2. An examination of the rating process in the revised IELTS Speaking Test

Author: Annie Brown
Ministry of Higher Education and Scientific Research, United Arab Emirates

Grant awarded Round 9, 2003

This study examines the validity of the analytic rating scales used to assess performance in the IELTS Speaking Test, through an analysis of verbal reports produced by IELTS examiners when rating test performances and their responses to a subsequent questionnaire.

ABSTRACT

In 2001 the IELTS interview format and criteria were revised. A major change was the shift from a single global scale to a set of four analytic scales focusing on different aspects of oral proficiency. This study is concerned with the validity of the analytic rating scales. Through a combination of stimulated verbal report data and questionnaire data, this study seeks to analyse how IELTS examiners interpret the scales and how they apply them to samples of candidate performance.

This study addresses the following questions:

- How do examiners interpret the scales and what performance features are salient to their judgements?
- How easy is it for examiners to differentiate levels of performance in relation to each of the scales?
- What problems do examiners identify when attempting to make rating decisions?

Experienced IELTS examiners were asked to provide verbal reports after listening to, and rating a set of the interviews. Each examiner also completed a detailed questionnaire about their reactions to the approach to assessment. The data were transcribed, coded and analysed according to the research questions guiding the study.

Findings showed that, in contrast with their use of the earlier holistic scale (Brown, 2000), the examiners adhered closely to the descriptors when rating. In general, the examiners found the scales easy to interpret and apply. Problems that they identified related to overlap between the scales, a lack of clear distinction between levels, and the inference-based nature of some criteria. Examiners reported the most difficulty with the Fluency and Coherence scale, and there were concerns that the Pronunciation scale did not adequately differentiate levels of proficiency.

IELTS RESEARCH REPORTS, VOLUME 6, 2006
Published by: IELTS Australia and British Council

National Library of Australia, cataloguing-in-publication data, 2006 edition, IELTS Research Reports 2006 Volume 6
ISBN 0-9775875-0-9

CONTENTS

AUTHOR BIODATA:

ANNIE BROWN

Annie Brown is Head of Educational Assessment in the National Admissions and Placement Office (NAPO) of the Ministry of Higher Education and Scientific Research, United Arab Emirates. Previously, and while undertaking this study, she was Senior Research Fellow and Deputy Director of the Language Testing Research Centre at The University of Melbourne. There, she was involved in research and development for a wide range of language tests and assessment procedures, and in language program evaluation. Annie's research interests focus on the assessment of speaking and writing, and the use of Rasch analysis, discourse analysis and verbal protocol analysis. Her books include *Interviewer Variability in Oral Proficiency Interviews* (Peter Lang, 2005) and the *Language Testing Dictionary* (CUP, 1999, co-authored with colleagues at the Language Testing Research Centre). She was winner of the 2004 Jacqueline A Ross award for the best PhD in language testing, and winner of the 2003 ILTA (International Language Testing Association) award for the best article on language testing.

1 RATIONALE FOR THE STUDY

The IELTS Speaking Test was re-designed in 2001 with a change in format and assessment procedure. These changes responded to two major concerns: firstly, that a lack of consistency in interviewer behaviour in the earlier unscripted interview could influence candidate performance and hence, ratings outcomes (Taylor, 2000); and secondly, that there was a degree of inconsistency in interpreting and applying the holistic band scales which were being used to judge performance on the interview (Taylor and Jones, 2001).

A number of studies of interview discourse informed the decision to move to a more structured format. These included Lazaraton (1996a, 1996b) and Brown and Hill (1998) which found that despite training, examiners had their own unique styles, and they differed in the degree of support they provided to candidates. Brown and Hill's study, which focused specifically on behaviour in the IELTS interview, indicated that these differences in interviewing technique had the potential to impact on ratings achieved by candidates (see also Brown, 2003, 2004). The revised IELTS interview was designed with a more tightly scripted format (using interlocutor "frames") to ensure that there would be less individual difference among examiners in terms of interviewing technique. A study by Brown (2004) conducted one year into the operational use of the revised interview found that generally this was the case.

In terms of *rating* consistency, a study of examiner behaviour on the original IELTS interview (Brown, 2000) revealed that while examiners demonstrated a general overall orientation to features within the band descriptors, they appeared to interpret the criteria differently and included personal criteria not specified in the band scales (in particular *interactional* aspects of performance, and *fluency*). In addition, it appeared that different criteria were more or less salient to different raters. Together these led to ratings variability. Taylor and Jones (2001) reported that "it was felt that a clearer specification of performance features at different proficiency levels might enhance standardisation of assessment" (2001: 9).

In the revised interview, the holistic scale was replaced with four analytic scales. This study seeks to validate the new scales through an examination of the examiners' cognitive processes when applying the scales to samples of test performance, and a questionnaire which probes the rating process further.

2 RATING BEHAVIOUR IN ORAL INTERVIEWS

There has been growing interest over the last decade in examining the cognitive process employed by examiners of second language production through the analysis of verbal reports produced during, or immediately after, performing the rating activity. Most studies have been concerned with the assessment of writing (Cumming, 1990; Vaughan, 1991; Weigle, 1994; Delaruelle, 1997, Lumley, 2000). But more recently, the question of how examiners interpret and apply scales in assessments of speaking has been addressed (Meiron, 1998; Brown, 2000; Brown, Iwashita and McNamara, 2005). These studies have investigated questions such as: how examiners assign a rating to a performance; what aspects of the performance they privilege; whether experienced or novice examiners rate differently; the status of self-generated criteria; and how examiners deal with problematic performances.

In her examination of the functioning of the now retired, IELTS holistic scale, Brown (2000) found that the holistic scale was problematic for a number of reasons. Different criteria appeared to be more or less salient at different levels; for example comprehensibility and production received greater attention at the lower levels and were typically commented on only where there was a problem. Brown found that different examiners attended to different aspects of performance, privileging certain features over others in their assessments. Also, some examiners were found to be more performance-

oriented, focusing narrowly on the quality of performance in relation to the criteria, while others were reported to be more inference-oriented, drawing conclusions about candidates' ability to cope in other contexts. The most recently trained examiner focused more exclusively on features referred to in the scales and made fewer inferences about candidates.

In the present study, of course, the question of weighting should not arise, although examiners may have views on the relative importance of the criteria. A survey of examiner reactions to the previous IELTS interview and holistic rating procedure (Merrylees and McDowell, 1999) found that most Australian examiners would prefer a profile scale. Another question then, given the greater detail in the revised, analytic scales, is whether examiners find them easier to apply than the previous one, or whether the additional detail and difficulty distinguishing the scales makes the assessment task more problematic.

Another question of concern when validating proficiency scales is the ease with which examiners are able to distinguish levels. While Merrylees and McDowell (1999) found that around half the examiners felt the earlier holistic scale used in the IELTS interview was able to distinguish clearly between proficiency levels, Taylor and Jones reported concern as to "how well the existing holistic IELTS rating scale and its descriptors were able to articulate key features of performance at different levels or bands" (2001: 9). Again, given the greater detail and narrower focus of the four analytic scales compared with the single holistic one, the question arises of whether this allows examiners to better distinguish levels. A focus in the present study, therefore, is the degree of comfort that examiners report when using the analytic scales to distinguish candidates at different levels of proficiency.

When assessing performance in oral interviews, in addition to a range of linguistic and production related features, examiners have also been found to attend to less narrowly linguistic aspects of the interaction. For example, in a study of Cambridge Assessment of Spoken English (CASE) examiners' perceptions, Pollitt and Murray (1996) found that in making judgements of candidates' proficiency, examiners took into account perceived maturity and willingness or reluctance to converse. In a later study of examiners' orientations when assessing performances on SPEAK (Meiron, 1998), *despite* it being a non-interactive test, Meiron found that examiners focused on performance features such as creativity and humour, which she described as reflecting a perspective on the candidate as an interactional partner.

Brown's analysis of the IELTS oral interview (2000) also found that examiners focused on a range of performance features, both specified and self-generated, and these included interactional skills, in addition to the more explicitly defined structural, functional and topical skills. Examiners noted candidates' use of interactional moves such as challenging the interviewer, deflecting questions and using asides, and their use of communication strategies such as the ability to self-correct, ask for clarification or use circumlocution. They also assessed candidates' ability to "manage a conversation" and expand on topics. Given the use in the revised IELTS interview of a scripted interview and a set of four linguistically focused analytic scales, rather than the more loosely worded and communicatively-oriented holistic one in the earlier format, the question arises of the extent to which examiners still attend to, and assess communicative or interactional skills, or any other features not included in the scales.

Another factor which has been found to impact on ratings in oral interviews is interviewer behaviour. Brown (2000, 2003, 2004) found that in the earlier unscripted quasi-conversational interviews, examiners took notice of the interviewer and even reported compensating when awarding ratings for what they felt was inappropriate interviewer behaviour or poor technique. This finding supported those of Morton, Wigglesworth and Williams (1997) and McNamara and Lumley (1997), whose analyses of score data combined with examiners' evaluations of interviewer competence also found that examiners compensated in their ratings for less-than-competent interviewers. Pollitt and Murray (1993) found

that examiners made reference to the degree of encouragement interviewers gave candidates. While it is perhaps to be expected that interviewer behaviour might be salient to examiners in interviews which allow interviewers a degree of latitude, the fact that the raters in Morton et al's study, which used a scripted interview (the access: oral interview), took the interviewer into account in their ratings, raises the question of whether this might also be the case in the current IELTS interview, which is also scripted, in those instances where interviews are rated from tape.

3 RESEARCH QUESTIONS

On the basis of previous research, and in the interests of seeking validity evidence for the current oral assessment process, this study focuses on the interpretability and ease of application of the revised, analytic scale, addressing the following sets of questions:

1. What performance features do examiners explicitly identify as evidence of proficiency in relation to each of the four scales? To what extent do these features reflect the "criteria key indicators" described in the training materials? Do examiners attend to all the features and indicators? Do they attend to features which are not included in the scales? How easy do they find it to apply the scales to samples of candidate performance? How easy do they find it to distinguish between the four scales?
2. What is the nature of oral proficiency at different levels of proficiency in relation to the four assessment categories? How easy is it for examiners to distinguish between adjacent levels of proficiency on each of the four scales? Do they believe certain criteria are more or less important at different levels? What problems do they identify in deciding on ratings for the samples used in the study?
3. Do examiners find it easy to follow the assessment method stipulated in the training materials? What problems do they identify?

4 METHODOLOGY

4.1 Data

The research questions were addressed through the analysis of two complementary sets of data:

- verbal reports produced by IELTS examiners as they rated taped interview performances
- the same examiners' responses to a questionnaire which they completed after they had provided the verbal reports.

The verbal reports were collected using the stimulated recall methodology (Gass and Mackey, 2000). In this approach, the reports are produced retrospectively, immediately after the activity, rather than concurrently, as the online nature of speaking assessment makes this more appropriate. The questionnaire was designed to supplement the verbal report data and to follow up any rating issues relating to the research questions which were not likely to be addressed systematically in the verbal reports. Questions focused on the examiners' interpretations of, application of, and reactions to, the scales. Most questions required descriptive (short answer) responses. The questionnaire is included as Appendix 1.

Twelve IELTS interviews were selected for use in the study: three at each of Bands 5 to 8. (Taped interviews at Band 4 level and below were too difficult to follow due to intelligibility and hence, interviews from Band 5 and above only were used.) The interviews were drawn from an existing data-set of taped operational IELTS interviews used in two earlier analyses: one of interviewer behaviour (Brown, 2003) and one of candidate performance (Brown, 2004). Most of the interviews were

conducted in Australia, New Zealand, Indonesia and Thailand in 2001-2, although the original set was supplemented with additional tapes provided by Cambridge ESOL (test centres unknown). Selection for the present study was based on ratings awarded in Brown's 2004 study, averaged across three examiners and the four criteria, and rounded to the nearest whole band.

Of the 12 interviews selected, seven involved male candidates and five female. The candidates were from the following countries: Bangladesh, Belgium, China (3), Germany, India, Indonesia (2), Israel, Korea and Vietnam. Table 1 shows candidate information and ratings.

Interview	Sex	Country	Averaged ratings
1	M	Belgium	8
2	F	Bangladesh	8
3	M	Germany	8
4	M	India	7
5	F	Israel	7
6	M	Indonesia	7
7	M	Vietnam	6
8	M	China	6
9	F	China	6
10	M	China	5
11	F	Indonesia	5
12	F	Korea	5

Table 1: Interview data

Six expert examiners (as identified by the local IELTS administrator) participated in the study. Expertise was defined in terms of having worked with the revised Speaking Test since its inception, and having demonstrated a high level of accuracy in rating.

Each examiner provided verbal reports for five interviews, see Table 2. (Note: Examiner 4 only provided four reports.) Prior to data collection they were given training and practice in the verbal report methodology.

The verbal reports took the following form. First, the examiners listened to the taped interview and referred to the scales in order to make an assessment. When the interview had finished, they stopped the tape and wrote down the score they had awarded for each of the criteria. They then started recording their explanation of why they had awarded these scores. Next they re-played the interview from the beginning, stopping the tape whenever they could comment on some aspect of the candidate's performance. Each examiner completed a practice verbal report before commencing the main study. After finishing the verbal reports, all of the examiners completed the questionnaire.

Interview	Examiner 1	Examiner 2	Examiner 3	Examiner 4	Examiner 5	Examiner 6
1	X		X			
2	X	X			X	
3			X	X		X
4	X				X	
5		X		X		
6			X			X
7	X		X		X	
8		X			X	
9				X		X
10	X				X	
11		X		X		X
12		X	X			X

Table 2: Distribution of interviews

4.2 Score data

There were a total of 29 assessments for the 12 candidates. The mean score and standard deviation across all of the ratings for each of the four scales is shown in Table 3. The mean score was highest on *Pronunciation*, followed by *Fluency and coherence*, *Lexical resource* and finally *Grammatical range and accuracy*. The standard deviation was smaller on *Pronunciation* than on the other three scales, which reflects the narrower range of band levels used by the examiners; there were only three ratings lower than Band 6.

Scale	Mean	Standard deviation
Fluency and coherence	6.28	1.53
Lexical resource	6.14	1.60
Grammatical range and accuracy	5.97	1.52
Pronunciation	6.45	1.30

Table 3: Mean ratings

4.3 Coding

After transcription, the verbal report data were broken up into units, a unit being a turn – a stretch of talk bounded by replays of the interview. Each transcript consisted of several units, the first being the summary of ratings, and the remainder being the talk produced during the stimulated recall. At times, examiners produced an additional turn at the end, where they added information not already covered, or reiterated important points.

Before the data was analysed, the scales and the training materials were reviewed, specifically the key indicators and the commentaries on the student samples included in the examiner training package

(UCLES, 2001). A comprehensive description of the aspects of performance that each scale and level addressed was built up from these materials.

Next, the verbal report data were coded in relation to the criteria. Two coders, the researcher and a research assistant undertook the coding with a proportion of the data being double coded to ensure inter-coder reliability (over 90% agreement on all scales). This coding was undertaken in two stages. First, each unit was coded according to which of the four scales the comment addressed: *Fluency and coherence*, *Lexical resource*, *Grammatical range and accuracy*, and *Pronunciation*. Where more than one was addressed the unit was double-coded. Additional categories were created, namely *Score*, where the examiner simply referred to the rating but did not otherwise elaborate on the performance; *Other*, where the examiner referred to criteria or performance features not included in the scales or other training materials; *Aside*, where the examiner made a relevant comment but one which did not directly address the criteria; and *Uncoded*, where the examiner made a comment which was totally irrelevant to the study or was inaudible. Anomalies were addressed through discussion by the two coders.

Once the data had been sorted according to these categories, a second level of coding was carried out for each of the four main assessment categories. Draft sub-coding categories were developed for each scale, based on the analysis of the scale descriptors and examiner training materials. These categories were then applied and refined through a trial and error process, and with frequent discussion of problem cases. Once coded, the data were then sorted in various ways and reviewed in order to answer the research questions guiding the study.

Of the comments that were coded as *Fluency and coherence*, *Lexical resource*, *Grammatical range and accuracy*, and *Pronunciation* (a total of 837), 28% were coded as *Fluency and coherence*, 26% as *Lexical resource*, 29% as *Grammatical range and accuracy* and 17% as *pronunciation*. Examiner 1 produced 18% of the comments; Examiner 2, 17%; Examiner 3, 10%; Examiner 4, 14%; and Examiners 5 and 6, 20% each.

The questionnaire data were also transcribed and analysed in relation to the research questions guiding the study. Where appropriate, the reporting of results refers to both sets of data.

5 RESULTS

5.1 Examiners' interpretation of the scales and levels within the scales

In this section the analysis of the verbal report data and relevant questionnaire data is drawn upon to illustrate, for each scale, the examiners' interpretations of the criteria and the levels within them. Subsequent sections will focus on the question of the discreteness of the scales and the remaining interview questions.

5.1.1 Fluency and coherence

5.1.1a Understanding the fluency and coherence scale

The *Fluency and coherence* scale appeared to be the most complex in that the scales, and examiners' comments, covered a larger number of relatively discrete aspects of performance than the other scales – hesitation, topic development, length of turn, and use of discourse markers.

The examiners referred often to the amount of hesitation, repetition and restarts, and (occasionally) the use of fillers. They noted uneven fluency, typically excusing early disfluency as "nerves". They also frequently attempted to infer the cause of hesitation, at times attributing it to linguistic limitations – a search for words or the right grammar – and at other times to non-linguistic causes – to candidates thinking about the content of their response, to their personality (shyness), to their cultural background, or to a lack of interest in the topic (having "nothing to say"). Often examiners were unsure whether language or content was the cause of disfluency but, because it was relevant to the ratings decision

(Extract 1), they struggled to decide. In fact, this struggle appeared to be a major problem as it was commented on several times, both in the verbal reports and in the responses to the questionnaire.

Extract 1

> *And again with the fluency he's ready, he's willing, there's still some hesitation. And it's a bit like 'guess what I'm thinking'. It annoys me between 7 and 8 here, where it says – I think I alluded to it before – is it content related or is it grammar and vocab or whatever? It says here in 7, 'some hesitation accessing appropriate language'. And I don't know whether it's content or language for this bloke. So you know I went down because I think sometimes it is language, but I really don't know. So I find it difficult to make that call and that's why I gave it a 7 because I called it that way rather than content related, so being true to the descriptor.*

In addition to the amount or frequency of hesitation and possible causes, examiners frequently also considered the impact of too much hesitancy on their understanding of the candidate's talk. Similarly, they noted the frequency of self-correction, repetition and restarts, and its impact on clarity. Examiners distinguished repair of the content of speech ("clarifying the situation", "withdrawing her generalisation"), which they saw as native-like, even evidence of sophistication, from repair of grammatical or lexical errors. Moreover, this latter type of repair was at times interpreted as evidence of limitations in language but at other times was viewed positively as a communication strategy or as evidence of self-monitoring or linguistic awareness.

Like repair, repetition could also be interpreted in different ways. Typically it was viewed as unhelpful (for example, one examiner described the candidate's repetition of the interviewer's question as "tedious") or as reducing the clarity of the candidate's speech, or as indicative of limitations in vocabulary, but occasionally it was evaluated positively, as a stylistic feature (Extract 2).

Extract 2

> *So here I think she tells us it's like she's really got control of how to…not tell a story but her use of repetition is very good. It's not just simple use; it's kind of drawing you … 'I like to do this, I like to do that' – it's got a kind of appealing, rhythmic quality to it. It's not just somebody who's repeating words because they can't think of others she knows how to control repetition for effect so I put that down for a feature of fluency.*

Another aspect of the *Fluency and coherence* scale that examiners attended to was the use of discourse markers and connectives. They valued the use of a range of discourse markers and connectives, and evaluated negatively their incorrect use and the overuse or repetitive use of only a few basic ones.

Coherence was addressed in terms of a) the relevance or appropriateness of candidates' responses and b) topic development and organisation. Examiners referred to candidates being on task or not ("answering the question"), and to the logic of what they were saying. They commented negatively on poor topic organisation or development, particularly the repetition of ideas ("going around in circles") or introduction of off-topic information ("going off on a tangent"), and on the impact of this on the coherence or comprehensibility of the response. At times examiners struggled to decide whether poor topic development was a content issue or a language issue. It was also noted that topic development favours more mature candidates.

A final aspect of *Fluency and coherence* that examiners mentioned was candidates' ability, or willingness, to produce extended turns. They made comments such as "able to keep going" or "truncated". The use of terms such as "struggling" showed their attention to the amount of effort involved in producing longer turns. They also commented unfavourably on speech which was disjointed or consisted of sentence fragments, and on speech where candidates kept on keep adding phrases to a sentence or when they ran too many ideas together into one sentence.

5.1.1b Determining levels within the fluency and coherence scale

To determine how examiners coped with the different levels within the *Fluency and coherence* scale, the verbal report data were analysed for evidence of how the different levels were interpreted. Examiners also commented on problems they had distinguishing levels. In the questionnaire the examiners were asked whether each scale discriminated across the levels effectively and, if not, why.

In general, hesitancy and repetition were key features at all levels, with levels being distinguished by the *frequency* of hesitation and repetition and its *impact* on the clarity or coherence of speech. At the higher levels (Bands 7–9), examiners use terms like "relaxed" and "natural" to refer to fluency. Candidates at these levels were referred to as being "in control".

Examiners appeared uncomfortable about giving the highest score (Band 9), and spent some time trying to justify their decisions. One examiner reported that the fact that Band 9 was "absolute" (that is, required *all* hesitation to be content-related) was problematic (Extract 3), as was distinguishing what constituted appropriate hesitation, given that native speakers can be disfluent. Examiners also expressed similar difficulties with the differences between Bands 7 and 8, where they reported uncertainty as to the cause of hesitation (whether it was grammar, lexis, or content related, see Extract 4).

Extract 3

> *Now I find in general, judgements about the borderline between 8 and 9 are about the hardest to give and I find that we're quite often asked to give them. And the reason they're so hard to give is that on the one hand, the bands for the 9 are stated in the very absolute sense. Any hesitation is to prepare the content of the next utterance for Fluency and coherence, for example. What've we got – all contexts and all times in lexis and GRA. Now as against that, you look at the very bottom and it says, a candidate will be rated on their average performance across all parts of the test. Now balancing those two factors is very hard. You're being asked to say, well does this person usually never hesitate to find the right word? Now that's a contradiction and I think that's a real problem with the way the bands for 9 are written, given the context that we're talking about average performance.*

Extract 4

> *It annoys me between 7 and 8 here. Where it says – I think I alluded to it before – is it content related or is it grammar and vocab or whatever? It says here in 7: 'Some hesitation accessing appropriate language'. And I don't know whether it's content or language for this bloke. So you know I went down because I think sometimes it is language, but I really don't know. So I find it difficult to make that call and that's why I gave it a 7 because I called it that way rather than content related, so being true to the descriptor.*

The examiners appeared to have some difficulty distinguishing Bands 8 and 9 in relation to topic development, which was expected to be good in both cases. At Band 7, examiners reported problems starting to appear in the coherence and/or the extendedness of talk.

At Band 6, examiners referred to a lack of directness (Extract 5), poor topic development (Extract 6) and to candidates "going off on a tangent" or otherwise getting off the topic, and to occasional incoherence. They referred to a lack of confidence, and speech was considered "effortful". Repetition and hesitation or pausing was intrusive at this level (Extract 6). As described in the scales, an ability to "keep going" seemed to distinguish a 6 from a 5 (Extract 7).

Extract 5

> *And I found that she says a lot but she doesn't actually say anything; it takes so long to get anywhere with her speech.*

Extract 6

6 for Fluency and coherence. She was very slow, very hesitant. I felt that her long searches, her low pauses were searching for right words. And I felt that there was little topic development; that she wasn't direct.

Extract 7

I ended up giving him a 6 for Fluency and coherence. I wasn't totally convinced but by-and-large he was able to keep going.

At Band 5, examiners commented on having to work hard to understand the candidate. All of them expressed an inability at times to follow what the candidate was saying. Other comments related to the degree of repetition, hesitation and pausing, the overuse of particular discourse markers, not answering the question, and occasional trouble keeping going, being able to elaborate or take long turns (Extracts 8-10).

Extract 8

So I've given it 5 for fluency and I guess the deciding factor there was the 4 says unable to keep going without noticeable pauses, and she was able to keep going. There were pauses and all that but she did keep going, so I had to give her a 5 there.

Extract 9

Overuse of certain discourse markers, connectives and other cohesive features. He was using the same ones again and again and again.

Extract 10

So she got a 5 for Fluency and coherence because she was usually able to keep going, but there was repetition and there were hesitations mid-sentence while she looked for fairly basic words and grammar, and then she would stop as if she had more to say but she couldn't think of the words. I think there is a category for that in the descriptor, but anyway…

Extract 11

It's interesting in this section, the fluency tends to drop away and I don't know whether it's just that he doesn't like the topic of soccer very much, so maybe I'm doing him an injustice but I'm going to end up marking him down to 7 on Fluency whereas before I was tending more to an 8 but I felt that if he was really fluent he'd be able to sustain it a little bit better.

5.1.1c Confidence in using the fluency and coherence scale

When asked to judge their confidence in understanding and interpreting the scales, no examiner selected lower than the mid-point on a scale of 1 (Not at all confident) to 5 (Very confident) for any of the scales (see Table 4). Examiners were marginally the least confident about *Fluency and coherence*, and the most confident about *Pronunciation*. When asked to elaborate on why they felt confident or not confident about the *Fluency and coherence* scale several commented, as they had also done in their verbal reports, that the focus on hesitation was problematic because it was necessary but not always possible to infer its cause – a search for content or language – in order to decide whether a rating of 7 or 8 should be given. One examiner commented that "there can at times be more witchcraft than science involved in discerning why hesitation is used". It was also noted that fluency can be affected by familiarity with or liking of a topic (Extract 11). Another commented that assessing whether speech is "situationally appropriate" is problematic given the restricted context of the interview, while another said that topic development being mentioned at Band 7 but not Band 8 is problematic. One examiner remarked that the *Fluency and coherence* descriptors are longer than the others and thus harder to "internalise".

Only one examiner reported that the *Fluency and coherence* scale was easy to apply, commenting that the key indicators for her were whether the candidate could or could not *keep going* and could or could not *elaborate*.

	Examiner						
	1	**2**	**3**	**4**	**5**	**6**	**Mean**
Fluency and coherence	4	5	4	3	4	3	3.8
Lexical resources	5	4	3	4	4	4	4.0
Grammatical range and accuracy	5	4	4	3	4	4	4.0
Pronunciation	4.5	5	5	5	3	3	4.3

Table 4: Confidence using the scales

When asked in the questionnaire whether the descriptors of the *Fluency and coherence* scale capture the significant performance qualities at each of the band levels and distinguished adequately between levels, most examiners reported problems. Bands 6 and 7 were considered difficult to distinguish in terms of the frequency or amount of hesitation, repetition and/or self-correction. The terms "some" (Band 7) and "at times" (Band 6) were said to be very similar. One examiner said it was difficult to infer intentions ("willingness" and "readily") to discriminate between Bands 6 and 7.

Distinguishing Bands 7 and 8 was also considered problematic for two reasons: firstly, because topic development is mentioned at Band 7 but not Band 8, but also because, as noted earlier, examiners found it difficult to infer whether disfluency was caused by a search for language (Band 7) or by the candidate thinking about their response (Band 8).

One examiner felt that Bands 4 and 5 were particularly difficult to distinguish because hesitation and repetition are "the hallmarks of both"; another reported problems distinguishing Band 4 and Band 6 ("even 6 versus 4 can produce problems … "*coherence may be lost*" at 6 and "*some breakdowns in coherence*" at 4"). Finally, it was noted that hesitation and pace may be an indication of limitations in language but often reflected individual habits of speech.

5.1.2 Lexical resource

5.1.2a Understanding the lexical resource scale

As was the case for the *Fluency and coherence* scale, examiners tended to refer to the range of features referred to in the descriptors and the key indicators. These included lexical errors, range of lexical resource (including stylistic choices and adequacy for different topics), the ability to paraphrase, and the use of collocations. One feature included in the key indicators but *not* referred to was the ability to convey attitude. Although not referred to in the scales, the examiners took candidates' lack of comprehension of interviewer talk as evidence of limitations in *Lexical resource*.

As expected, there were numerous references to the sophistication and range of the lexis used by candidates, and to inaccuracies or inappropriate word choice. When they referred to inaccuracies or inappropriateness, examiners commented on their frequency ("occasional errors"), their seriousness ("a small slip"), the type of error ("basic", "simple", "non-systematic") and the impact the errors had on comprehensibility. Examiners also commented on the appropriateness or correctness of collocations, and on morphological errors **(**the use of *dense* instead of *density*). They commented unfavourably on candidates' inability to "find the right words", a feature which overlapped with assessments of fluency.

While inaccuracies or inappropriate word choice were typically taken as evidence of lexical limitations, it was also recognised that unusual lexis or use of lexis may in fact be normal in the candidate's dialect or style. This was particularly the case for candidates from the Indian sub-continent. The evidence suggests, however, that determining whether a particular word or phrase was dialectal or inappropriate was not necessarily straightforward (Extract 12).

Extract 12

> *That's her use of "in here" and she does it a lot. I don't know whether it's a dialect or whether it's a systematic error.*

As evidence of stylistic control, examiners commented on a) the use of specific, specialist or technical terms, and b) the use of idiomatic or colloquial terms. They also evaluated the adequacy of candidates' vocabulary for the type of topic (described in terms of *familiar*, *unfamiliar*, *professional*, etc). There was some uncertainty as to whether candidates' ability to use specialist terms within their own professional, academic or personal fields of interest was indicative of a broad range of lexis or whether, because the topic was 'familiar', it was not. Reference was also made to the impact of errors or inappropriate word use on comprehensibility. Finally, although there were not a huge number of references to learned expressions or 'formulae', examiners typically viewed their use as evidence of vocabulary limitations (Extract 13), especially if the use of relatively sophisticated learned phrases contrasted with otherwise unsophisticated usage.

Extract 13

> *Very predictable and formulaic kind of response: "It's a big problem" and "I'm not sure about the solution" kind of style, which again suggests very limited lexis and probably pre-learnt.*

Examiners also attended to candidates' ability to paraphrase when needed (Extract 14). They drew attention to specific instances of what they considered to be successful or creative circumlocution, "my good memory moment" or "the captain of a company".

Extract 14

> *He rarely attempts paraphrase, he sort of stops, can't say it and he doesn't try and paraphrase it; he sort of repeats the bit that he didn't say right.*

5.1.2b Determining levels within the lexical resource scale

The verbal report and questionnaire data were next analysed for evidence of how examiners distinguished levels within the *Lexical resource* scale and what problems they had distinguishing them.

Examiners tended to value "sophisticated" or idiomatic lexical use at the higher end (Extract 15), although they tended to avoid Band 9 because of its 'absoluteness'. Band 8 was awarded if they viewed non-native usage as "occasional errors" (Extract 16), and Band 9 if they considered them to be dialectal or "creative". Precise and specific use of lexical items was also important at the higher levels, as per the descriptors.

Extract 15

> *... and very sophisticated use of common, idiomatic terms. He was clearly 8 in terms of lexical resources.*

Extract 16

Then with Lexical resource, occasionally her choice of word was slightly not perfect and that's why she didn't get a 9 but she really does nice things that shows that she's got a lot of control of the language – like at one stage she says that something "will end" and then she changed it and said it "might end", and that sort of indicated that she knew about the subtleties of using; the impact of certain words.

At Band 7 examiners noted style and collocation. They still looked for sophisticated use of lexical items, although in contrast with Band 8, performance was considered uneven or patchy (Extract 17). They also noticed occasional difficulty elaborating or finding the words at Band 7.

Extract 17

So unusual vocabulary there; it's suitable and quite sophisticated to say "eating voluptuously", so eating for the joy of eating. So this is where my difficulty in assessing her lexical resource. She'll come out with words like that which are really quite impressive but then she'll say "the university wasn't published", which is quite inappropriate and distracting. So yes, at this stage I'm on a 7 for Lexical resource.

Whereas Band 7 required appropriate use of idiomatic language, at Band 6 examiners reported errors in usage (Extract 18). Performance at Band 6 was also characterised by "adequate" or "safe" use of common lexis.

Extract 18

Lexical resource was very adequate for what she was doing. She used a few somewhat unusual and idiomatic terms and there were points where therefore I was torn between a 6 and a 7. The reason I erred on the side of the 6 rather than the 7 was because those idiomatic and unusual terms were sometimes themselves not used quite correctly and that was a bit of a giveaway, it just wasn't quite the degree of comfort that I'd have expected with a 7.

A Band 5 was typically described in terms of the range of lexis ("simple"), the degree of struggle involved in accessing it, and the inability to paraphrase. At this level candidates were seen to struggle for words and there was some lack of clarity in meaning (Extract 19).

Extract 19

It's pretty simple vocabulary and he's struggling for words, at times for the appropriate words, so I'd say 5 on Lexical resource.

Examiners awarded Band 4 when they felt candidates were unable to elaborate, even on familiar topics (Extract 20) and when they were unable to paraphrase (Extract 21). They also noted repetitive use of vocabulary.

Extract 20

So she can tell us enough to tell us that the government can't solve this problem but she hasn't got enough words to be able to tell us why, so it's like she can make the claims but she can't work on the meaning to build it up, even when she's talking about something fairly familiar.

Extract 21

I did come down to a 4 because resource was 'sufficient for familiar topics' but really only basic meaning on unfamiliar topics, which is number 4. 'Attempts paraphrase' – well she didn't really, she couldn't do that. So I felt that she fitted a 4 with the Lexical resource.

5.1.2c Confidence in using the lexical resource scale

The examiners reported being slightly more comfortable with the *Lexical resource* scale than they were with the *Fluency and coherence* scale (Table 4). Three of them noted that it was clear, and the bands easily distinguishable. One noted that it was easy to check "depth" or "breadth" of lexical knowledge with a quick replay of the taped interview, focusing on the candidate's ability to be specific. When asked to elaborate on what they felt the least confident about, examiners commented on:

- the lack of interpretability of terms used in the scales (terms such as *sufficient*, *familiar* and *unfamiliar*)
- the difficulty they had distinguishing between levels (specifically, the similarity between Band 7 *Resource flexibly used to discuss a variety of topics* and Band 6 *Resource sufficient to discuss at length*), and
- the difficulty distinguishing between *Fluency and coherence* and *Lexical resource* (discussed in more detail later).

In relation to this last point, one examiner remarked that spoken discourse markers and other idiomatic items such as adverbials ("possibly", "definitely"), emphatic terms ("you know", "in a sense") and intensifiers or diluters ("really", "somewhat", "quite") are relevant to both *Lexical resources* and *Fluency and coherence.* One examiner commented that paraphrasing is difficult to assess as not all candidates do it, and that indicators such as repetition and errors are more useful. In contrast, another commented that the paraphrase criterion *was* useful, particularly across Bands 4 to 7. Another remarked that it is difficult to assess lexical resources in an interview and that the criteria should focus more on the relevance or appropriateness of the lexis to the context.

When asked whether the descriptors of the *Lexical resource* scale capture the significant performance qualities at each of the Band levels and discriminate across the levels effectively, most examiners said that they felt that this was the case. One said the scale developed well – from *basic meaning* at 4 through *sufficient* at 5 to *meaning clear*, and then higher levels of *idiom* and *collocation* etc. One felt that a clearer distinction was needed in relation to paraphrase for Bands 7 and 8, and another that Bands 5 and 6 were difficult to distinguish because *the ability to paraphrase*, which seemed to be a key 'cut off', was difficult to judge. Another felt that deciding what was a familiar or unfamiliar topic was problematic, particularly across Bands 4 and 5. One examiner did not like the use of the term "discuss" at Band 5, as this for her implied dealing in depth with an issue, something she felt was unlikely at that level. She suggested the term "talk about". Another commented that some candidates have sophisticated vocabulary relating to specific areas of work or study yet lack more general breadth.

5.1.3 Grammatical range and accuracy

5.1.3a Understanding the grammatical range and accuracy scale

In general, the examiners were very true to the descriptors and all aspects of the *Grammatical range and accuracy* scale were addressed. The main focus was on error frequency and error type on the one hand, and complexity of sentences and structures on the other. Examiners appeared to balance these criteria against each other.

In relation to grammatical errors, examiners referred to density or frequency, including the number of error-free sentences. They also noted the type of error – those viewed as simple, basic, or minor included articles, tense, pronouns, subject-verb agreement, word order, plurals, infinites and participles – and whether they were systematic or not. They also noted the impact of errors on intelligibility.

The examiners commented on the range of structures used, and the flexibility that candidates demonstrated in their use. There was reference, for example, to the repetitive use of a limited range of structures, and to candidates' ability to use, and frequency of use of, complex structures such as passive, present perfect, conditional, adverbial constructions, and comparatives. Examiners also noted

candidates' ability to produce complex sentences, the range of complex sentence types they used, and the frequency and success with which they produced them. Conversely, what they referred to as *fragmented* or *list-like* speech or the inability to produce complete sentences or connect utterances (a feature which also impacted on assessments of coherence) was taken as evidence of limitations in grammatical resources.

5.1.3b Determining levels within the grammatical range and accuracy scale

To determine how examiners coped with the different levels within the *Grammatical range and accuracy* scale, the verbal report data were analysed for evidence of how the different levels were interpreted. Again Band 9 was used little. This seemed to be because of its "absolute" nature; the phrase "at all times" was used to justify *not* awarding this Band (Extract 22). Examiners did have some problems deciding whether non-native usage was dialectal or error. At Band 8, examiners spoke of the complexity of structures and the "flexibility" or "control" the candidates displayed in their use of grammar. At this level errors were expected to be both occasional non-systematic, and tended to be referred to as "inappropriacies" or "slips", or as "minor", "small", or "unusual" (for the candidate), or as "non-native like" usage.

Extract 22

> *And again I think I'm stopping often enough for these grammatical slips for it on average, remembering that we are always saying that, for it on average to match the 8 descriptor which allows for these, than the 9 descriptor which doesn't.*

Overall, Band 7 appeared to be a default level; not particularly distinguishable but more a middle ground between 8 and 6, where examiners make a decision based on whether the performance is as good as an 8 or as bad as a 6. Comments tended to be longer as examiners tended to argue for a 7 and against a 6 and an 8 (Extract 23). At this level inaccuracies were expected but they were relatively unobtrusive, and some complex constructions were expected (Extract 24).

Extract 23

> *I thought that he was a 7 more than a 6. He definitely wasn't an 8, although as I say, at the beginning I thought he might have been. There was a 'range of structures flexibly used'. 'Error free sentences frequent', although I'm not a hundred per cent sure of that because of pronunciation problems. And he could use simple and complex sentences effectively, certainly with some errors. Now when you compare that to the criteria for 6: 'Though errors frequently occur in complex structures these rarely impede communication …'*

Extract 24

> *For Grammatical range and accuracy, even though there was* [sic] *certainly errors, there was certainly still errors, but you're allowed that to be a 7. What actually impressed me here … he was good on complex verb constructions with infinitives and participles. He had a few really quite nice constructions of that nature which, I mean there we're talking about sort of true complex sentences with complex verbs in the one clause, not just subordinate clauses, and I thought they were well handled. His errors certainly weren't that obtrusive even though there were some fairly basic ones, and I think it would be true to say that error-free sentences were frequent there.*

At Band 6 the main focus for examiners was the type of errors and whether they impeded communication. While occasional confusion was allowed, if the impact was too great then examiners tended to consider dropping to a 5 (Extract 25). Also, an inability to use complex constructions successfully and confidently kept candidates at 6 rather than a 7 (Extract 26).

Extract 25

A mixture of short sentences, some complex ones, yes variety of structures. Some small errors, but certainly not errors that impede communication. But not an advanced range of sentence structures. I'll go for a 6 on the grammar.

Extract 26

Grammatical range and accuracy was also pretty strong, relatively few mistakes, especially simple sentences were very well controlled. Complex structures. The question was whether errors were frequent enough for this to be a 6, there certainly were errors. There were also a number of quite correct complex structures. I did have misgivings I suppose about whether this was a 6 or a 7 because she was reasonably correct. I suppose I eventually felt the issue of flexible use told against the 7 rather than the 6. There wasn't quite enough comfort with what she was doing with the structures at all times for it to be a 7.

At Band 5 examiners noted frequent and basic errors, even in simple structures, and errors were reported as frequently impeding communication. Where attempts were made at more complex structures these were viewed as limited, and tended to lead to errors (Extract 27) Speech was fragmented at times. Problems with the verb 'to be' or sentences without verbs were noted.

Extract 27

She had basic sentences, she tended to use a lot of simple sentences but she did also try for some complex sentences, there were some there, and of course the longer her sentences, the more errors there were.

The distinguishing feature of Band 4 appeared to be that basic and systematic errors occurred in most sentences (Extract 28).

Extract 28

Grammatical range and accuracy, I gave her a 4. Even on very familiar phrases like where she came from, she was missing articles and always missed word-ending 's'. And the other thing too is that she relied on key words to get meaning across and some short utterances were error-free but it was very hard to find even a basic sentence that was well controlled for accuracy.

5.1.3c Confidence in using the grammatical range and accuracy scale

When asked to comment on the ease of application of the *Grammatical range and accuracy* scale, one examiner remarked that is easier to notice specific errors than error-free sentences, and another that errors become less important or noticeable if a candidate is fluent. Three examiners found the scale relatively easy to use.

Most examiners felt that the descriptors of the scale captured the significant performance qualities at each of the Band levels. One examiner said that he distinguished levels primarily in terms of the degree to which errors impeded communication. Another commented that the notion of "error" in speech can be problematic as natural speech flow (ie native) is often not in full sentences and is sometimes grammatically inaccurate.

When asked whether the *Grammatical range and accuracy* scale discriminates across the levels effectively, three agreed and three disagreed. One said that terms such as *error-free*, *frequently*, and *well controlled* are difficult to interpret ("I ponder on what per cent of utterances were frequently error-free or well controlled"). Another felt that Bands 7 and 8 were difficult to distinguish because he was not sure whether a minor systematic error would drop the candidate to 7, and that Bands 5 and 6 could also be difficult to distinguish. Another felt that the Band 4/5 threshold was problematic because some candidates can produce long turns (Band 5) but are quite inaccurate even in basic sentence forms

(Band 4). Finally, one examiner remarked that a candidate who produces lots of structures with a low level of accuracy, even on basic ones, can be hard to place, and suggested that some guidance on "risk takers" is needed.

5.1.4 Pronunciation

5.1.4a Understanding the pronunciation scale

When evaluating candidates' pronunciation, examiners focused predominantly on the impact of poor pronunciation on intelligibility, in terms of both *frequency* of unintelligibility and the *amount of strain* for the examiner (Extract 29).

Extract 29

> *I really do rely on that 'occasional strain', compared to 'severe strain'. [The levels] are clearly formed I reckon.*

When they talked about specific aspects of pronunciation, examiners referred most commonly to the production of sounds, that is, vowels and consonants. They did also, at times, mention stress, intonation and rhythm, and while they again tended to focus on errors there was the occasional reference to the use of such features to enhance the communication (Extract 30).

Extract 30

> *And he did use phonological features in a positive way to support his message. One that I wrote down for example was 'well nobody was not interested'. And he got the stress exactly right and to express a notion which was, to express a notion exactly. I mean he could have said 'everybody was interested' but he actually got it exactly right, and the reason he got it exactly right among other things had to do with his control of the phonological feature.*

5.1.4b Determining levels within the pronunciation scale

Next the verbal report and questionnaire data were analysed for evidence of how the different levels were interpreted and problems that examiners had distinguishing levels. While they attended to a range of phonological features – vowel and consonant production but also stress and rhythm – *intelligibility*, or the *level of strain* involved in understanding candidates appeared to be the key feature used to determine level (Extract 33). Because only even numbered bands could be awarded, it seemed that examiners took into account the impact that the *Pronunciation* score might have on overall scores (Extract 34).

Extract 31

> *I really do rely on that 'occasional strain', compared to 'severe strain'.*

Extract 32

> *So I don't know why we can't give those bands between even numbers. So, just as I wanted to give a 5 to the Indian I want to give a 9 to this guy. Because you see the effect of 9, 9, 8, 8 will be he'll come down to 8 probably, I'm presuming.*

At Band 8 examiners tended to pick out isolated instances of irregular pronunciation, relating the impact of these on intelligibility to the descriptors: *minimal impact* and *accent present but never impedes communication*. Although the native speaker was referred to as the model, it was recognised that native speakers make occasional pronunciation errors (Extract 35). Occasional pronunciation errors were generally considered less problematic than incorrect or non-native stress and rhythm (Extract 36) One examiner expressed a liking for variety of tone or stress in delivery and noted that she was reluctant to give an 8 to a candidate she felt sounded bored or disengaged.

Extract 33

Because I suppose the truth is, as native speakers, we sometimes use words incorrectly and we sometimes mispronounce them.

Extract 34

It's interesting how she makes errors in pronunciation on words. So she's got "bif roll" and "steek" and "selard" and I don't think there is much of a problem for a native speaker to understand as if you get the pauses in the wrong place, if you get the rhythm in the wrong place… so that's why I've given her an 8 rather than dropping her down because it says 'L1 accent may be evident, this has minimal effect on intelligibility', and it does have minimal effect because it's always in context that she might get a word mispronounced or pronounced in her way, not my way.

Band 6 appeared to be the 'default' level where examiners elect to start. Examiners seemed particularly reluctant to give 4; of the 29 ratings, only three were below 6. Bands 4 and 6 are essentially determined with reference to listener strain, with severe strain at Band 4 and occasional strain at Band 6 (Extract 37).

Extract 35

Again with Pronunciation I gave her a 6 because I didn't find patches of speech that caused 'severe strain', I mean there was 'mispronunciation causes temporary confusion', some 'occasional strain'.

At Band 4 most comments referred to severe strain, or to the fact that examiners were unable to comprehend what the candidate had said (Extract 38).

Extract 36

I actually did mark this person down to Band 4 on Pronunciation because it did cause me 'severe strain', although I don't know whether that's because of the person I listened to before, or the time of the day but there were large patches, whole segments of responses that I just couldn't get through and I had to listen to it a couple of times to try and see if there was any sense in it.

5.1.4c Confidence in using the pronunciation scale

When asked to judge their confidence in understanding and interpreting the scales, the examiners were the most confident about *Pronunciation* (see Table 4). However, there was a common perception that the scale did not discriminate enough (Extract 31). One examiner remarked that candidates most often came out with a 6, and another that she doesn't take pronunciation as seriously as the other scales. One examiner felt that experience with specific language groups could bias the assessment of pronunciation (and, in fact, there were a number of comments in the verbal report data where examiners commented on their familiarity with particular accents, or their lack thereof). One was concerned that speakers of other Englishes may be hard to understand and therefore marked down unfairly (Extract 32). Volume and speed were both reported in the questionnaire data and verbal report data as having an impact on intelligibility.

Extract 37

And I would prefer to give a 5 on Pronunciation but it doesn't exist. But to me he's somewhere between 'severe strain', which is the 4, and the 6 is 'occasional strain'. He caused strain for me nearly 50% of the time, so that's somewhere between occasional and severe. And this is one of the times where I really wish there was a 5 on Pronunciation because I think 6 is too generous and I think 4 is too harsh.

Extract 38

> *I think there is an issue judging the pronunciation of candidates who may be very difficult for me to understand, but who are fluent/accurate speakers of recognised second language Englishes, (Indian or Filipino English). A broad, Scottish accent can affect comprehensibility in the Australian context and I'm just not sure therefore, whether an Indian or Filipino accent affecting comprehensibility should be deemed less acceptable.*

While pronunciation was generally considered to be the easiest scale on which to distinguish Band levels because there are fewer levels, four of the six examiners remarked that there was too much distinction between levels, not too little, so that the scale did not discriminate between candidates enough. One examiner commented that as there is really no Band 2, it is a decision between 4, 6, or 8, and that she sees 4 as "almost unintelligible". In arguing for more levels they made comments like: "Many candidates are Band 5 in pronunciation – between *severe strain* for the listener and *occasional.* Perhaps *mild strain quite frequently*, or *mild strain in sections of the interview*. One examiner felt a Band 9 was needed (Extract 39).

Extract 39

> *Levels 1,3,5,7 and 9 are necessary. It seems unfair not to give a well-educated native speaker of English Band 9 for pronunciation when there's nothing wrong with their English, Australian doctors going to UK.*

Examiners commented at times on the fact that they were familiar with the pronunciation of candidates of particular nationalities, although they typically claimed to take this into account when awarding a rating (Extract 40).

Extract 40

> *I found him quite easy to understand but I don't know that everybody would and there's a very strong presence of accent or features of pronunciation that are so specifically Vietnamese that they can cause other listeners problems. So I'll go with a 6.*

5.2 The discreteness of the scales

In this section, the questionnaire data and, where relevant, the analysis of the verbal report data were drawn upon to address the question of the ease with which examiners were able to distinguish the four analytic scales – *Fluency and coherence* (F&C); *Grammatical range and accuracy* (GRA); *Lexical resource* (LR); and *Pronunciation* (P).

The examiners were asked how much overlap there was between the scales on a range of 1 (*Very distinct*) to 4 (*Almost total overlap*), see Table 5. The greatest overlap (mean 2.2) was reported between *Fluency and coherence* and *Grammatical range and accuracy*. Overall, *Fluency and coherence* was considered to be the least distinct and *Pronunciation* the most distinct scale.

	Examiner						
Scale overlap	**1**	**2**	**3**	**4**	**5**	**6**	**Mean**
F&C and LR	1	2	2	2	3	2	2.0
F&C and GRA	3		2	2	2	2	2.2
F&C and P	2		2	2	1	2	1.8
LR and GRA	2	2	2	1	2	2	1.8
LR and P	1		1	1	1	1	1.0
GRA and P	1		1	1	1	1	1.0

Table 5: Overlap between scales

When asked to describe the nature of the overlap between scales, the examiners responded as follows. Comments made during the verbal report session supported these responses,

Overlap: Fluency and coherence / Lexical resource

Vocabulary was seen as overlapping with fluency because "to be fluent and coherent [candidates] need the lexical resources", and because good lexical resources allow candidates to elaborate their responses. Two examiners pointed out that discourse markers (and, one could add, connectives), which are included under *Fluency and coherence,* are also lexical items. Another examiner commented that the use of synonyms and collocation helps fluency.

Overlap: Fluency and coherence / Grammatical range and accuracy

Grammar was viewed as overlapping with fluency because if a candidate has weak grammar but a steady flow of language, coherence is affected negatively. The use of connectives ("so", "because") and subordinating conjunctions ("when", "if") was said to play a part in both sets of criteria. Length of turn in *Grammatical range and accuracy* was seen as overlapping with *the ability to keep going* in *Fluency and coherence* (Extract 41).

Extract 41

> *Again I note both with fluency and with grammar the issue of the length of turns kind of cuts across both of them and I'm sometimes not sure whether I should be taking into account both of them or if not which for that, but as far as I can judge it from the descriptors, it's relevant to both.*

One examiner remarked that fluency can dominate the other criteria, especially grammar (Extract 42).

Extract 42

> *Well I must admit that I reckon if the candidate is fluent, it does tend to influence the other two scores. If they keep talking you think 'oh well they can speak English'. And you have to be really disciplined as an examiner to look at those other – the lexical and the grammar – to really give them an appropriate score because otherwise you can say 'well you know they must have enough vocab I could understand them'. But the degree to which you understand them is the important thing. So even as a 4 I said that I think there also needs to be some other sort of general band score. It does make you focus on those descriptors here.*

Overlap: Lexical resource / Grammatical range and accuracy

Three examiners wondered whether errors in expressions or phrases (preposition phrases, phrasal verbs, idioms) were lexical or grammatical ("If a candidate says *in the moment* instead of *at the moment*, what is s/he penalised under?" and "*I'm one of those lucky persons* – Is it lexical? Is it expression?") Another examiner saw the scales as overlapping in relation to skill at paraphrasing.

Overlap: Fluency and coherence / Pronunciation

Two examiners pointed out that if the pronunciation is hard to understand the coherence will be low. Another felt that slow speech (disfluent) was often more clearly pronounced and comprehensible, although another felt that disfluent speech was *less* comprehensible if there was "a staccato effect".

One examiner remarked that if pronunciation is unintelligible it is not possible to accurately assess *any* of the other areas.

5.3 Remaining questions

5.3.1 Additional criteria

As noted earlier, during the verbal report session, examiners rarely made reference to features not included in the scales or key criteria. Those that examiners did refer to were:

- the ability to cope with different functional demands
- confidence in using the language, and
- creative use of language.

In response to a question about the appropriateness of the scale contents, the following additional features were proposed as desirable: voice; engagement; demeanour; and paralinguistic aspects of language use. Three examiners criticised the test for not testing "communicative" language. One examiner felt there was a need for a holistic rating in addition to the analytic ratings because global marking was less accurate than profile marking "owing to the complexity of the variables involved".

5.3.2 Irrelevant criteria

When asked whether any aspects of the descriptors were inappropriate or irrelevant, one examiner remarked that candidates may not exhibit all aspects of particular band descriptors. Another saw conflict between the "absolute nature of the descriptors for Bands 9 and 1 and requirement to assess on the basis of 'average' performance across the interview".

When asked whether they would prefer the descriptors to be shorter or longer, most examiners said they were fine. Three remarked that if a candidate must fully fit all the descriptors at a particular level, as IELTS instructs, it would create more difficulties if descriptors were longer. One examiner said that the *Fluency and coherence* descriptors could be shorter and should rely less on discerning the cause of disfluency, whereas another remarked that more precise language was needed in *Fluency and coherence* Bands 6 and 7. Another referred to the need for more precise language in general. One examiner suggested that key 'cut off' statements would be useful, and another that an appendix to the criteria giving specific examples would help.

5.3.3 Interviewing and rating

While they acknowledged that it was challenging to conduct the interview and rate the candidate simultaneously, the examiners did not feel it was inappropriately difficult. In part, this was because they had to pay less attention to managing the interaction and thinking up questions than they did in the previous interview, and in part because they were able to focus on different criteria in different sections of the interview, while the monologue turn gave then ample time to focus exclusively on rating. When asked whether they attended to specific criteria in specific parts of the interview, some said "yes" and some "no".

They also reported different approaches to arriving at a final rating. The most common approach was to make a tentative assessment in the first part and then confirm this as the interview proceeded (Extract 43). One reported working down from the top level, and another making her assessment after the interview was finished.

Extract 43

> *By the monologue I have a tentative score and assess if I am very unsure about any of the areas. If I am, I make sure I really focus for that in the monologue. By the end of the monologue, I have a firmer feel for the scores and use the last section to confirm/disconfirm. It is true that the scores do change as a candidate is able to demonstrate the higher level of language in the last section. I do have some difficulties wondering what weight to give to this last section.*

When asked if they had other points to make, two examiners remarked that the descriptors could be improved. One wanted a better balance between "specific" and "vague" terms, and the other "more distinct cut off points, as in the writing descriptors". Two suggested improvements to the training: the use of video rather than audio-recordings of interviews, and the provision of examples attached to the criteria. Another commented that "cultural sophistication" plays a role in constructing candidates as more proficient, and that the test may therefore be biased towards European students ("some European candidates come across as better speakers, even though they may be mainly utilising simple linguistic structures").

6 DISCUSSION

The study addressed a range of questions pertaining to how trained IELTS examiners interpret and distinguish the scales used to assess performance in the revised IELTS interview, how they distinguish the levels within each scale, and what problems they reported when applying the scales to samples of performance.

In general, the examiners referred closely to the scales when evaluating performances, quoting frequently from the descriptors and using them to guide their attention to specific aspects of performance and to distinguish levels. While there was reference to all aspects of the scales and key criteria, some features were referred to more frequently than others. In general, the more 'quantifiable' features such as amount of hesitation (*Fluency and coherence*) or error density and type (*Lexical resource* and *Grammatical range and accuracy*) were the most frequently mentioned, although it cannot be assumed that this indicates greater weighting of these criteria over the less commonly mentioned ones (such as connectives or paraphrasing). Moreover, because examiners are required to make four assessments, one for each of the criteria, it seems that there is less likelihood than was the case previously with the single holistic scale of examiners weighting these four main criteria differentially.

There were remarkably few instances of examiners referring to aspects of performance not included in the scales, which is in marked contrast to the findings of an examination of functioning of the earlier holistic scale (Brown, 2000). In that study Brown reported while some examiners focused narrowly on the criteria, others were "more inference-oriented, drawing more conclusions about the candidates' ability to cope in other contexts" (2000: 78). She noted also that this was the case more for more experienced examiners.

The examiners reported finding the scales relatively easy to use, and the criteria and their indicators to be generally appropriate and relevant to test performances, although they noted some overlap between scales and some difficulties distinguishing levels.

It was reported that some features were difficult to notice or interpret. Particularly problematic features included:

- the need to infer the cause of hesitation (*Fluency and coherence*)
- a lack of certainty about whether inappropriate language was dialectal or error (*Lexical resource* and *Grammatical range and accuracy*)
- a lack of confidence in determining whether particular topics were familiar nor not, particularly those relating to professional or academic areas (*Lexical Resource*).

Difficulty was also reported in interpreting the meaning of "relative" terms used in the descriptors, such as sufficient, adequate, etc. There was some discomfort in the "absoluteness" of the Band 9 descriptors across the scales.

The most problematic scale appeared to be *Fluency and coherence*. It was the most complex in terms of focus and was also considered to overlap the most with other scales. Overlap resulted from the impact of a lack of lexical or grammatical resources on fluency, and because discourse markers and connectives (referred to in the *Fluency and coherence* scale) were also lexical items *and* a feature of complex sentences. Examiners seemed to struggle the most to determine band levels on the *Fluency and coherence* scale, perhaps because of the broad range of features it covers, and the fact that the cause of hesitancy, a key feature in the scale at the higher levels, is a high-inference criterion.

The *Pronunciation* scale was considered the easiest to apply, however the examiners expressed a desire for more levels for *Pronunciation*. They felt it did not distinguish candidates sufficiently and the fewer band levels meant the rating decision carried too much weight in the overall (averaged) score.

As was found in earlier studies of examiner behaviour in the previous IELTS interview (Brown, 2000) and in prototype speaking tasks for Next Generation TOEFL (Brown, Iwashita and McNamara, 2005), in addition to 'observable' features such as frequency of error, complexity and accuracy, examiners were influenced in all criteria by the impact of particular features on *comprehensibility*. Thus they referred frequently to the impact of disfluency, lexical and grammatical errors and non-native pronunciation on their ability to follow the candidate or the degree of strain it caused them.

A marked difference in the present study from that of Brown (2000) was the relevance of interviewer behaviour to ratings. Brown found that a considerable number of comments were devoted to the interviewer and reports that the examiners "were constantly aware of the fact that the interviewer is implicated in a candidate's performance" (2000:74). At times, the examiners even compensated for what they perceived to be unsupportive or less-than-competent interviewer behaviour (see also Brown 2003, 2004). While there were one or two comments on interviewer behaviour in the present study, they did not appear to have any impact on ratings decisions. In contrast, however, some of the examiners did report a level of concern that the current interview and assessment criteria focused less on "communicative" or interactional skills than previously, a result of the use of interlocutor frames.

Finally, although the examiners in this study were rating taped tests conducted by other interviewers, they reported feeling comfortable, (and more comfortable than was the case in the earlier unscripted interview), with simultaneously conducting the interview and assessing it, despite the fact that they were required to focus on four scales rather than one. This seemed to be because they no longer have to manage the interview by developing topics on-the-fly and also have the opportunity during Part 2 (the long turn) to sit back and focus entirely on the candidate's production.

7 CONCLUSION

This study set out to investigate examiners' behaviour and attitudes to the rating task in the IELTS interview. The study was designed as a follow-up to an earlier study (Brown, 2000), which investigated the same issues in relation to the earlier IELTS interview. Two major changes in the current interview are: the use of interlocutor frames to constrain unwanted variation amongst interviewers; and the use of a set of four analytic scales rather than the previous single holistic scale.

The study aimed to derive evidence for or against the validity – the interpretability and ease of application – of these revised scales within the context of the revised interview. To do this, the study drew on two sets of data, verbal reports and questionnaire responses provided by six experienced IELTS examiners when rating candidate performances.

On the whole, the evidence suggested that the rating procedure works relatively well. Examiners reported a high degree of comfort using the scales. The evidence suggested there was a higher degree of consistency in examiners' interpretations of the scales than was previously the case; a finding which is perhaps unsurprising given the more detailed guidance that four scales offer in comparison with a single scale. The problems that were identified – perceived overlap amongst scales, and difficulty distinguishing levels – could be addressed in minor revisions to the scales and through examiner training.

REFERENCES

Brown, A, 1993, 'The role of test-taker feedback in the development of an occupational language proficiency test' in *Language Testing,* vol 10 no 3*,* pp 277-303

Brown, A, 2000, 'An investigation of the rating process in the IELTS Speaking Module' in *Research Reports 1999, vol 3,* ed R Tulloh, ELICOS, Sydney, pp 49-85

Brown, A, 2003a, 'Interviewer variation and the co-construction of speaking proficiency', *Language Testing,* vol 20, no 1, pp 1-25

Brown, A, 2003b, 'A cross-sectional and longitudinal study of examiner behaviour in the revised IELTS Speaking Test', report submitted to IELTS Australia, Canberra

Brown, A, 2004, 'Candidate discourse in the revised IELTS Speaking Test', *IELTS Research Reports 2006, vol 6* (the following report in this volume), IELTS Australia, Canberra, pp 71-89

Brown, A, 2005, *Interviewer variability in oral proficiency interviews*, Peter Lang, Frankfurt

Brown, A and Hill, K, 1998, 'Interviewer style and candidate performance in the IELTS oral interview' in *Research Reports 1997, vol 1*, ed S Woods, ELICOS, Sydney, pp 1-19

Brown, A, Iwashita, N and McNamara, T, 2005, *An examination of rater orientations and test-taker performance on English for Academic Purposes speaking tasks*, TOEFL Monograph series MS-29, Educational Testing Service, Princeton, New Jersey

Cumming, A, 1990, 'Expertise in evaluating second language compositions' in *Language Testing,* vol 7, no 1, pp 31-51

Delaruelle, S, 1997, 'Text type and rater decision making in the writing module' in *Access: Issues in English language test design and delivery*, eds G Brindley and G Wigglesworth, National Centre for English Language Teaching and Research, Macquarie University, Sydney, pp 215-242

Gass, SM and Mackey, A, 2000, *Stimulated recall methodology in second language research*, Lawrence Erlbaum, Mahwah, NJ

Green, A, 1998, *Verbal protocol analysis in language testing research: A handbook*, (*Studies in language testing 5),* Cambridge University Press and University of Cambridge Local Examinations Syndicate, Cambridge

Lazaraton, A, 1996a, 'A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE)' in *Performance testing, cognition and assessment: Selected papers form the 15th Language Testing Research Colloquium*, eds M Milanovic and N Saville, Cambridge University Press, pp 18-33

Lazaraton, A, 1996b, 'Interlocutor support in oral proficiency interviews: The case of CASE' in *Language Testing,* vol 13, pp 151-172

Lewkowicz, J, 2000, 'Authenticity in language testing: some outstanding questions' in *Language Testing,* vol 17 no 1, pp 43-64

Lumley, T and Stoneman, B, 2000, 'Conflicting perspectives on the role of test preparation in relation to learning' in *Hong Kong Journal of Applied Linguistics,* vol 5 no 1, pp 50-80

Lumley, T, 2000, 'The process of the assessment of writing performance: the rater's perspective', unpublished doctoral thesis, The University of Melbourne

Lumley, T and Brown, A, 2004, 'Test-taker response to integrated reading/writing tasks in TOEFL: evidence from writers, texts and raters', unpublished report, The University of Melbourne

McNamara, TF and Lumley, T, 1997, 'The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings' in *Language Testing,* vol 14, pp 140-156

Meiron, BE, 1998, 'Rating oral proficiency tests: a triangulated study of rater thought processes', unpublished Masters thesis, University of California, LA

Merrylees, B and McDowell, C, 1999, 'An investigation of Speaking Test reliability with particular reference to the Speaking Test format and candidate/examiner discourse produced' in *IELTS Research Reports Vol 2*, ed R Tulloh, IELTS Australia, Canberra, pp 1-35

Morton, J, Wigglesworth, G and Williams, D, 1997, 'Approaches to the evaluation of interviewer performance in oral interaction tests' in *Access: Issues in English language test design and delivery,* eds G Brindley and G Wigglesworth, National Centre for English Language Teaching and Research, Macquarie University, Sydney, pp 175-196

Pollitt, A and Murray, NL, 1996, 'What raters reall*y* pay attention to' in *Performance testing, cognition and assessment, (Studies in language testing 3)*, eds M Milanovic and N Saville, Cambridge University Press, Cambridge, pp 74-91

Taylor, L and Jones, N, 2001, *University of Cambridge Local Examinations Syndicate Research Notes 4*, University of Cambridge Local Examinations Syndicate, Cambridge, pp 9-11

Taylor, L, 2000, *Issues in speaking assessment research, (Research notes 1)*, University of Cambridge Local Examinations Syndicate, Cambridge, pp 8-9

UCLES (2001) *IELTS examiner training material*, University of Cambridge Local Examinations Syndicate, Cambridge

Vaughan, C, 1991, 'Holistic assessment: What goes on in the rater's mind?' in *Assessing second language writing in academic contexts,* ed L Hamp-Lyons, Ablex, Norwood, New Jersey, pp 111-125

Weigle, SC, 1994, 'Effects of training on raters of ESL compositions' in *Language Testing,* vol 11, no 2, pp 197-223

APPENDIX 1: QUESTIONNAIRE

A Focus of the criteria

1. Do the four criteria cover features of spoken language that can be readily assessed in the testing situation?
Yes / No Please elaborate _____

_____

2. Do the descriptors relate directly to key indicators of spoken language? Is anything left out?

Yes / No Please elaborate _____

_____

3. Are any aspects of the descriptors inappropriate or irrelevant?

Yes / No Please elaborate _____

_____

B Interpretability of the criteria

4. Are the descriptors easy to understand and interpret? How would you rate your confidence on a scale of 1-5 in using each scale?

	Not at all confident				**Very confident**
Fluency and coherence	1	2	3	4	5
Lexical resource	1	2	3	4	5
Grammatical range and accuracy	1	2	3	4	5
Pronunciation	1	2	3	4	5

5. Please elaborate on why you felt confident or not confident about each of the scales:

Fluency and coherence _____

_____

Lexical resource _____

_____

Grammatical range and accuracy _____

_____

Pronunciation _____

_____

6. How much overlap do you find among the scales?

	Very distinct	Some overlap	A lot of overlap	Almost total overlap
F&C and LR	1	2	3	4
F&C and GRA	1	2	3	4
F&C and P	1	2	3	4
LR and GRA	1	2	3	4
LR and P	1	2	3	4
GRA and P	1	2	3	4

7. Could you describe this overlap? _____

_____

8. Would you prefer the descriptors to be shorter / longer?

Please elaborate _____

_____

C Level distinctions

9. Do the descriptors of each scale capture the significant performance qualities at each of the band levels?

Fluency and coherence Yes / No Please elaborate _____

_____

Lexical resource Yes / No Please elaborate _____

_____

Grammatical range and accuracy Yes / No Please elaborate _____

_____

Pronunciation Yes / No Please elaborate _____

_____

10. Do the scales discriminate across the levels effectively? (If not, for each scale which levels are the most difficult to discriminate, and why?)

Fluency and coherence Yes / No Please elaborate _____

_____

Lexical resource Yes / No Please elaborate _____

_____

Grammatical range and accuracy Yes / No Please elaborate _____

_____

Pronunciation Yes / No Please elaborate _____

_____

11. Is the allocation of bands for pronunciation appropriate?

Yes / No Please elaborate

12. How often do you award flat profiles?

Please elaborate

D The rating process

13. How difficult is it to interview and rate at the same time?

Please elaborate

14. Do you focus on particular criteria in different parts of the interview?

Yes / No Please elaborate

15. How is your final rating achieved? How do you work towards it? At what point do you finalise your rating?

Please elaborate

Final comment

Is there anything else you think you should have been asked or would like to add?