform for the candidate to display a range of lexical and grammatical structures, of pronunciation and discoursal features, a launchpad for the generation of rateable features for easy matching to descriptors. In practice, however, the examiner's questions require the candidate to provide both a direct answer to the question and the not-too-short but not-too-long development of the topic inherent in the question. This means that in the transcripts we do not generally find candidates producing clearly delineated characteristics of talk which could clearly and neatly be matched to those characteristics of talk specified in the grading criteria. This is precisely because of the discourse involvement load created by the topic-scripted QA adjacency pair. This is not in any way a criticism of the structure: all varieties of talk inevitably involve participants in a specific discoursal structure. The topic-scripted QA adjacency pair, as previously noted, is extremely efficient in generating differential performance between candidates. Rather, this 'discourse involvement hypothesis' attempts to explain why it is not straightforward to match the band descriptors directly to features of talk in the IST. The level of discourse organisation intervenes and mediates the talk, thereby transforming that which is institutionally intended into what actually happens. In a similar way, Seedhouse (2004, 252) suggests that L2 classroom interaction has a level of discourse organisation which mediates between pedagogy and learning and transforms the task-as-workplan into the task-in-process, the intended talk into the actual talk.

We now consider how the discoursal requirements of the IST relate to the band descriptors. In parts 1 and 3 of the IST, the topic-scripted QA adjacency pair imposes discoursal requirements on the candidate, specifically to answer the examiner's question, but this is not specified as such in the grading criteria. Descriptors do mention fluency, coherence and 'discussion of topics', but the ability to answer questions is not specifically included. What is unclear is the extent to which the candidate's ability to meet the specific discoursal requirements of the IST is being oriented to by examiners as a contributory factor in the ratings process. This was not an issue mentioned specifically by raters in Brown's study of the IST ratings process, although they did refer to the additional criterion of "the ability to cope with different functional demands" (Brown, 2006b, 62). However, if it were the case that candidates are being partially assessed on their ability to participate in the unique speech exchange system of the IST, this would have implications for candidate preparation. Kasper (2013, 279) suggests that real-world pragmatic competence does not necessarily transfer directly to the demands of OPIs: "…real-world pragmatic competence gets in the way in the OPI at moments where the purpose of language assessment requires a different kind of pragmatic competence, that is, to understand and act upon the institutionally critical focus of the interviewers' task instructions". Discoursal participation in the IST may be a skill which needs to be learnt and it may be a skill which forms part of the ratings process in practice. This is an issue which may be worth investigating in future research, which should involve examiner perspectives. This study has suggested that band ratings relate more clearly to clusters of features than to individual features. However, no evidence has been provided that examiners orient to clusters in this way, since examiner perspectives have not formed part of the methodology.

We recommend that consideration be given to the explicit inclusion in the ratings process of the quality of candidate discoursal participation or interactional competence. In particular, we recommend including in the band descriptors 'the ability to answer questions'. This is because of their central importance to the discourse structure. Furthermore, examiner instructions for part 1 state: "The exact words in the frame should be used. If a candidate misunderstands the question, it can be repeated once but the examiner cannot reformulate the question in his or her own words. If misunderstanding persists, the examiner should move on to another question in the frame". (Instructions to IELTS Examiners, p 5). This implies that the issue of whether a candidate can understand and respond to a question is of importance. Brown (2006b, 49) reports examiners referring to candidates being on task or not "answering the question" in relation to the fluency and coherence scale, so our recommendation would mean making explicit what may be existing practice for examiners.

The Band Descriptors for Examiners state (Note i) "A candidate must fully fit the positive features of the descriptor at a particular level". We recommend that this instruction be reviewed. Both our quantitative and qualitative analyses suggest that examiners have not been following these instructions in their practice. Furthermore, the manner in which speaking features are distributed in clusters means that it would be very difficult to follow this instruction in practice.

REFERENCES

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.
- IETLS Handbook, 2005.

Brown, A, 2006a, 'Candidate discourse in the revised IELTS Speaking Test', *IELTS Research Reports Vol 6*, IELTS Australia and British Council, Canberra, pp 71-89

Brown, A, 2006b, 'An examination of the rating process in the revised IELTS Speaking Test', *IELTS Research Reports Vol 6*, IELTS Australia and British Council, Canberra, pp 41-70

Drew, P, and Heritage, J, 1992, 'Analyzing talk at work: an introduction' in *Talk at Work: Interaction in Institutional Setting,* eds P Drew and J Heritage, Cambridge University Press, Cambridge, pp 3-65

Douglas, D, 1994, 'Quantity and quality in speaking test performance', *Language Testing,* 11, pp 125-144

Ellis, R, and Barkhuizen, G, 2005, *Analysing Learner Language,* Oxford University Press, Oxford

Foster, P, Tonkyn, A, and Wigglesworth, G, 2000, 'Measuring spoken language: a unit for all reasons', *Applied Linguistics*, 21, 3, pp 354-375

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

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

Hopkins, D, and Cullen, P, 2007, *The IELTS Grammar Preparation Book*, Cambridge University Press, Cambridge

Johnson, R, Onwuegbuzie, A, and Turner Lisa, A, 2007, 'Towards a definition of mixed methods research', *Journal of Mixed Methods Research*, 1, pp 112-138

Kasper G, 2013, 'Managing task uptake in oral proficiency interviews' in *Assessing Second Language Pragmatics, eds S* Ross and G Kasper, Palgrave Macmillan, Basingstoke, pp 258-287

Lazaraton, A, 1998, *An analysis of differences in linguistic features of candidates at different levels of the IELTS Speaking Test*, Report prepared for the EFL Division, University of Cambridge Local Examinations Syndicate, Cambridge

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

Luoma, S, 2004, *Assessing Speaking*, Cambridge University Press, Cambridge

Mehnert, U, 1998, 'The effects of different lengths of time for planning on second language performance', *Studies in Second Language Acquisition*, 20, pp 52-83

Psathas, G, 1995, '*Conversation Analysis: The Study of Talk-in-Interaction*', Sage, London

Richards, K, Ross, S, and Seedhouse, P, eds, 2012, *Research Methods for Applied Language Studies*, Routledge, Oxon

Schegloff, EA, 1993, 'Reflections on quantification in the study of conversation', *Research on Language and Social Interaction*, 26, 1, pp 99-128

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

Seedhouse, P, and Harris, A, 2010, 'Topic Development in the IELTS Speaking Test', *IELTS Research Reports Vol 12*, IDP: IELTS Australia and British Council, Melbourne, pp 55-110

Skehan, P, and Foster, P, 1999, 'The influence of task structure and processing conditions on narrative retellings', *Language Learning*, 49, pp 93-120

Taylor, L, ed, 2011, *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Cambridge University Press, Cambridge

van Lier, L, 1989, 'Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversations', *TESOL Quarterly*, 23, pp 480-508

Weir, CJ, Vidakovic, and Galaczi, ED, 2013, *Measured Constructs: A History of Cambridge English Language Examinations 1913-2012*, Cambridge University Press, Cambridge

Young, R, 1995, 'Conversational styles in language proficiency interviews', *Language Learning*, 45, 1, pp 3-42

Young, RF, and He, A, eds, 1998, *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*, Benjamins, Amsterdam

Yuan, F, and Ellis, R, 2003, The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production, *Applied Linguistics*, 24, 1, pp 1-27

APPENDICES

Appendix 1: Operationalising the complexity measure

The following rules apply when coding the transcript:

AS units

"AS unit is identified as an independent clause or sub-clausal unit together with any subordinate clauses(s) associated with either" (Foster el at, 2000, p 365)

AS units are syntactic units so the boundaries of an AS unit are those of a full utterance including the main and subordinate clauses.

- An independent clause is a clause with at least a finite verb

Example 1
C: [My surname and middle name is quite unusual]
-- AS units (independent clauses)

- An independent sub-clausal unit is either one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation. This means that an independent clause might not have a verb but will allow for one by recovery of the ellipted elements.

Example: 2
E: so what's your job then?

C: a nurse -- AS unit (independent sub-clausal unit)

A units

A units are subordinate clauses which have at least a finite or non-finite verb element plus at least one other clause element such as subject, object, complement, or adverbial (Foster et al, 2000, p 366). Thus, contrary to AS units, A units must have a verb.

Example 3
1. C: /I'm basically/ /since I'm a freshman/ I do general nursing -- two A units although the first one doesn't have a complement but it has a subject and a finite verb

2. C: /when i use them for lon::g/: they sprain the eyes -- one AS unit with one A unit and an independent clause

Identifying boundaries of AS and A units

As a rule, the existence of falling or rising intonation followed by (0.5) pause identifies the start of a new AS unit. When there are cases of doubt, count the utterance as a separate A unit to show complexity.

Difficulty with identifying the boundary of AS units

Example:
C: [I've always loved business][it's something/ I've always wanted to do/ :: /since I was a little girl// I used to pretend like// I was business woman/]

This utterance could be analysed in two ways, either that it's something she wanted to do since she was a little girl so 'since I was a little girl' is a subordinate clause within the preceding AS unit, or 'since I was a little girl' is the start of a new AS unit. In such a case, we analyse it to the benefit of the student, i.e. as an A unit to demonstrate more complexity.

Difficulty with identifying the boundary of A unit

In the case of the non-finite progressive participle, there should be at least one other clause element to be considered an A unit.

Example:
C: /the things I've recently done//that have really put me// into thinking about a coffee shop -- 3 A units

into thinking is followed by about a coffee shop so it is considered a separate A unit

C: /the things I've recently done //that have put me into thinking -- 2 A units

Independent clauses

Whenever we have a lot of independent clauses, if there is rising or falling intonation and (0.5) pause, it breaks the AS unit and we start a new one.

Non-finite phrases

A noun phrase without a verb is considered a separate AS unit if it is separated from the following phrase by falling intonation and a pause of (0.5)

Example:
C: [and some children] ↓(0.5) [they are playing the ball]
-- two AS units

Inaudible

If the AS or A unit rely on the inaudible to be identified as an AS or A, we are not counting it, we are counting the AS and A units even when they have inaudible if the inaudible doesn't make a difference.

Example:
E: Can I see your ID please?

C: here's (inaudible) -- AS unit because we are still able to reconstruct the structure to a full clause regardless of the inaudible

Ellipted clauses

When counting ellipted clauses, we ask the question 'could we reconstruct a whole clause from the ellipted utterance?'

If yes and we can relate it to what comes before it, we count the utterance as an AS unit

Example:
C: (name omitted) I prefer my (inaudible)

E: (name omitted) is it?

C: yes (we count it as an AS unit)

If not, we don't count it.

Example:
C:[/it is a hospital/ /located at ::: (0.4) erm (Kesin city/)]
-- 1 AS unit, 1 A unit

E: I see

C: yeah (we don't count it as AS unit)

The rule above applies only if the ellipted forms are produced as full utterances like the example above. However, if the ellipted utterances are produced with other clauses, we use (0.5) pause to identify the boundary of an AS unit.

For example:
C: [no:. (0.4) as much as my mother wanted to]
--1 AS unit

C: [no:. (o.6) as much as my mother wanted to]
-- 2 AS units

Interruption

The examiner interrupts the utterance. If what he is saying is backchannelling and there is less than 0.5 space, we consider what is said after as part of the previous utterance.

Example:
C: [/I'm basically// since I'm a freshman/ and (inaudible)

E: yep

(0.4)

C: erm i: do general nursing/]
-- one As unit, 2 A units within one AS unit

If what the examiner is performing is a complete turn, the structure is broken and what is said after is a new clause.

Hesitation markers

If there are three hesitation markers or more, they also break the boundary of an AS unit and we start a new one after the hesitation markers.

Example:
C: It's something I'd love and .hhh so:: (.) it's following my dream. Two AS units

Repetition

- False starts, repetition and self corrections are not counted.
- Repetition is not counted if it is within the same turn.

Difficulty with implementing the repetition analysis is that repetition is rarely exact.

They are repeating a clause but it is not exactly the same and there is a distance, we do count it.

Example:
C: I've always loved business it's something I've always wanted to do since I was a little girl I used to pretend like I was a business woman sit around with a suit wear some glasses pretend like I'm doing statistics so yeah it's something I've always wanted to do as my dream

In this example, repetition is in fact a summary and it does show complexity so we do count it although it is in the same turn.

They are repeating the same clause but the two clauses have different functions, we count both.

Example:
C: I do love my name: : [e .hh] hh I love my name because

In this example, the first instance is an answer to the question while the second is the start of an explanation so the two clauses have different functions

- We do not count fillers or use of phatic communion such as 'you know, you see, well', as they don't show complexity and AS units are syntactic units.

Coordination

Coordinating clauses are counted as two AS units if there is a pause of (.5) and the first one is marked with rising or falling intonation.

Example:
C: you have/ to go upstairs/ and you have /to take the stairs -- two independent clauses, 1 AS unit, 2 A units

C: [Last year I just graduated from bachelor of science in nursing] (0.5) and [right now I was hired by the national transplant institute] -- two AS units

If the two coordinating elements are verb phrases and the first one has falling or rising intonation and is followed by a pause of 0.5 second or more, we consider them two AS units.

If we have a subordinate clause followed by 'and' and another clause which we could relate to the main clause, we are counting them as two A units (subordinate clauses).

Example:
C: there's also the variety of seafood because we do live in a gulf country and we basically live in the sea

Total number of clauses

When calculating the total number of clauses, the same rules above apply.

Do not count repetition (exception listed above).

Do not count phatic communion.

Do not count false starts.

Appendix 2: Verb forms for grammatical range

Present simple	He promotes
Present continuous	He is promoting
Past simple	He promoted
Past continuous	He was promoting
Used to (repeated action)	He used to promote
Would (repeated action)	He would promote
Present perfect simple	He has promoted
Present perfect continuous	He has been promoting
Past perfect simple	He had promoted
Past perfect continuous	He had been promoting
Will future	He will promote
Going to future	He is going to promote
Future continuous	He will be promoting
Future perfect simple	He will have promoted
Future perfect continuous	He will have been promoting

Modals

Can	He can promote
Could	He could promote
May	He may promote
Might	He might promote
Must	He must promote
Will (willingness and habits)	He will promote
Would (willingness, future in past)	He would promote
Shall	He shall promote
Should	He should promote
Ought to	He ought to promote
Need	He needs to promote

Passive

Present simple passive	He is promoted
Present continuous passive	He is being promoted
Past simple passive	He was promoted
Past continuous passive	He was being promoted
Present perfect passive	He has been promoted
Past perfect passive	He had been promoted
Going to passive	He is going to be promoted
Will passive	He will be promoted
Modal passive	He could be promoted

(also can, may, might, must, will, would, shall, should, ought to, need)

Verb plus verb patterns

Verb + to-infinitive
He decided to promote

Verb + -ing
He prefers promoting

Verb +preposition + -ing
He's thinking about promoting

Verb + object + infinitive without to
He made us promote

Conditional

Zero conditional
If you heat water to 100C, it boils

First conditional
If I invest my money, it will grow

Second conditional
If I invested my money, it would grow

Third conditional
If I had invested my money, it would have grown

Reported speech
(Many verbs can be used to report instead of 'said')

Past simple reported
She said he promoted

Past continuous reported
She said he was promoting

Past perfect reported
She said he had promoted

Past perfect continuous reported
She said he been promoting

Will reported
She said he would promote

Is going to reported
She said he was going to promote

Modal reported
She said he could promote

(also can, may, might, must, will, would, shall, should, ought to, need)

Appendix 3: Transcription conventions

A full discussion of Conversation Analysis (CA) transcription notation is available in Atkinson and Heritage (1984). Punctuation marks are used to capture characteristics of speech delivery, **not** to mark grammatical units.

[	indicates the point of overlap onset
]	indicates the point of overlap termination
=	a) turn continues below, at the next identical symbol b) if inserted at the end of one speaker's turn and at the beginning of the next speaker's adjacent turn, it indicates that there is no gap at all between the two turns
(3.2)	an interval between utterances (3 seconds and 2 tenths in this case)
(.)	a very short untimed pause
word	underlining indicates speaker emphasis
e:r the:::	indicates lengthening of the preceding sound
-	a single dash indicates an abrupt cut-off
?	rising intonation, not necessarily a question
!	an animated or emphatic tone
,	a comma indicates low-rising intonation, suggesting continuation
.	a full stop (period) indicates falling (final) intonation
CAPITALS	especially loud sounds relative to surrounding talk
° °	utterances between degree signs are noticeably quieter than surrounding talk
↑ ↓	indicate marked shifts into higher or lower pitch in the utterance following the arrow
> <	indicate that the talk they surround is produced more quickly than neighbouring talk
( )	a stretch of unclear or unintelligible speech.
((inaudible 3.2))	a timed stretch of unintelligible speech
(guess)	indicates transcriber doubt about a word
.hh	speaker in-breath
hh	speaker out-breath
hhHA HA heh heh	laughter transcribed as it sounds
→	arrows in the left margin pick out features of especial interest

Additional symbols

ja ((tr: yes))	non-English words are italicised, and are followed by an English translation in double brackets
[gibee]	in the case of inaccurate pronunciation of an English square brackets
[æ]	phonetic transcriptions of sounds are given in square brackets
< >	indicate that the talk they surround is produced slowly and deliberately (typical of teachers modelling forms)
C:	Candidate
E:	Examiner

Appendix 4: IELTS speaking band descriptors

IELTS Speaking band descriptors (public version)

Band	Fluency and Coherence	Lexical Resource	Lexical Resource	Pronunciation
9	▪ speaks fluently with only rare repetition or self correction; any hesitation is content-related rather than to find words or grammar ▪ speaks coherently with fully appropriate cohesive features ▪ develops topics fully and appropriately	▪ uses vocabulary with full flexibility and precision in all topics ▪ uses idiomatic language naturally and accurately	▪ uses a full range of structures naturally and appropriately ▪ produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	▪ uses a full range of pronunciation features with precision and subtlety ▪ sustains flexible use of features throughout ▪ is effortless to understand
8	▪ speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language ▪ develops topics coherently and appropriately	▪ uses a wide vocabulary resource readily and flexibly to convey precise meaning ▪ uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies ▪ uses paraphrase effectively as required	▪ uses a wide range of structures flexibly ▪ produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	▪ uses a wide range of pronunciation features ▪ sustains flexible use of features, with only occasional lapses ▪ is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	▪ speaks at length without noticeable effort or loss of coherence ▪ may demonstrate language-related hesitation at times, or some repetition and/or self-correction ▪ uses a range of connectives and discourse markers with some flexibility	▪ uses vocabulary resource flexibly to discuss a variety of topics ▪ uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices ▪ uses paraphrase effectively	▪ uses a range of complex structures with some flexibility ▪ frequently produces error-free sentences, though some grammatical mistakes persist	▪ shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	▪ is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation ▪ uses a range of connectives and discourse markers but not always appropriately	▪ has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies ▪ generally paraphrases successfully	▪ uses a mix of simple and complex structures, but with limited flexibility ▪ may make frequent mistakes with complex structures, though these rarely cause comprehension problems	▪ uses a range of pronunciation features with mixed control ▪ shows some effective use of features but this is not sustained ▪ can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	▪ usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going ▪ may over-use certain connectives and discourse markers ▪ produces simple speech fluently, but more complex communication causes fluency problems	▪ manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility ▪ attempts to use paraphrase but with mixed success	▪ produces basic sentence forms with reasonable accuracy ▪ uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	▪ shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	▪ cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction ▪ links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	▪ is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice ▪ rarely attempts paraphrase	▪ produces basic sentence forms and some correct simple sentences but subordinate structures are rare ▪ errors are frequent and may lead to misunderstanding	▪ uses a limited range of pronunciation features ▪ attempts to control features but lapses are frequent ▪ mispronunciations are frequent and cause some difficulty for the listener
3	▪ speaks with long pauses ▪ has limited ability to link simple sentences ▪ gives only simple responses and is frequently unable to convey basic message	▪ uses simple vocabulary to convey personal information ▪ has insufficient vocabulary for less familiar topics	▪ attempts basic sentence forms but with limited success, or relies on apparently memorised utterances ▪ makes numerous errors except in memorised expressions	▪ shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	▪ pauses lengthily before most words ▪ little communication possible	▪ only produces isolated words or memorised utterances	▪ cannot produce basic sentence forms	▪ speech is often unintelligible
1	▪ no communication possible ▪ no rateable language			
0	▪ does not attend			