1 Introduction

Paul Seedhouse
School of Education, Communication and Language Sciences, Newcastle University

Fumiyo Nakatsuhara
Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire

Introduction

This chapter starts by providing background information on the IELTS Speaking test (IST) for readers who are not familiar with it. We then relate the IST to the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) and position the study in relation to the aims of English Profile (EP), considering issues such as reference level descriptions, language competencies and learner profiles. The aims and structure of the volume are outlined and we then introduce the main issues and themes which emerge during the course of the volume. A review of relevant literature on the IST and on interview tests in general is provided. Finally, the study is positioned in terms of the research gap and relevance to the fields of applied linguistics, language testing, language acquisition and discourse studies.

1.1 Background information on the IELTS Speaking test

The IST is one of the four components of the International English Language Testing System (IELTS), the most widely used English proficiency test for overseas applicants to British and Australian universities. In 2017, over 7,000 certified examiners administered over 3 million ISTs annually at more than 1,100 centres, in over 140 countries around the world (www.ielts.org). For candidates, this can be a very high-stakes test in that it can determine their access to the degree programme of their choice. IELTS 'measures the language proficiency of people who want to study or work where English is used as a language of communication' (www.ielts.org). The IST aims to evaluate how well a language learner might function in a target context, often an academic one. There are two types of IELTS: IELTS Academic and IELTS General Training, but the Speaking and Listening components are common

1

The Discourse of the IELTS Speaking Test

to both types (see Chapter 2, this volume). There is a 9-band grading system from 1 (Non User) to 9 (Expert). IELTS includes tests of Listening, Reading, Writing and Speaking. The total test length is 2 hours 44 minutes, made up of Listening (30 minutes), Reading (1 hour), Writing (1 hour) and Speaking (11–14 minutes).

The IST is administered face-to-face with a trained examiner. Examiners award a band score for each of four criterion areas: *Fluency and Coherence*, 25%; *Lexical Resource*, 25%; *Grammatical Range and Accuracy*, 25%; and *Pronunciation*, 25%. ISTs are spoken encounters between one candidate and one examiner and are designed to take between 11 and 14 minutes. Before the test starts, some time is taken up with the administrative business of verifying the candidate's identity. There are three main parts. Each part fulfils a specific function in terms of interaction pattern, task input and candidate output, as follows.

- **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 himself/herself and confirms the candidate's identity. The examiner interviews the candidate using verbal questions selected from familiar topic frames. This part lasts between 4 and 5 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 1 minute to prepare before speaking at length, for between 1 and 2 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, for example 'can you compare ideas that architects and the general public have about buildings?'.

Examiners receive detailed directives in order to maximise test reliability and validity. The most relevant and important instructions to examiners are as follows: 'Standardisation plays a crucial role in the successful management of the IELTS Speaking test' (*Instructions to IELTS Examiners*, 2001:11)[1]. 'The IELTS Speaking test involves the use of an examiner frame which is **a script that must be followed** . . . Stick to the rubrics – do not deviate in any way . . . If asked to repeat rubrics, do not rephrase in any way . . . Do not make any unsolicited comments or offer comments on performance' (*IELTS Examiner Training Material*, 2001:5, emphases in original). The degree of

---

1 A number of cited publications do not appear in the references as they are confidential and not publicly available. These have been superseded in the current (2017) test version. However, the nature of the new guidelines cannot be specified.

control over the phrasing differs in the three parts of the test as follows: 'The wording of the frame is carefully controlled in Parts 1 and 2 of the Speaking test to ensure that all candidates receive similar input delivered in the same manner. In Part 3, the frame is less controlled so that the examiner's language can be accommodated to the level of the candidate being examined. In all parts of the test, examiners are asked to follow the frame in delivering the script . . . Examiners should refrain from making unscripted comments or asides' (*Instructions to IELTS Examiners*, 2001:5). Research has shown that the speech functions which occur regularly in a candidate's output during the Speaking test are: 'Providing personal information; expressing a preference; providing non-personal information; comparing; expressing opinions; summarising; explaining; conversation repair; suggesting; contrasting; justifying opinions; narrating and paraphrasing; speculating; analysing. Other speech functions may emerge during the test, but they are not forced by the test structure' (UCLES and British Council 2007:12).

Detailed rating scales or Band Descriptors (the public version is available in Appendix A) have been developed which describe spoken performance at the nine IELTS bands, based on the following criteria.

> 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. The key indicators of coherence are logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices (e.g. connectors, pronouns and conjunctions) within and between sentences.
>
> Lexical Resource 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.
>
> (*IELTS Handbook*, 2005:11)

The Discourse of the IELTS Speaking Test

Equal weighting is given to each of the criteria. This is an analytic or profile approach (Taylor and Galaczi 2011) in which several performance features are evaluated separately on their own subscale prior to combining subscores to produce an overall score. Test items are calibrated to a common scale using Rasch methodology and for that reason results are reported on a common 9-band scale to the nearest half-band (Lim, Geranpayeh, Khalifa and Buckendahl 2013:7). The design and historical development of the IST and its predecessors are described in detail in Chapter 2, this volume.

Many courses around the world are run to prepare students for the IELTS test and many course materials and preparatory tests are available. A great deal of further information on the IST is available on the IELTS website: www.ielts.org.

Figure 1.1 The face-to-face IELTS Speaking test

Source: www.youtube.com/watch?v=5T6zglM1Onc&list=PLSAx4faA_rjNEb16E7 FGDnQTSn9dAuDPv

1.2 The CEFR and English Profile

The Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) is now the most influential frame of reference which is widely used within Europe for describing language ability. It was designed to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabuses and curriculum guidelines, the design of teaching and learning materials, and the assessment of foreign language proficiency. As Salamoura and Saville (2011) point out, the CEFR is an intuitively helpful descriptive scheme for researchers, curriculum designers, teachers, teacher

4

trainers and language testers. However, in many cases the existing scales and related descriptors have not proved to be operationally adequate as they stand, lacking in specific detail in relation to specific languages. The CEFR therefore needs to be adapted or developed further for each specific context in which it is to be used.

Therefore, Cambridge Assessment English established English Profile (EP) in 2005 with the aim of bringing more specific content to the CEFR in relation to the English language (Saville and Milanovic 2014). EP is a collaborative programme of interdisciplinary research whose goal is to transpose the CEFR for the English language and for teaching and assessment purposes where English is the language being learned (Salamoura and Saville 2011). The intended output is a 'profile' of English language learners in terms of the six proficiency bands of the CEFR. Research to date has tended to focus on grammatical, functional and lexical features of learner English in relation to the six proficiency bands of the CEFR. EP has provided detailed descriptions of the criterial features which can distinguish between levels in the areas of syntax, morphology, lexis and language functions (Green 2012, Hawkins and Filipović 2012, Salamoura and Saville 2011).

1.3 IELTS, the CEFR and EP

As explained above, one aim of EP is to transpose the CEFR into English for the purposes of assessment. Given IELTS' status as a global test of English, it is important to understand their relationship at this point. A number of studies have discussed the relationship between the IELTS test and the CEFR (Davidson and Fulcher 2007, Lim et al 2013, Milanovic 2009, Taylor 2004a, 2004b, Weir 2005b). A rudimentary mapping of the IELTS scale onto the CEFR is available in Appendix B, but comparisons between the two are not straightforward for a number of reasons, as detailed in the publications cited previously. There are six CEFR levels and IELTS has a 9-point band rating system. The CEFR is intended to promote needs and proficiency profiling (North 2014:13), whereas IELTS provides summative scores for gatekeeping purposes. Chapter 2, this volume, explains the historical development of the ELTS/IELTS tests, but here we may note that these pre-exist the CEFR by at least 15 years and the two were developed for different contexts and purposes. From a historical perspective, then, it is easy to see why the 9-band IELTS descriptors and the 6-level CEFR descriptors would not exactly align.

Lim et al (2013) describe empirical work to relate the IELTS test to the CEFR, providing the most detailed discussion to date of the relationship. They reported on a standard-setting study and an external validation study for IELTS. Although the results of the two studies did not completely agree with each other, the study empirically suggested CEFR–IELTS cut scores from the two sources of evidence. For IST, IELTS 5.5 was suggested as the cut score

The Discourse of the IELTS Speaking Test

for B1/B2, IELTS 6.0–7.0 for B2/C1, and IELTS 7.5–8.0 for C1/C2. In addition, these cut scores were slightly different from those of other CEFR linking studies with international examinations including IELTS (Educational Testing Service 2010, Pearson Test of English 2009, Tannenbaum and Wylie 2008). Lim et al (2013) carefully considered the challenges that standard setting to the CEFR poses due to the purpose and nature of the CEFR (e.g. Weir 2005b), and discussed the validity of these inconsistent outcomes. They also argue that the purpose and nature of the CEFR pose challenges to the theory and practice of standard setting. Nonetheless, they conclude that the CEFR is still suited to identify some degree of synergy between the target levels of a test and the CEFR levels, and that standard setting to the CEFR is a meaningful activity.

It is important to note that the IELTS scores mentioned in this study are for the IST component only and not the overall band scores on the IELTS test as a whole. However, for the purposes of this volume, and purely as a point of reference, the examples of high-scoring ISTs employed in Chapter 6, this volume, are of IELTS 9.0 and 8.0, which very roughly correspond to CEFR Levels C2 to borderline C2/C1. The examples of low-scoring ISTs employed are of IELTS 5.0 and below, which very roughly corresponds to CEFR borderline B1/B2, moving downwards into B1.

1.4 This volume and EP

This volume aims to complement previous EP research by examining the micro-detail of the interaction in ISTs in order to understand how the organisation of the interaction generates differential candidate performance. The volume also examines the relationship between features of candidate discourse and the scores allocated to candidates. This is an area which undoubtedly merits more research and the study aims to contribute to the development of learner profiles of oral production at different proficiency levels, which is applicable to the CEFR as well as to IELTS.

This study argues that it is important to understand how the interactional organisation of an interview test relates to the type of talk generated by candidates. This relationship in turn influences how candidate talk might be related to descriptors of talk, such as those found in both the IST and the CEFR. The CEFR is intended as a framework of reference to promote profiling in different European languages (North 2014) rather than a prescriptive basis for test production. However, it is worth considering how the CEFR descriptors might be operationalised by test designers to develop interview tests. How might the discourse structure of interview tests be designed in relation to descriptors? For example, the CEFR descriptor for communicative fluency Level B2+ states that users 'relate own contributions skilfully to others' follow-up statements and inferences, and give feedback' (North 2014:29). If this descriptor were to be operationalised within an interview

test, one might assume that the interview test structure would allow extended two-way interaction in which an examiner responded to the candidate's turns and the candidate responded in turn to the examiner's response. If the interview test did not permit a candidate to follow up on an examiner's response, evidence of this particular candidate skill would not be available for assessment. The CEFR descriptors for communicative fluency Level B1+ state that users 'deal with difficult, less routine situations' (North 2014:30). It might prove difficult for an interview test to operationalise the concept of a 'less routine situation' within its discourse structure. The CEFR descriptors for strategic competence include one on 'asking for clarification' (North 2014:92), but in the IST examiners are trained to respond only to candidate requests for repetition, not clarification. The CEFR collated scale for interaction and production Level C1 includes the action 'get the floor, gain time and keep the floor while thinking' (North 2014:163), which would be applicable to interview tests in which multiple candidates compete for the floor. However, in interview tests in which the examiner asks questions and the candidate is automatically allocated a turn, it would be difficult to evaluate this skill. Another skill mentioned in the CEFR collated scales for interaction and production (Level B2) is 'ask follow-up questions to check understanding' (North 2014:163), whereas some interview tests do not currently include the possibility of candidates asking questions.

So the point is that there needs to be a match between the candidate talk that can be expected to be produced as part of the discourse structure of a specific interview test and the type of candidate talk and skill envisioned in CEFR descriptors which have been used to develop an interview test. It would be invalid to expect candidates to produce speech moves which are not actually available to them in the interactional organisation of a specific interview test. This volume attempts to contribute to understanding of this issue by making explicit the interactional organisation of a specific interview test and uncovering how this relates to the assessment of candidate proficiency. This may be particularly important in relation to CEFR because of the prevalence of 'Can Do' descriptors. The question then becomes whether the discourse structure of a particular interview test enables candidates to demonstrate what they can do in relation to Can Do descriptors and micro-descriptors.

1.5 The aims and structure of the volume: Interactional design and practice[2]

The volume aims to combine a detailed portrayal of the design of a face-to-face Speaking test with a detailed description of its actual implementation

---

2   The following section introduces a number of terms which are unpacked more fully in the following chapters.

The Discourse of the IELTS Speaking Test

in interactional terms. The studies presented in the volume therefore reveal the relationship between *interactional design* and *practice* in interview tests, as well as tackling the relationship between candidate talk and score. Using many empirical extracts of interaction from authentic ISTs, the volume illustrates how the interaction is organised in relation to the institutional aim of ensuring valid assessment. The overarching question investigated is how interactional design and practice are related in the IST. How was IST interaction designed and why? How is IST interaction actually organised in practice and how are the interaction and the scoring process related?

Chapter 2 covers the interactional design of the IST, explaining the principles and procedures underlying its development. To offer a better understanding of the development of the IST, the chapter first considers the Speaking component of the English Language Testing Service (ELTS) test, the predecessor of the IELTS test, which was launched in 1980. It then describes the original design of the IELTS Speaking test introduced in 1989 and how this was subsequently revised in 2001 in response to research outcomes from studies of IELTS and other Speaking tests. Such studies had applied discourse and conversation analysis techniques to actual performance samples in order to investigate the interactional features of spoken discourse, and their findings significantly informed the design and development of the revised IST in 2001. This historical review chapter shows how the ELTS and IELTS Speaking tests have dynamically evolved over years in response to continuing advances in the fields of language testing, applied linguistics, language pedagogy, and educational measurement.

Chapter 3 explains how to relate instances of talk to institutional design. The methodology employed is conversation analysis (CA). Studies of institutional interaction have focused on how the organisation of the interaction is related to the institutional aim and on the ways in which this organisation differs from the benchmark of 'ordinary' conversation[3]. The basic mechanisms, concepts and terms of CA are introduced in this chapter in relation to the baseline of ordinary conversation as a prelude to showing (in the following chapters) how these are transformed in relation to the institutional goal of the IST. Information is also provided on the corpora which underlie the study. Finally, we explain how complexity theory is employed in this study in a complementary way to CA to conceptualise the characteristics and function of the interactional system of the IST as a whole. The combination of methodologies is intended to allow a simultaneous focus on micro-detail in interaction as well as on the macro-level functioning of the IST interactional system in relation to the institutional goal.

Chapters 4 and 5 describe and analyse the IST in terms of *interactional practice*, based on empirical evidence. Chapter 4 describes the overall

---

3   The contested concept of 'ordinary' conversation will also be discussed in Chapter 3.

8

interactional organisation of the IST, illustrating the exposition with inter-actional extracts. The organisation of turn-taking, sequence and repair are tightly and rationally organised in relation to the institutional goal of ensur-ing valid assessment of English speaking proficiency. Examiner questions contain two components: a) an adjacency pair component, which requires the candidate to provide a response; and b) a topic component, which requires the candidate to develop a specific topic. This organisation may be termed a 'topic-scripted Q–A adjacency pair'. The overall organisation of the interac-tion is highly constrained in terms of who may say what and when, although there are some differences in the different parts of the test. There is considera-tion of how and why interactional trouble arises and how it is repaired. The organisation of repair has a number of distinctive characteristics in that it is conducted according to strict specified rules, in which the examiners have been briefed and trained.

Chapter 5 is devoted to *topic*. The organisation of topic must be under-stood as entwined with the organisation of turn-taking, sequence and repair and as directly related to the institutional goal. In this specific setting, topic has developed a 'dual personality' in service to the institutional goal. 'Topic-as-script' is the statement of topic in the form of a script which examiners read out to candidates; this is the same for all candidates or 'homogenised' input. By contrast, 'topic-as-action' refers to the very diverse or heteroge-neous ways in which candidates talk a topic into being, produce assessable output and therefore achieve diverse scores. This movement from 'topic' as a single homogeneous script to a heterogeneous series of responses by dif-ferent candidates (topic-as-action) enables differential ratings of their per-formances, and is the main focus of interest in this chapter. The examiner must match the (heterogeneous) features of individual candidate talk against the (homogenised) features of talk described in a specific Band Descriptor in order to allocate the appropriate grade to each candidate.

Chapter 6 tackles the relationship between interaction and score. The chapter presents the findings of a mixed methods study looking at how features of candidate discourse relate to scores allocated to candidates; the overall aim was to identify candidate speaking features that distinguish profi-ciency levels in the IST. The first research question noted that grading criteria distinguish between Bands 5, 6, 7 and 8 in the ways described in the Speaking Band descriptors, and asked to what extent these differences are evident in test performances at those levels. In order to answer this research question, quantitative measures for some of the constructs in the grading criteria were selected and applied to the spoken data (fluency, grammatical range, com-plexity and accuracy). The second question asked which speaking features distinguish test performances rated at Bands 5, 6, 7 and 8 from each other. This question was answered by working inductively from the spoken data, applying CA.

Chapter 7 pulls together the issues which have emerged from the previous chapters, draws implications and makes recommendations. There is discussion of the tension between the need to standardise interactional input to ensure scoring validity and the need to allow candidates freedom to express themselves in the interactional output. We consider the role of empirical research in test validation and also in uncovering challenges in the data and finding solutions for them; examples of these are provided. Issues are identified for further research and we consider the implications of the study for the CEFR and EP.

1.6 Issues and themes

The leitmotif of this study is the relationship between interactional design and practice; between the intentional design of IST talk and evidence of practice as revealed by test talk data. This is part of the test validation process: to what extent does the talk feature the anticipated characteristics? To what extent does the talk enable examiners to perform valid assessment by matching talk features to Band Descriptors? A pervasive theme is therefore the relationship between candidate talk and score. Chapter 2 reveals how this relationship was conceptualised and realised in the design of the IST. The empirical analyses in Chapters 4 and 5 contrast the features of candidate talk which receive high scores and low scores. Chapter 6 develops this theme in detail, using a mixed methods approach to consider which speaking features distinguish test performance at different levels. Throughout the volume we also try to reveal which characteristics of interactional design enable differential performance by candidates and differential grading by examiners. The key interactional mechanism in this regard is the topic-scripted Q–A adjacency pair. This is characteristic of the overall test design in relation to score: the examiner's input is scripted and homogenised in a format which enables the heterogeneous output by the candidate required for the assessment mechanism to function. So another persistent theme is the tension between the homogeneity required of the examiner input and the heterogeneity of candidate output required for the assessment mechanism to function as a system. The balance between homogeneity and heterogeneity is another recurring theme in our analysis; the IST creates a balance between homogeneity and heterogeneity in that the examiner must match the (heterogeneous) features of individual candidate talk against the (homogenised) features of talk described in a specific Band Descriptor in order to allocate the appropriate grade to each candidate.

Talk in the IST is studied as a complex adaptive system, which means that it is more than the sum of its parts; its behaviour emerges from the interaction of its components. It is presented as a system in which the mechanisms of spoken interaction have been adapted to the institutional goal of ensuring valid assessment. Therefore, we consider in detail the individual mechanisms