# 3. An Investigation of the Rating Process in the IELTS Oral Interview

*Annie Brown*
*Language Testing Research Centre*
*The University of Melbourne*

## Abstract

Holistic assessments of oral language proficiency are often made in relation to performance in conversational language proficiency interviews, one such example of which is the IELTS Oral Interview. This study seeks to explore the rating practices of trained and accredited IELTS raters when judging candidates' performance in IELTS interviews. In particular, it aims to address questions such as:

- How do raters cope with the task of having to base an assessment of *ability* on a single *performance*?
- What is the relationship of linguistic and non-linguistic aspects of the performance?
- How is the *interlocutor's* performance dealt with in the assessment of the *candidate's* ability?
- Do raters focus on criteria other than those specifically mentioned in the descriptors?
- How salient are the stated criteria?
- Does the same performance elicit judgements of the same kind from different raters?

This study adds to a small but growing body of qualitative research into the judgements made in assessments of second language speaking proficiency. Using data (taped IELTS interviews) collected in an earlier study (Brown and Hill, 1998), eight IELTS raters each rated four interviews selected from a set of eight using the IELTS bandscales. For each interview they provided a verbal protocol where they first summarised the reasons for the score they had awarded and then reviewed the tape in order to identify those features of the rating procedure which influenced their scoring. This methodology is known as stimulated verbal recall (di Pardo, 1994). In these, the raters were asked talk about the judging process and to identify the salient decision-making points of the interview.

The raters were all accredited and practicing IELTS interviewers. The candidates were all overseas students drawn from a pre-university (Foundation) course. At the time of the interviews they were preparing to take IELTS prior to submitting applications for tertiary study in Australia.

The protocols were transcribed and coded. Findings are discussed and implications are drawn regarding the validity of this test format.

IELTS™
English for International Opportunity

## Publishing details

# International English Language Testing System (IELTS)

## Research Reports 2000

## Volume 3

Editor: Robyn Tulloh

IELTS Australia Pty Limited
ABN 84 008 664 766
Incorporated in the Australian Capital Territory
Web: www.ielts.org



National Library of Australia
Cataloguing-in-Publication Data
2000 ed
IELTS Research Reports 2000 Volume 3
ISBN 0 86403 036

# 1.0 Introduction

The conversational language proficiency interview, a face-to-face interview in which an interviewer questions a learner on a number of specified topics, is a popular technique for the assessment of oral language proficiency. The popularity of this technique derives to a large extent from the belief that it provides a context in which candidates' *communicative* and *interactional* skills can be tested. The IELTS Oral Interview is one example of this test genre.

The discourse produced in conversational language proficiency interviews has been the focus of a number of studies, often in response to questions of authenticity or the conversational nature of the interaction (see, for example, Neeson 1985; Perrett 1990; Ross 1992; Ross and Berwick 1992; Young and Milanovic 1992; Filipi 1994; Cafarella 1994; Lazaraton 1993, 1996, 1997). However, despite claims that interactional skills and communicative skills (for example the ability to negotiate meaning, the ability to maintain a conversation) are tapped in conversational interviews, there are as yet relatively few studies of the *rating* process, investigating just what raters take into account when awarding scores, despite a growing interest in general in what raters do (see Pollitt and Murray 1993; Brown 1995; Chalhoub-Deville 1995; Lazaraton 1993, 1996; McNamara and Lumley 1997, Meiron 1998). In particular, in contrast with research into raters' decision making processes in the assessment of writing, there are as yet few published studies which use verbal protocols.

Verbal protocol studies can provide valuable information on aspects of the rating process which quantitative studies of test scores cannot necessarily explore. For example: How do raters cope with the task of having to base a general assessment of *ability* on a single, co-constructed *performance*? How is the *interlocutor's* performance dealt with in the assessment of a *candidate's* ability? and What is the relationship between *linguistic* and *non-linguistic* aspects of the performance?

This study adds to a small but growing body of qualitative research into the judgements made in assessments of second language speaking proficiency. Retrospective verbal protocols provided by a group of trained IELTS raters are analysed in order to investigate how the construct of oral language ability is understood, how linguistic and other criteria contribute to raters' judgements, and which aspects of candidates' performances are salient to these judgements. In other words, it seeks to shed some light on the question *What does it mean to be proficient?* in the context of the IELTS oral interview

The study seeks in particular to respond to a range of questions raised in earlier studies of both speaking and writing assessment. Researchers have commented, for example, on the existence of 'implicit' criteria, criteria which are not explicitly stated in the band descriptors. They have also commented on the fact that of the stated criteria, some may be more salient than others, and that judgements may in fact be based on one or two particular language behaviours rather than on the whole range of features included in the band descriptors. It appears also to be the case that different features may be more or less salient at different levels of proficiency.

As noted, conversational interviews are generally considered appropriate means of assessing not only traditional linguistic criteria (such as accuracy, syntactic and vocabulary breadth, and pronunciation) but also aspects of what is commonly termed *communicative competence.* The influence of less narrowly linguistic factors (such as sensitivity to audience, interactive skill, personal style etc.) in performance-based language assessment has long been acknowledged and discussed by language testers (see for example, Upshur 1979; Jones 1985; McNamara 1990; Wesche 1992), although there is considerable disagreement on what should or shouldn't

be included in second language proficiency tests. Absalom and Brice (1997), for example, consider pragmatic skills such as affecting and responding to an interlocutor, expressing one's self (ideas and emotions), initiating and controlling dialogue, cuing topic shifts and listening actively to be important aspects of the oral proficiency construct. Similarly, Bennett and Slaughter refer to the importance of interactional skills in determining 'conversational proficiency' over and above the 'linguistic skills', for example in ensuring coherence through 'the provision of adequate background information and specific pronoun reference' (1983:19). Others argue that not all aspects of the performance are necessarily relevant to the construct of second language proficiency. 'Interpersonal skills and other affective components', for example, are rejected by Stansfield and Powers (1983) as dimensions of second language communicative competence. De Jong and van Ginkel (1992:187) similarly argue that 'productive skills are observable, but not everything that can be observed in performance data is necessarily skill related'.

A few studies have attempted to identify aspects of the construct of second language speaking proficiency within the context of specific tests. Hadden (1991), for example, found linguistic ability to be but one of five factors contributing to global assessments of oral communicative proficiency, the others being comprehensibility, social acceptability, personality and body language, and argues that there is, therefore, a lack of a direct relationship between linguistic ability and communicative proficiency. Chalhoub-Deville (1995) found that as well as the more linguistic features (grammar and pronunciation) raters focused upon creativity and content (for example, the extent to which the speaker engages the listener) and on detail (for example, the ability to provide information unassisted, the length of the answer and the amount of elaboration). However, while such studies depend upon the analysis of analytic scores, many language proficiency interviews (like IELTS) are based upon a single holistic rating which is not amenable to such analysis.

This shift away from a focus on narrowly linguistic skills towards communicative skill appears to have created an assessment climate where raters, in order to make judgements about learners' communicative skills, need to make *inferences* about candidates on the basis of their communicative behaviours. Pollitt and Murray (1996), in a study of Cambridge Proficiency Examination raters' perceptions using a type of verbal protocol, found that many of the raters' statements consisted of inferences about candidates based on their behaviour. Raters referred, for example, to the candidates' exam-consciousness, apparent lack of intelligence, maturity, willingness or reluctance to converse and sex-related comfort or discomfort. In fact, Pollitt and Murray conclude, raters are 'as concerned with their interpretation of what they observed as with those objective features evident in the performances and equally accessible to all judges'.

Whilst most would agree that inferences are not a suitable basis for judgements, it is nevertheless clear that the assessment of communicative skill is a complex task, made all the more complex because of the general lack of agreement and clarity about what aspects of performance are relevant. Shohamy and Walton (1992) point out, 'The degree of uncertainty about which categories are relevant [to judging the success of the communication] and which kinds of distinctions should be made only increases as we move further away from a purely linguistic description'. We believe that in the IELTS oral interview, which espouses a communicative model (Ingram and Wylie 1996) and which aims to evaluate candidates' ability to cope with the communicative demands of tertiary study, non-linguistic aspects of the performance will inevitably be drawn into the raters' judgements. One aim of this study will be to identify those aspects of the performance and the performer, both linguistic and non-linguistic, which contribute to the raters' perceptions of proficiency.

## 3.1 The Development of the IELTS Oral Interview

Ingram and Wylie (1996) report on the development of the IELTS oral interview in what appears to be the most comprehensive publicly available document pertaining to the interview. The following excerpts provide something of a picture of the construct from the 'task' aspect (the complementary aspect to this being the criteria contained in the scales which will be discussed subsequently):

> ' The three main phases of the interview were sequenced to give candidates the initiative from the start, to encourage them to become active participants in the conversational exchange rather than just provide minimal responses to a series of questions, and to enable them to demonstrate their ability to produce a variety of eliciting functions. ..... Phase 3 [later Phase 2] was designed to give candidates the opportunity to produce extended speech, describing, narrating, explaining or speculating on a familiar topic generally relating to their own experience. Phase 4 was to be a 'dialogue', a classic oral interview situation in which interviewers used brief 'c.v.' forms that had been filled in by candidates before the interview as a basis on which to engage them in discussions (including speculative discussions) about future intentions. This phase was intended to personalise the test, provide something familiar on which candidates could be questioned and could respond at length, and allow scope for more complex, speculative language.' (1996:3-4)

> ' the principal reason for the test was to require candidates to take the initiative, seek information, and speak at length' . (1996:4)

> Phase 4: ' Activities require the candidate to speculate; to express ideas, attitudes, and plans with some precision; to demonstrate the ability to switch register; and to use language relevant to their particular academic, vocational, or other interests.' (1996:11)

As we can deduce from these excerpts, the oral interview was based largely upon a functional view of language. A range of functions are nominated, and the distinction between phases of the interview is based primarily on the different functions to be elicited from the candidate in each phase.

The expectation is also stated that the candidates will demonstrate 'interactional' skills, such as taking the initiative. However, exactly what 'active participant' means is unclear and is perhaps what lies behind the criticisms of oral interviews in general as 'conversations' (cf. van Lier, 1989, etc.). Taking the initiative and being 'active' imply some sort of equality in determining the flow of the conversation, yet this has been argued to be unlikely, to say the least, in an institutional event such as a test where the interviewer is the more powerful participant (eg. Perrett, 1990; Neeson, 1985)

As well as functions and interactional skills, there is also a focus on the complexity of language ('more complex, speculative language'). Additionally, mention is made of precision in expressing ideas and of ability to vary register.

In summary, oral proficiency as interpreted from the description provided in Ingram and Wylie (1996) may be seen as being on a continuum with the following aspects to it :

less evidence of ⟷ more evidence of

functional range
complexity of language
initiative
extended speech
precision of meaning
register variation
field-specific vocabulary

## 3.2 The Test

### 3.2.1 The Interview Structure[1]

The IELTS Speaking Module takes between 10 and 15 minutes. It consists of an oral interview, a conversation between the candidate and a trained interviewer / assessor. There are five sections:

| | |
|---|---|
| *Introduction* | The candidate is encouraged to talk briefly about his/her life, home, work and interests. |
| *Extended Discourse* | The candidate is encouraged to speak at length about some very familiar topic either of general interest or of relevance to their [sic] culture, place of living, or country of origin. This will involve explanation, description or narration. |
| *Elicitation* | The candidate is given a task card with some information on it and is encouraged to take the initiative and ask questions either to elicit information or to solve a problem. Tasks are based on 'information gap' type activities. |
| *Speculation and attitudes* | The candidate is encouraged to talk about their [sic] future plans and proposed course of study. Alternatively the examiner may choose to return to a topic raised earlier. |
| *Conclusion* | The interview is concluded. |

The present study is concerned particularly with the assessment of interview skills and for this reason the Phase 3 role-play was not included in the interview (see Brown and Hill 1998).

[1] This information is taken from the IELTS Handbook (IELTS, 1997).
Editor: The format of the IELTS Speaking test will change from 1 July 2001, see Appendix 6.1.

### 3.2.2 The Band Descriptors

As in any oral test, the task itself is only one half of the story. The other half is the criteria or scales, which are designed to 'exert control on observations both through directing the observer and by providing the language with which to describe an observation' (Griffin and McKay, 1992:17). In this respect the criteria *are* the construct.

The IELTS scales include the following features:

- effectiveness of communication (in relation to a specified range of topic types)
- grammatical range and accuracy
- the ability to talk at length
- functional range.

Other features referred to at specific levels only are circumlocution, accent/pronunciation and fluency. The study will investigate the status of these nominated criteria vis-à-vis other (non-specified) linguistic and non-linguistic criteria in the assessments made by the raters. In particular it seeks to determine what is understood by the term 'effective communicator'.

## 4.0 Methodology

### 4.1 Protocol Analysis

Protocol analysis has long been acknowledged as a suitable technique for investigating the construct validity of tests (see Cronbach 1970, 1971, for example). The application of verbal protocols in language test validation is discussed by Cohen and Hosenfeld (1981), and a range of studies report on their use in investigations of rater perceptions of composition or writing ability (eg. Cumming 1990; Huot 1990; Vaughan 1991; Milanovic, Saville and Shen 1993; Milanovic and Saville 1994; Weigle 1994; Delaruelle 1997).

Of the various types of verbal protocol, *concurrent* verbal reports have been widely used studies involving test data, especially in the investigation of reading skills and the judging of written scripts. This study however, uses a type of *retrospective* verbal protocol known as stimulated verbal recall (Smagorinsky 1994; di Pardo 1994). Stimulated verbal recalls are claimed to have 'a unique capacity to probe the reasons for particular decisions' (Smagorinsky, 1994: xiv). They have been widely used in studies in the fields of psychology, sociology, anthropology and linguistics. The validity of this methodology is premised on the belief that the subject is likely to remember or relive the original behaviour if presented with the same stimulus (Ericsson and Simon, 1984). The advantage of retrospective over concurrent protocols is that they are less intrusive; they allow access to the participants thoughts while avoiding interruption (and hence possible contamination) of the behaviour of interest. This is particularly of concern in the present study where raters cannot be expected to monitor the performance *at the same time as* verbalising their thoughts, so that protocols could only be gathered *concurrently* with constant stopping and starting of the taped interview; verbalisations are likely to interrupt the 'on-line' listening and rating process and seriously distort it.

As the scores themselves are awarded under normal conditions, that is without the interruption of verbalisations, we can assume that the ratings and processes of rating will be consistent with normal rating behaviour as it is undertaken in rater training and re-accreditation, for example. On the other hand, in one respect the ratings do not reflect

operational IELTS ratings. Operational ratings are awarded by the interviewer herself, so in the present study there is likely to be an additional focus on the interviewer which is not present in IELTS interviews which are rated 'live'.

The fact that the protocols are gathered immediately after each rating allows us to assume that raters will still have access to their 'working memory' (Green, 1997:6). We are not, however, claiming the protocol to be an exact replication of the cognitive processes of the interviewer while rating; the task is far to complex for this to be possible. Nevertheless, we can reasonably assume that comments made during these protocols will have some basis in the earlier rating event. Raters were first asked to nominate and justify a score; this justification can only be made by drawing on their earlier thoughts and perceptions. In addition, they were explicitly requested to point out aspects of the performance which contributed to their judgement in the subsequent review of the tape.

The retrospective protocol procedure does, of course, have drawbacks. Obviously time is a consideration: the more delayed the recall, the more likely the subject is to 'reinvent' her/his earlier behaviour rather than remember. Green discusses this in terms of two phenomena: tidying up one's comments, and saying what one thinks the interviewer wants to hear. In this study we anticipated particularly that raters would tidy up their comments in order to appear to be adhering to the criteria (the band descriptors). Steps were taken to ensure that this did not happen by indicating to participants beforehand that it was expected that they would consider features not included in the scales, and that one of the purposes of the study was to find out exactly *what* experienced raters considered relevant. Care was taken to refer to the raters as the experts, and the study was framed as an investigation of the nature of this expertise. In this way the importance of conforming to the scales was down played. In fact, the range of features referred to in raters' comments, and the fact that they at times explicitly acknowledged that they considered factors other than those mentioned in the band descriptors, indicated that this strategy worked.

## 4.2 Procedure

This study is linked to an earlier one investigating interviewer variability (Brown and Hill, 1998), and draws on the same data. The test candidates are overseas students taking part in a pre-university Foundation Program. At the time of the interviews they were preparing to take IELTS prior to submitting applications for tertiary study in Australia. For the present study, a sub-set of eight from the total of forty-two interviews was selected.

Eight accredited IELTS examiners were recruited by letter to take part in the study. They had been IELTS raters for between one and nine years. Each was to rate and provide a verbal protocol for four of the eight tapes, a total of 32 protocols in all.

The raters were scheduled to provide the protocols individually. At the start of each rater's session they were told that they would be asked to listen to four interviews and rate them in the normal way. After each one had been rated they would then be asked to talk about their reasons for awarding the score they gave. As well as these verbal instructions, they were also given them in written form (Appendix 3.1), which they were asked to read through before asking any clarificatory questions. They were also given a copy of the IELTS band descriptors to read through before listening to the first tape. The room was set up with two tape recorders, one to play back the IELTS interview tape for rating, and one to record the subsequent retrospective verbal protocol.

During each protocol session, that is after nominating the score awarded and providing a brief justification, the rater was invited to replay the tape from the beginning, stopping wherever she/he felt some comment was in order. The researcher was present during these events, providing an audience for the comments, but minimal intervention. Most of the researcher's participation consisted of minimal feedback and encouragement to continue. At times, however, intervention was necessary, for example where a comment was unclear, or where the interviewer appeared to react strongly to something in the interview but did not stop the tape[2].

A short break was offered between each protocol session and the rating of the next interview. Each rater's full session lasted for between three and four hours.

# 5.0 The Data

## 5.1 Scores

Table 1 shows the ratings awarded to the performances. As can be seen, and is perhaps to be expected given the nature of the assessment (a single rating using an holistic scale), there was a considerable level of disagreement amongst raters. Variation in scores awarded to individual candidates ranged from 2 band levels (Tapes 48, 57 and 66) to three band levels (all other tapes).

| | Raters | | | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| **Tape** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **Range** | **Score** |
| **8** | | 8 | | 6 | | 8 | | 8 | 6-8 | 7 |
| **32** | 6 | | 6 | | 8 | | 6 | | 6-8 | 6.5 |
| **40** | 6 | 4 | | | | 5 | 5 | | 4-6 | 5 |
| **44** | | | 5 | 6 | | | 7 | 6 | 5-7 | 6 |
| **48** | | | 4 | 5 | | | 5 | 4 | 4-5 | 4.5 |
| **50** | 6 | 5 | | | 5 | 6 | | | 4-6 | 5 |
| **57** | | 5 | | 5 | | 4 | | 5 | 4-5 | 4.75 |
| **66** | 5 | | 6 | | 6 | | 6 | | 5-6 | 5.75 |

*Table 1: Ratings*

## 5.2 Protocols

All but one of the protocols (Interviewer 4, Interview 8) were recorded successfully. The data set consists therefore of 31 protocols. The shaded cell in Table 1 indicates the missing protocol.

At the start of each protocol session the rater started by nominating a score for the candidate and briefly justifying it. Further comments were invited once raters had completed the stimulated recall. These comments, which served to sum up the reasons for the particular score awarded are referred to henceforth as summary comments. All other comments, ie. those which took place *during* the stimulated recall, the review of the interview, are referred to as review turns.

[2] As we shall see, the result of this is that some comments which appeared obvious at the time were less meaningful later. The question of how much intervention and clarification should be allowed in verbal protocols is a difficult and unresolvable issue.

Contributions varied enormously, with the longest protocol, 2207 words being produced by Rater 3 in response to Interview 44, and the shortest, 326 words, being produced by Rater 4 in response to Interview 48. Rater 3 produced on average the longest protocols (1542 words) which was more than twice the average amount produced by Raters 4 (718 words) and 2 (756 words).

The number of review turns (that is, the number of times the rater stops the tape to comment) also varies enormously, ranging from 7 for Rater 2 (Interview 50) to 28 for Rater 3 (Interview 32). In fact, Rater 2 produced on average the shortest reviews (363 words) and Rater 3 the longest (1281 words). Averages for each rater are presented in Table 2. The number of summary words (the justification of score) varied from a low of 72 (Rater 4, Interview 48) to a high of 753 (Rater 8, also Interview 48). Rater 4 produced on average the shortest summaries (255 words) and Rater 8 the longest (558 words).

| Rater | Av. protocol length (words) | Av. summary length (words) | Av. Review length (words) | Av. number of review turns | Av. review turn length (words) |
|---|---|---|---|---|---|
| 1 | 1374 | 482 | 892 | 18 | 50 |
| 2 | 756 | 393 | 363 | 9 | 40 |
| 3 | 1542 | 269 | 1281 | 28 | 46 |
| 4* | 718 | 255 | 462 | 11 | 42 |
| 5 | 1425 | 539 | 886 | 18 | 49 |
| 6 | 1100 | 368 | 732 | 13 | 56 |
| 7 | 1066 | 403 | 663 | 14 | 47 |
| 8 | 1239 | 558 | 1047 | 17 | 62 |

*Table 2: Averages for each rater* * 3 interviews only

In summary, Rater 3 has the most to say during the reviews. Although she doesn't have the most to say in the summaries (in fact she produces the second shortest on average), she compensates for this with frequent (average 28) stops for comments during the review. Raters 2 and 4, in contrast, between them produce the two shortest protocols, with Rater 4 producing the shortest summary turns, and Rater 2 the shortest reviews. In addition, Raters 2 and 4 produce the least number of turns per interview on average, and the shortest review turns (as measured by average number of words produced).

## 6.0 The Analysis

Transcripts of the protocol session were reviewed carefully in order to get a feeling for both possible units of analysis and possible coding categories, although there was clearly an expectation that these would reflect, at least to some extent, the contents of the bandscales. The unit of analysis decided upon was 'a single or several utterances with a single aspect of the event as the focus' (Green, 1997). Additional items which elaborated on the central comment in some way, for example providing justification, amelioration, evaluation and exemplification) were not treated as separate units for the purposes of this analysis. Because of the complexity of the comments and the overlap between categories (the result to some extent of a lack of clarity or ambiguity in raters' comments, but also attributable to a difficulty in separating aspects of performance conceptually, such as the organisation and content of candidates' contributions), the process of coding was an iterative process, requiring constant revision until most comments were classifiable in a way which appeared intuitively adequate and was also relatively straightforward to do.

In general, three types of comment occurred - *evaluative*, which focused on some aspect of the candidates' language; *non-evaluative*, which referred often to affective aspects of the interview such as the relationship between the two participants, and *interviewer-focused*, consisting of comments on the interviewer or their behaviour. We consider first the evaluative comments.

A total of 413 evaluative comments were made. Evaluative comments include both explicit and implicit evaluations. Examples[3] of explicit evaluations include:

40-6 So she explains all that quite clearly

66-7 Yeah, *I think my cousin or my sister is in fifth year.* That's alright, that's okay.

44-8 Now that's not a bad answer: *a lot of development and a good place to study*

32-5 A bit inappropriate, *animal bashing*

Implicit evaluations were less frequent, but occurred particularly in relation to sentence level syntax and vocabulary, and tended to include quotations of errors (negative evaluations) or sophisticated language (positive evaluations). The evaluative nature of the comment was often to be inferred from the way the rater uttered the comment, or from the context in which it appeared:

57-8 *One of my uncles are engineering*

32-5 See that? Another aside *It's got lots of shops, quite expensive*

8-8 *They're pretty old*

32-1 *Not really hotel*

Of the evaluative comments, 151 (37%) were positive and 262 (63%) were negative, in other words the majority of comments were negative. Starting initially with aspects of linguistic skill included in the band descriptors and following an iterative procedure, comments were ultimately grouped according to the following categories: (sentence level) syntax, discourse, vocabulary, production, comprehensibility, use of strategies, and comprehension. Each of these will be discussed in more detail in following sections.

## 6.1 Validity of the Protocols

A check was made upon the validity of the retrospective protocol data as a representation of raters' actual assessment processes. We hypothesised that the proportion of positive comments would increase as the score increased. The ranking of the eight interviews according to their mean score was compared with their rankings based on the proportion of 'positive' to 'negative' comments.

[3] In all extracts the first number refers to the tape or interview number, the second to the rater. Thus 40-6 refers to Tape 40 Rater 6. Within the extracts, direct quotes from candidate speech are in italics.

| Interview | Total no. of evaluative comments | Negative comments No. | % | Positive comments No. | % | Ranking | Mean Score | Mean Score ranking |
|---|---|---|---|---|---|---|---|---|
| 8 | 42 | 27 | *64* | 15 | *36* | 4 | 7 | 1 |
| 32 | 83 | 41 | *49* | 42 | *51* | 1 | 6.5 | 2 |
| 40 | 57 | 38 | *67* | 19 | *33* | 6 | 5 | 5 |
| 44 | 53 | 33 | *62* | 20 | *38* | 3 | 6 | 3 |
| 48 | 41 | 32 | *78* | 9 | *22* | 8 | 4.5 | 8 |
| 50 | 43 | 26 | *60* | 17 | *40* | 2 | 5 | 5 |
| 57 | 44 | 32 | *73* | 12 | *27* | 7 | 4.75 | 7 |
| 66 | 50 | 33 | *66* | 17 | *34* | 5 | 5.75 | 4 |

*Table 3: Protocol validation: rankings*

The distribution of positive and negative comments on the whole reflected the rankings based on scores. We can reasonably conclude therefore that the comments are adequately representative of raters' views. There was only one interview where the score ranking appeared to be out of line with the polarity of the comments, and that was for Interview 8 (mean score ranking = 1, ranking according to polarity of comments = 4). For this interview, the scores awarded were 8, 8 and 6. (The fourth score, awarded by Rater 4, is not considered here as the protocol recording was faulty and could not be included in the analysis), and the candidate was hence ranked the highest by score, but sixth based on polarity of comments. We reviewed the comments themselves in order to seek a reason for this discrepancy.

We found that the two raters who gave the highest scores actually presented more negative comments than positive (Table 3). While the main reason Rater 2 gives in her summary statement for awarding an 8 was a certain 'nativeness', particularly in the use of markers such as *like* and *hopefully*, this is not in fact mentioned in the band descriptors. Perhaps this is why she avoided further mention of this in the review section, choosing instead to comment overwhelmingly on the candidate's syntax and fluency, both being categories which are explicitly mentioned in the descriptors. She comments positively 5 times, all on syntax, and negatively 6 times, all on fluency. However, each time she comments on fluency she provides non-linguistic justification for the candidate's disfluency - embarrassment ('She doesn't know how to put it delicately'), thinking of ideas ('I think it's difficult to speak fluently and readily about the same topic for ... yeah, you're running out of ideas'), or personal style of speaking ('It's a personal trait probably'; ' That's her style of speaking'). In short, although the comments were negative, they did not lead to a negative evaluation of candidate ability.

Rater 6 also awarded an 8, yet provided 8 negative comments and only 2 positive ones. Again, this raters judgement was not, as she acknowledged, something that was clearly based on the scales, but was instead to do with the extent to which she would have to modify her speech: 'If I were talking to her, I wouldn't adjust my language .... it doesn't say anything like that in the bands, but that's a sort of a gut feeling you have when you first listen to someone'. In fact the negative comments, the weaknesses that she points out, are probably examples of the 'few inappropriacies' she refers to in her summary. And again the candidate's hesitancy is perceived as non-linguistic: 'the sort of hesitancy that native speakers have just speaking appropriate words and searching through the brain'.

## 6.2 The Comments by Category

The largest group of comments (31% of all evaluative comments) relates to sentence level syntax, and just over half of these (55%) are negative (Table 4). The heavy focus on grammar reflects the findings of a number of other studies. McNamara (1990), for example, in an analysis of the relationship between an 'overall' score and specific linguistic analytic criteria in a speaking test for medical professionals, found that grammar contributed more than any other category to the overall assessments. This may well be because grammar is quantifiable and systematically taught, so that for a language expert, as Wall, Clapham and Alderson (1994:334) point out, 'grammar is less difficult to judge than the language skills'. Comments on the discourse (including content) account for 22% and are the second largest category, and again over half (60%) are negative. Production is the third largest category, with 18% of all comments, of which an overwhelming majority (81%) are negative.

| | Syntax | | Discourse | | Production | | Compre-hensibility | | Vocabulary | | Strategies | | Compre-hension | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | 130 | | 89 | | 75 | | 39 | | 36 | | 33 | | 11 | |
| *%* | *31* | | *22* | | *18* | | *9* | | *9* | | *8* | | *3* | |
| Polarity | + | - | + | - | + | - | + | - | + | - | + | - | + | - |
| Total | 58 | 72 | 36 | 53 | 14 | 61 | 2 | 37 | 12 | 24 | 27 | 6 | 2 | 9 |
| *%* | *45* | *55* | *40* | *60* | *19* | *81* | *5* | *95* | *33* | *67* | *82* | *18* | *18* | *82* |

*Table 4: Positive and negative comments by category*

The next three groups each account for just less than 10% of all comments - comprehensibility, strategies and vocabulary. While the overwhelming majority of comprehensibility-related comments are negative (95%), the comments on strategies are overwhelmingly positive (82%). Comments on vocabulary are also mainly negative (67%). The candidate's comprehension accounts for only 3% of comments, and most of these are negative.

The fact that comments in the production, comprehensibility, comprehension and strategies categories are mainly negative deserves comment. This will be done as each category is discussed in turn.

### 6.2.1 Syntax

Positive reference was made to syntactic accuracy and maturity, and negative reference to syntactic error, immaturity and limited range. Whereas some comments made reference to the candidate's overall syntactic ability, others made reference to occurrences of specific aspects of syntax. Thus positive general comments tended to refer to 'structural competence', to infrequency or lack of impact of errors on comprehensibility, or to sophistication, naturalness or maturity of expression. Negative but general comments tended to refer to lack of structural control, the occurrence of errors or the narrowness of the range of structures used. Positive specific comments referred typically to the occurrence of structures which were presumably considered evidence of a developing syntactic maturity (discussed below) and negative specific comments typically referred to syntactic errors (also discussed below). Examples of comments in the syntax category include:

40-1 But still, I mean, she's keeping utterances going without making any terrible mistakes, without mistakes which really do interfere with communication. They're pretty thin on the ground. (*general positive*)

66-3 He doesn't display any degree of flexibility in, or creativity in his sentence structures. They're very sort of simple, really. (*general negative*)

50-2 *More easily than other subjects*, I thought was quite good. (*specific positive*)

32-1 Again *youth hostel* there wasn't an article. (*specific negative*)

In order to increase our understanding of what *specific* aspects of grammar raters consider relevant, comments were coded according to the aspect of grammar referred to. Positive comments referred to conditionals and verb tense (7 comments each), adverbs (6 comments), relative clauses (4), modals (3), and comparative and use of connectives (2 each). Negative comments, references to grammatical errors, overwhelmingly concerned verbs, and in particular tense (18 comments). Other negative comments concerned the comparative (5 comments), connectives, articles and prepositions (3 each), word order, adverbs and pronouns (2 each), and adjective order, conditionals, relative clauses, reported speech and the subjunctive (1 each).

While the use of connectives is explicitly referred to in the band descriptors, and the use of the conditional is implied (as a task feature in relation to the function of speculation), the other grammatical categories commented upon here derive presumably from teachers' experience of and expectations regarding the acquisition of English grammar. While the number of comments is admittedly low for many categories, it seems reasonable to assume that the fact they are commented on indicates that the occurrence and accurate use of these specific aspects of grammar is considered to be an indicator of syntactic maturity. It is interesting to note that tense appears to be a most salient indicator, being commented on at some point by all raters.

### 6.2.2 Discourse

Comments in this category included reference both to the *discourse structure and organisation* and to the *content* of the candidate's speech. Positive comments made reference to the adequacy of the sample of speech in relation to specific functions (narrating, describing, speculating, hypothesising, and so on), to the ability to produce extended discourse, or to the sophistication or maturity of the ideas or their organisation. While some of the comments were readily identifiable as one or the other of these three categories - functional skill, discourse complexity, and maturity of ideas - many of the raters' comments did not appear to distinguish content and means of expression, perhaps because the two notions, sophistication of content and discoursal sophistication, tend to go hand-in hand, or perhaps because it is not always possible to disambiguate content and organisation[4].

Because of the difficulty in assigning all comments to one or another of these discourse-related categories, this analysis does not distinguish between them. A further analysis, or perhaps a different study where raters are asked to elaborate and expand on their comments, may be able to tease out more subtle distinctions than has been possible here given the scope

[4] On the other hand it may be that the raters themselves were clear on what they were commenting upon, but were simply not explicit enough. As was mentioned earlier, a decision was made not to ask for clarification or elaboration as it was felt that this intervention may influence the direction of subsequent talk by the interviewer. Such are the difficulties of this type of research!

of the current research project. The examples below reflect the range of comments in the discourse category:

Discourse: Positive comments

66-1 I suppose again he's managing to get out quite a complex discussion there about the advantages and disadvantages of pharmacy. It's taking him a long time to get it out, but it's reasonably sophisticated

57-6 He's not too bad there. He explains it ... he's linking his ideas and explaining why he chose it, so you know, it's not too bad.

50-1 Okay, well there she challenges the interviewer, which I thought was sort of fairly critical because I think that is part of communicating effectively, that you are able to challenge.

32-5 You can see that she's trying to say something more than the obvious. You know, she's trying to think of something: *well everyone knows it's big, and everyone knows it's got lots of cars ...What else can I say that's, you know, interesting, that people don't already know*, you know, she's sort of excusing herself for saying something so ordinary when she says *of course.*

50-1 Again, that's sort of, it's reasonable reasoning. Okay, maybe business is culturally bound, but accounting? It seems a reasonable suggestion that accounting is-

Discourse: Negative comments

8-8 See, she could have expanded there. She could have said: Yeah my grandfather, my father, even though, you know, my grandfather is older and would normally have more respect because my father was an important businessman or something'. She didn't. She had a chance to say more there, but she didn't.

32-3 See the descriptions fall down

32-3 Okay so when he's here and she says *Tell me something about it,* he starts saying *it's very beautiful and it's a beautiful island.* Then he says *there's a lot of Chinese,* and then he says *the food's good,* and it just seems to me that these are the words he knows. It's not a very sophisticated way of describing where you live. You could see he was - perhaps if he started to talk about the Chinese if he could get involved in the politics of what it feels like to be living there, and he never gets there so you think okay, he knows those words, that's why he's using them. So I really felt he was limited in what he could explain.

40-2 She can manage a conversation, but still the nature of what she's saying is not going much- not advancing.

44-8 So she says it reasonably well, but there's no sort of opening general comment following, oh you know: *Oh when I go to Singapore, oh there's*, you know, *lots to do*, or, you know, *I do a whole range of things*, you know, nothing like that. It's just, you know, *I visit my auntie and I-* It's like kids.

66-1 He gave very minimalist answers. He was very unforthcoming.

While functional skill and extended discourse are explicitly referred to in the band descriptors and/or in the test and task specifications, the quality of ideas is not explicitly referred to. It is interesting that some raters, particularly Raters 4 and 8 made relatively frequent reference to the maturity of ideas expressed by the candidate (or lack thereof), particularly in the more cognitively demanding functions of hypothesising and speculating. It seems that for such raters content is indeed an aspect of the construct. This may well be because the purpose of

the test, to screen tertiary applicants, leads some raters at least to consider intellectual maturity as well as linguistic maturity, the two being relevant to success at university.

The complexity of the relationship between length of output, complexity of ideas and complexity/precision of expression is further compounded when we note that while in some instances immaturity of ideas was attributed to a lack of linguistic resources, in other instances the apparent youth or immaturity of the candidate was used to justify the lack of extended or complex response. In fact, inferences about the candidates' personality, maturity, world knowledge, and so on were frequently in comments falling into this category. This is perhaps inevitable given the references in the band descriptors to 'effectiveness of communication' and 'precision of meaning', terms which are abstract, which do not themselves make the distinction between language and content clear. A lack of extended discourse was attributed variously:

to the interviewer's style:

44-3 I think she jumps in pretty quickly

to the interviewer's failure to elicit extended discourse:

44-3 And again those yes/no questions ... rather than general questions 'Tell me about your life'

50-5 You know, they're not questions that are making her actually- you know, you just need something like 'Tell me', or 'Go on and tell me a bit more about that' or- Yeah, I think he's trying to make her talk by switching topic, but actually, that's making it worse because every time there's a new topic, it's only a short answer again instead of maybe digging deeper and saying 'Well tell me about that then' or 'Tell me in detail' .... Yeah, because until a topic is established, you don't really know what's relevant and what isn't, but he keeps changing topics and there's no time to work out what we're going to talk about

to the candidate's personality or youth:

44-4 It could be, again, a young person who doesn't really like to talk too much

44-3 There's never any attempt to fill in the details the whole way through. So, again, it's a bit of immaturity too a little bit I think, but she's just answering the questions rather than filling in any of the details or describing or explaining or, you know-

to affective factors:

8-7 She had a chance to say more there, but she didn't. But I think it was a confidence thing by the end of the interview

to test wiseness:

50-6 She just didn't add information, and I don't know whether that was a cultural thing because she, maybe she was shy, but I don't think it was that. Maybe lack of preparation.

and to the choice of topic:

44-4 The topic doesn't really extend them either.

Even within the same performance, raters do not agree in their interpretations of particular behaviours. We draw on the summary comments provided for interview 44 to illustrate this. Rater 3 who gave the lowest score (5) felt that the candidate was *not able* to produce extended

discourse or speculation, whereas of the two raters who awarded a score of 6, Rater 6 justified the limited discourse with reference to the skill of the *interviewer* and Rater 8 with reference to the *candidate's youth* or immaturity. Rater 7, who awarded a score of 7, attributed the lack of extended discourse to *both* the skill of the interviewer *and* the youth of the candidate.

Raters frequently made reference to the functional skills displayed by the candidates, a feature of the band descriptors. However, again, inferences were made regarding non-satisfaction of the functional demands, and reasons given were both linguistic and non-linguistic (see Section 6.2 2 Discourse). Rater 1, for example, interpreted the candidate's failure to speculate in Interview 32 as a result of linguistic weaknesses; he justifies this interpretation on the grounds that the candidate has (to him) clearly thought about the issue:

32-1 Judging from what she said about her commitment against animal testing, I felt that it probably wasn't a lack of having actually thought about the issue, which was causing a problem here ... It was some difficulty in presenting complex ideas and language that was causing it to break down.

In other instances, non-linguistic reasons were inferred for lack of speculative language. These drew on maturity:

40-1 taking into account the fact that X was 17 years of age, and that therefore, sort of cognitively, just in terms of real world knowledge her ability to speculate would be a little bit limited

the difficulty of the questions:

32-8 Okay, she's having some difficulty speculating about how she can help her country ... I think it's quite a difficult question if you haven't thought about it before

lack of speculative questions:

44-7 I really feel like the interviewer doesn't challenge enough in terms of the speculative, argumentative

66-3 there wasn't a lot of speculative language elicited

and lack of comparison of real world and test context:

32-9 I always think about it in context of the university situation, and I think, usually you do have time to prepare for those sorts of responses, and you're dealing with the issue all the time, so you're becoming very familiar with the vocabulary, it's all at the tip of your tongue. Just off the cuff like that it's hard to think about those things.

It was also clear that raters perceive the status of functions, particularly speculation, as an assessment focus to be somewhat problematic. This appears to be because it is not entirely clear whether they are to focus on the *linguistic* exponents of the function (eg. the conditional, the subjunctive) or the ability to respond to such questions with *appropriate content.* The reasons for this is undoubtedly because of the common association of certain grammatical features with particular functional uses of language in the teaching of language. The use of the conditional, for example, is typical of speculation and hypothesising, and at times raters commented specifically on the candidate's use of the conditional:

40-2 I think I heard one conditional, and that was a first conditional.

Comments such as these have been coded in the syntax category (see Section 6.2.1 Syntax). Other comments refer more generally to the candidates' failure to respond to a speculative

question, seeing it as a consequence of immaturity, rather than linguistic limitations, in one case, :

40-2 It's all description even though she's asked her what- about the future and her plans, and even-. She is actually in theory talking about the future, but it's still description.

32-2 the younger ones tend not to be able to deal with that sort of speculation very easily. So I think the more mature candidates do have an advantage in that way.

It appears then that different raters appear to look for different evidence of ability to hypothesise or speculate; some for specific linguistic structures, others for fulfilment of the functional task (ie. answering the interviewers question adequately in terms of meaning) *regardless* of the linguistic forms used. The tension between these two perspectives is evident to at least some of the raters themselves, as the following excerpt shows:

66-3 He's tried to ask him some speculative questions, and rather than using conditionals or hypothetical language, he actually just takes off 'Because all my sisters are over here', and that's quite natural way of speaking really. So I don't know whether he can use it or not, but it's quite natural. 'Would you have studied over here anyway?' 'Yeah, because my brothers have'. I wouldn't say 'Yes I would have studied even if I hadn't done' - you know, that sounds unnatural in the conversation that's going on. So, even though there wasn't a lot of speculative language elicited, I think that he dealt with it. Whether he could, his response was appropriate, even if we're not sure whether he can use speculation.

Whilst this ambiguity is, as noted, of concern in the classification of comments, more importantly it highlights an underlying lack of consistency in the ways raters focus on functional skills, which in itself reflects the lack of explicitness in the assessment guidelines.

### 6.2.3 Production

Comments in this category referred to fluency, rhythm and intonation, and pronunciation.

Fluency

Fluency was in some (5) instances referred to in a non-specific way, for example:

44-8 Yeah, so she's sort of quite fluent with sort of answering the questions

44-7 See this is tending to lose fluency here.

40-1 There you see, on the one hand it's definitely not fluent,

Other references (a total of 56) were more specific and concerned features such as hesitation or speed of delivery, the use of fillers, and repetition:

Hesitancy and speed of delivery (30 comments)

Most comments in this category were negative (27) and every interview received at least one or two:

40-1 And also there was a certain hesitancy always there.

32-3 It's a little bit slow I guess for an interview process.

8-2 The only thing that is irritating is that it takes her so long to actually spit it out. She's taking a lot of thinking time.

The three positive comments were all made by the same rater, Rater 7, two of them in relation to the same interview:

44-7 And she's just, she's like a native speaker in her retorts so quickly and with her amusement. I mean she, there's no hesitation. She's very quick.

44-7 See, when she's asked *What are the main things you've learned?* she says *Independence*. She says it very quickly. I mean it's just a- she knows the response straightaway, and she can articulate it.

66-7 Okay now see, he responds to that quickly and capably. Now is that because he's been asked that question five thousand times before, and he's got the answer down pat, or is it something that he ... not like that and he just formulates it quite capably? I don't know what it is.

It is interesting to note that the raters often made inferences about the reasons for hesitation. Lack of fluency was at times seen as a linguistic feature, that is, the candidate was searching for words or structures:

32-3 I think her biggest limitation is a lack of vocabulary and she tries to cover that a lot by using phrases like *something like* and *you know*

48-2 There's so much hesitation as she's trying to find words or the form of the words as with, *Malay, Malay*

It was also frequently attributed to non-linguistic features such as

personality:

8-2 I think that's a personality thing rather than a linguistic thing

8-6 I know a lot of slow native speakers

affective aspects of the encounter:

8-2 Maybe she's embarrassed here

57-4 And so much hesitation, you know ... but that's stage fright

interest in the topic:

8-2 In terms of content, I think that it's difficult to speak fluently and readily about the same topic for ... yeah, you're running out of ideas

8-2 She's speeding up a little bit now because she's got something different to say.

a result of (native-like) cognitive planning:

8-6 The sort of hesitancy that native speakers have just speaking appropriate words and searching through the brain

32-7 See, all this hesitation I feel is because she's thinking, not because she's trying to think of the word.

8-6 So there's a lot of pausing here, but I think this is really hard for her to explain.

It seems therefore that, as predicted, raters routinely infer the reason for particular behaviours and, moreover, that they realise that they have to make inferences. At one point one rater says:

66-7 ... but why is he hesitating? I don't know why he's hesitating so much. ... and I'm asking myself why is he hesitating so much

and later on she comments:

66-7 I don't know whether he's buying time- he repeats, you know, *How long have I*, that technique in conversation when you repeat the question like you want to buy time to formulate your answer. Now are you doing that because you're trying to do it to think of an opinion or think of answer or because you're just thinking I'm trying to process these words that you're giving to me? See the two different things that might be happening and I'm trying to think about which is happening because obviously it affects how you score it.

A major problem with performance tests is the fact that while evidence of a particular behaviour can clearly be taken as an indication of mastery, lack of evidence cannot always be assumed to indicate non-mastery. So in the case of fluency the question arises 'Is the lack of fluency evidence of linguistic shortcomings (ie. a search for words) or is simply evidence of cognitive planning, a consequence of the type of task or question?' Whatever the case, it is likely that the inference drawn by the rater as to the cause of hesitation will affect the way the perception of fluency is integrated into the final judgement. The same issue of how to interpret non-production of particular grammatical features (in this case the lack of a conditional) applies also to the comment by Rater 3 in relation to the use of 'would' (in the section on discourse above). Non-use of 'would' in the context of a hypothesising statement may indicate that the candidate is *unable* to produce conditionals, it may however simply be that she has *chosen* not to use this particular form (as, in fact, Rater 3 assumes).

Fillers (12 comments)

Comments on the use of fillers were made in relation to four of the interviews. These are particularly interesting, as raters appear to draw different conclusions about candidate proficiency when fillers are a feature of their speech. Some cases were considered to be native-like, indicative of a certain ease with the language, and hence a positive feature. In these instances the assumption is that fillers are used as native-speakers use them, that is, while the speaker is thinking *what* to say. In other cases, however, the use of fillers is interpreted as evidence of limited vocabulary. In these cases the assumption is that they indicate that the candidate is thinking *how* to say it, ie. searching for words. The following two comments are illustrative of these two viewpoints (and both were made in relation to the same interview):

32-3 Okay, there when she's filling in she continues to do this all during the tape. I think her biggest limitation is a lack of vocabulary and she tries to cover that a lot by using phrases like *something like* and *you know* and she's got lots of fillers like that so she's fluent enough to be able to use those but I think really does hide a limited vocabulary and not being able to extend herself,

32-7 *What I mean is like...* The fact she says *like*, shows a degree of sophistication. You know, she's heard Australians talking, or she's picked that up - *like I can da da da.*

Stress, rhythm and intonation (10 comments)

Five positive comments were made regarding intonation, and five negative. It is interesting that all the positive comments were made in relation to one interview, Interview 32, and were made by all four raters. These positive assessments of intonation were typically associated with nativeness:

32-5 Now it's really- that's a classic (rising intonation) and you do this or you do that, that's really Australian intonation. You get carried into her conversation.

32-7 See, there where she's talking about animal bashing, the intonation rises just the way that a native speaker does when we're trying to, you know- ... You know, she's picked up those little nuances of native speaker language..

32-3 ... her intonation is also very good, so and that's just the rhythm of her speech ... I like the way she used that *of course*. ... Quite naturalistic sort of...

Negative assessments were made in relation to Interviews 44 (2 comments) and 40 (3 comments) and were typically associated with interference from the L1:

40-5 I reckon that intonation is just annoying me. It's just *na na na na* (undulating). It's Hong Kong.

44-8 it's a combination of the stress and the way she occasionally leaves a word out that gives her this machine gun effect rather than a nice smooth speaking style.

Repetition

Only four comments were found in relation to repetition, all of them negative. Three were made by the one interviewer in relation to a single interview (Interview 48, Rater 7). Repetition appears to be interpreted as a failed self-repair strategy:

48-7 See, she has to try about three ways to say something, you know, *Some people, some students, some....* And then there's a lot of rephrasing until she finally- and even then she doesn't necessarily get it right. Whereas I'd take it as some form of mastery that they could self-correct quickly and get it right. But she's still at that stage where she's exploring three or four options and still not necessarily coming out with the right one.

Pronunciation. (14 comments)

Comments in this category referred both to general traits:

57-4 And his pronunciation is a little bit difficult.

57-5 you're sort of initially thrown off guard because his pronunciation's bad

and to specific instances where the pronunciation was noticed as being problematic:

48-7 And- first of all I thought she said, *no clothes*, and couldn't work out what she was saying

40-1 Yes, the play/pray problem.

The fact that there was only one instance of a positive comment

48-3 but the pronunciation's really good

provides evidence for the claim that pronunciation is likely to be salient to the rater only when it causes problems. All but one of the interviews received one or two comments regarding the quality of pronunciation. Interview 57 received the most, five (negative) comments made by three of the raters.

### 6.2.4 Vocabulary

Comments falling into this category were of three broad types:

drawing attention to specific words:

8-8 She seemed to have trouble describing the company, like she didn't know the vocabulary for the- she couldn't say it was a stationery or whatever it was ... Yeah, she had to say what it was. She couldn't generalise about it, but then she countered that with something that was quite- what did she say after that? *She told me that she was going to quit.* Now even that is pretty natural sort of English. Usually they don't know that, you know, the bad ones don't know the word *quit.*

32-1 *Small house*, so again, and this was the point where I decided well this is definitely not a 8.

32-5 A bit inappropriate, *animal bashing.*

drawing attention to lexical sets in relation to a particular topic:

57-8 Okay, so that's not very impressive, you know, Describe it ... *beautiful island, surrounded.* That *surrounded* was good, but he hasn't got very many adjectives, so his vocabulary is not too fantastic.

66-5 He's got words like *increase, the chance, applying, job,*

general comments on vocabulary range or usage:

40-6 So she's got limited vocabulary.

44-3 She misuses the vocab too occasionally.

Only in one instance is there general comment on the same vocabulary item – 'seldom'. Otherwise raters appear to be idiosyncratic in choosing lexical items to comment on.

44-8 *Seldom*'s a rather good word ((laughs))

44-4 Yeah, I love this little use of *seldom* that she has there.

44-7 *I seldom go*, you know, that's sophisticated. I mean who says *seldom*? I mean she's learned English well, I feel.

Negative comments on vocabulary (24) outweighed positive ones (12), and three interviews in particular received a rather large proportion of negative ones, Interview 40, Interview 32 and Interview 50.

### 6.2.5 Comprehensibility

The comments included in this category relate primarily to the effect on the listener, unlike those classified in the production category which do not explicitly refer to the raters' understanding[5]. They are, however, as for pronunciation, almost entirely negative; out of a total of 40 comments only two were positive assessments. While it is in some cases possible to infer where the cause of the comprehensibility problem lies, it is more often not clear. Mishearing may be due to an attention lapse of the rater as much as to the candidate's production:

[5] Obviously such comments are related; their classification into different categories is unfortunately a consequence of the necessity of imposing categorical distinctions on such speech data.

### 6.2.7 Comprehension

Nine negative and two positive comments were made regarding the candidate's comprehension of the interviewer. Negative assessments of comprehension refer to simple mishearing as well as language-related inability to understand (such as not knowing specific vocabulary), and miscomprehension of the intent of the interviewer's question (misinterpretation). It is clear also from the comments below that raters appear not to put too much weight on the importance of individual instances of miscomprehension:

50-6 That's right. So, okay, so she misunderstood what he said, but that's not an issue, I don't think. He came back to it, and she answered it.

48-7 See, she lacks the- she doesn't understand that question, which is what's important. I mean she's relating it to herself. I don't think she understood that at all.

44-7 I think she misunderstands here. She says *What's your first impression?* and she said *Oh, I came here before*. She thinks she means the first trip. So that comprehension thing comes into it, but I think it's a minor sort of error in communication

32-7 See, she's just lost- hasn't understood that at all, has been thrown by the concept of stages and got a bit lost I think in what the interviewer was saying because it's quite a long explanation of what she meant, and hasn't answered appropriately, but I don't think we can give her the chop for that.

Comprehension appears to have an ambiguous status as an assessment focus. One rater (in fact the rater who comments the most on comprehension) comments explicitly on this:

32-7 I always have a problem with the issue of comprehension because when we did our training I got the distinct impression that we shouldn't be testing the comprehension, that comprehension was looked after in the listening section, but I don't see how you can possibly cut your mind off from comprehension even though it doesn't say anything here about comprehension. And I have a problem with that. Well if you ignore that they haven't comprehended what you've said, then it isn't real communication, and this is supposed to be a test of some sort of communicative ability, so yeah- ...

We turn now to the two categories of non-evaluative comments, the interviewer and affective factors.

### 6.2.8 The Interviewer

A considerable number of comments (95) were devoted to the interviewer. These included reference to the difficulty of questions, the lack of speculative questions (Phase 4), the number of closed questions, the interviewer 'talking down' to the candidate, the labouring of particular topics, the inappropriateness of certain topics ('delicate', 'boring') the interviewer's interrupting the candidate, the time allowed for the candidate to respond, and the interviewer's failure to pick up on points made by the candidate. Examples include:

44-7 I don't think it really reflects an extended conversation, and I really feel like the interviewer doesn't challenge enough in terms of the speculative, argumentative.

8-2 It's getting a bit laboured now. I'm ready to move on.

8-2 I think that there has been a missed opportunity here to get her to talk about why she wants to do the multicultural course and where that might lead her then in the future. I'm not saying that the interviewer's done the wrong thing. It's just I would like to

have heard her expand on that because she might have been a bit more enthusiastic, but *now we're going to talk about Melbourne*

32-6 That is the first of a series of interruptions which I found really off-putting at the beginning. She does it three or four times, and she interrupts so badly that she even apologises at one stage. She goes *Oh sorry.*

32-7 See, she's just lost- hasn't understood that at all, has been thrown by the concept of *stages* and got a bit lost I think in what the interviewer was saying because it's quite a long explanation of what she meant,

40-6 I wish the interviewer had said more 'Tell me about the subjects you studied'. She's giving her the opportunity give single answers all the time

40-6 So disapproving. Don't you think it's a really sort of, not- ... Yeah, she can't work out what to do. So then she does something and then the interviewer doesn't like it. So then she says *BUT, you're seventeen.* What do you do? ...

44-4 but the examiner really gives me the shits, that she's very condescending

50-1 That's a rather delicate question, and he then goes on with it.

50-5 Students hate talking about architecture usually. Maybe it's because they don't have the vocab. That could be a reason as well. And often, people don't walk around looking at their architecture, like people living in modern cities and people who are coming to Australia, they're more interested in, you know, high-rise and shopping complexes,

66-5 She leaves plenty of time for him to talk. She's just silent (laughs) letting him get over the gaps..... I think the fact that she leaves a silence and he has to fix it means he's got more chance of showing that he's a 6.

66-3 She's already asked this question. He's already answered the question previously. He already said there's shortage of engineers in Malaysia, therefore it'd be easy to get a job. So she's lost concentration, which throws the candidate a bit.

### 6.2.9 Affective Comments

A number of non-evaluative comments concerning the attitude of the candidate or the relationship between the two participants were made, for example:

44-3 She seems to be relaxed. The candidate seems to be relaxing a little bit more here when she's talking about her mother and her home. She's obviously feeling easier and feels comfortable talking about that sort of thing

40-6 Okay. When she starts, she actually interacts - tries to interact - with the interviewer - she says *How are you?* So that makes you think 'okay, she's got a bit of confidence', and that immediately puts her in maybe- okay, she's not going to be down low. And then she starts talking, and the tone of her voice initially is confident. So I'm still thinking because she sounds confident. I know you shouldn't think that, but I- it does prejudice me if someone sounds confident. I think okay they're not going to be too ....

8-8 They've lost the momentum in this interview. They started off quite well and she sounded happy and the interviewer sounded happy but they've both lost the momentum a bit, and it's tapering off to a nothing interview. Do you know what I mean? It's like she's lost interest in what they're talking about, and the interviewer doesn't sound very interested any more either. You know how some... Yeah, yeah, I don't know exactly what's going on, but somehow the interview's not working now. She's lost confidence, the girl, the interviewee, the Japanese girl's lost confidence,

it is not an explicit indicator in the IELTS bandscales, raters still focus on fluency. Where there is disfluency they tend to make judgements about candidate ability on the basis of their inferences regarding the reasons for disfluency. In this respect fluency is a problem category - clearly salient to raters and yet potentially performance-specific and differently interpretable.

Comprehension was also commented upon only where candidates experienced problems understanding the interviewer, but, as they are instructed, they tend not to penalise students for misunderstanding.

Finally, it is interesting that raters comment on the communication strategies used by candidates, particularly as the elicitation of such strategies is neither prescribed within the interview format, not consistent across performances, and, indeed, is generally a result of problems in the communication. Whilst most of the comments are positive it is not clear to what extent the raters take their occurrence into account in awarding scores. Whilst there is an explicit acknowledgement of their relevance in the band descriptors, at least in terms of circumlocution, they are not addressed systematically either in the test or the descriptors. It may be time to consider either their explicit and systematic inclusion or their removal from consideration in general.

## 7.3 Interpretations of Candidate Behaviour

We would concur with the findings reported by Pollitt and Murray (1993) that many of the raters' comments consisted of inferences based on the candidates' behaviour, and that these inferences often differed from rater to rater. We have commented more than once before now on the amount of interpretation that raters engage in. Inferences were frequent, and typically used to excuse or explain certain patterns of behaviour and to justify certain scores. They occurred particularly in relation to fluency, the use of speculative language and the production of extended discourse, and were often concerned with the candidates' maturity.

While we would agree with Pollitt and Murray's statement that 'Given the subjective nature of the interpretative process, there is, then, clearly room for variability in the ways in which different judges perceive a performance', we feel it is important to consider why it happens and, in the interests of fairness, how it can be constrained. As with the assessment of any complex performance, some ambiguity arises when certain required aspects of performance are not demonstrated: Is the candidate not capable of this skill, or is the candidate capable but did not display for another reason? Given the frequent (and typically negative) evaluations of the interviewer, coupled with the interpretations referred to above, it appears to be the case here that raters tended to give the candidate the benefit of the doubt, particularly in relation to lack of evidence of extended discourse and speculation. Perhaps the time has come to tighten up the elicitation process in performance tests such as IELTS in order to ensure that *all* candidates are required to demonstrate specific skills, even those candidates who are not familiar with the test requirements, so that assessments can be based more directly on what does occur rather than what doesn't occur. This would mean, for example, that candidates are *explicitly* instructed to produce extended discourse, rather than its production being a result of test wiseness or personality, as appears often to be the case now.

## 7.4 The Interviewer

The raters in this study were constantly aware of the fact that the interviewer is implicated in a candidate's performance. Given the extent to which interviewers' behaviour varies from interview to interview, part of the raters' dilemma in tape-based assessments is an attempt to disentangle the two so that a score can be awarded to the candidate *alone*. A further complication lies in the fact that each performance (each *interaction*) is unique, and that certain behaviours which will be noticed by the rater may occur in one performance and not in another, for example, failure of the candidate to comprehend, the chance/need to demonstrate certain communication strategies. In addition, the choice of topic and the way it is addressed will vary, as will the interviewers' interviewing style, and these will have implications for the candidates performance.

Whilst operational ratings of IELTS do not require raters to assess from tape so the question of how to compensate for the interviewer does not arise, it does lead us to ask to what extent performances elicited by two different interviewers will differ, and what the implications of this may be for candidates. This is the subject of a current study (Brown 1998).

## 7.5 The Band Scales

Finally, as this study has shown, the rating of complex communicative performances such as that exhibited in the IELTS interview is a difficult task, especially where they are guided by brief and necessarily vague holistic band scales. It is and will always remain an 'imprecise science' and raters deserve to be given credit for their attempts to make sense of the interaction and quantify it as they are required to do. We are perhaps over-ambitious to expect patterns of rating which are consistent across interviews and across raters. Perhaps we must in the end accept that raters of performances such as these must be allowed their individuality and their internal variability, and that the best we can hope for is that they ultimately conform to some notional standard of *reliability*. Perhaps, we should look for other ways to ensure fairness for candidates. One way, of course, is to use more constrained and explicit tasks and criteria, but the danger here is the potential loss of communicativeness, or at least interactiveness. Another is the use of multiple ratings, which would avoid putting the entire responsibility on a single rater and expecting them to perform the impossible and produce a replicable and justifiable single score.

## Bibliography

Alderson, J.C. & Clapham, C. (1992) *Examining the ELTS test: An account of the first stage of the ELTS revision project. IELTS Research Report 2.* The British Council, The University of Cambridge Local Examinations Syndicate & the International Development Program of Australian Universities and Colleges.

Bachman, L. (1988) Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition* 10: 149-164.

Bachman, L. & Savignon, S. (1986) The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal* 70: 380-390.

Brown, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12: 1-15.

Brown, A. (1998) Interviewer style and candidate performance in the IELTS Oral Interview. Paper presented at the Language Testing Research Colloquium, Monterey, CA.

Brown, A. and Hill, K. (1988) Interviewer style and candidate performance in the IELTS oral interview. S. Wood (Ed.) *IELTS Research Reports 1.* Sydney: ELICOS.

Brown, A. and Lumley, T. (1997) Interviewer variability in specific-purpose language performance tests. In Kohonen, V., Huhta, A., Kurki-Suonio, L. & Luoma, S. (Eds.) *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96.* Jyväskylä: University of Jyväskylä and University of Tampere.

Cafarella, C. (1994) Assessor accommodation in the V.C.E. Italian oral test. *Australian Review of Applied Linguistics* 20: 21-41.

Chalhoub-Deville, M. (1995) Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12: 16-35.

Cohen, A.D. and Hosenfeld, C. (1981) Some uses of mentalistic data in second language research. *Language Learning* 31: 285-313.

Criper, C. and Davies, A. (1988) *ELTS validation project report. ELTS Research Report 1(i).* Hertford, UK: The British Council & University of Cambridge Local Examinations Syndicate.

Cronbach, L.J. (1970) *Essentials of psychological testing.* New York: Harper and Row.

Cronbach, L.J. (1971) Test validation. In R.L. Thorndike (Ed.) *Educational measurement. Second edition.* Washington, DC: American Council on Education.

Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing* 7: 31-51.

de Jong, J.H.A.L. and van Ginkel. L.W. (1992) Dimensions in oral foreign language proficiency. In L. Verhoeven and J.H.A.L. de Jong (Eds.) *The construct of language proficiency* (pp.187-205). Amsterdam: John Benjamins.

Delarulle, S. (1997) Text type and rater decision-making in the writing module. In Brindley, G. & Wigglesworth, G. (Eds.) *Access: Issues in English language test design and delivery* (pp. 215-242). Sydney: National Centre for English Language Teaching and Research.

DiPardo, A. (1994) Stimulated recall in research on writing: An antidote to ' I don't know, it was fine'. In Smagorinsky, P. (Ed.) *Speaking about writing: Reflections on research methodology.* Thousand Oaks, CA: Sage.

Ericsson, K. and Simon, H. (1984) *Protocol Analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Filipi, A. (1994) Interaction in an Italian oral test: The role of some expansion sequences. In Gardner, R. (Ed.) *Spoken interaction studies in Australia: Australian Review of Applied Linguistics Series S, No. 11*: 119-136.

Green, A.J.K. (1997) *Verbal protocol analysis in language testing research: Studies in Language Testing 5.* Cambridge: Cambridge University Press.

Griffin, P. and McKay, P. (1992) Assessment and reporting in ESL Language and Literacy on Schools project. In P. McKay (Ed.) *ESL development: Language and literacy in schools: Tapping the potential Vol. 2* (pp. 9-28). Canberra: Commonwealth of Australia..

Hadden, B.L. (1991) Teacher and nonteacher perceptions of second-language communication. *Language Learning* 41: 1-24.

Halleck, G. and Reed, D. (1996) Probing above the ceiling in oral interviews: what's up there. In Kohonen, V., Huhta, A., Kurki-Suonio, L. & Luoma, S. (Eds.) *Current Developments*

*and Alternatives in Language Assessment: Proceedings of LTRC 96* (pp. 225-238). Jyväskylä: University of Jyväskylä and University of Tampere.

Huot, B. (1990) Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication* 41: 201-213.

IELTS (2000) *IELTS Handbook.* University of Cambridge Local Examinations Syndicate, British Council and IDP Education Australia Ltd.

Ingram, D. and Wylie, E. (1996) The general modules: Speaking. In Clapham, C. & Alderson, J.C. (Eds.) *Constructing and trialling the IELTS test. IELTS Research Report 3.* The British Council, The University of Cambridge Local Examinations Syndicate & the International Development Program of Australian Universities and Colleges.

Kelly, G.A. (1955) *The psychology of personal constructs. Vols 1 and 2.* Norton: New York.

Lazaraton, A. (1996a) Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing* 13: 151-172.

Lazaraton, A. (1996) A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). In Milanovic, M. & Saville, N. (Eds.) *Performance Testing, Cognition and Assessment. Studies in Language Testing 3* (pp. 18-33). Cambridge, UK: Cambridge University Press.

Lazaraton, A. (1997) Preference organisation in oral proficiency interviews: The case of language ability assessments. *Research on Language and Social Interaction* 30: 53-72.

McNamara, T. and Lumley, T. (1997) The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing* 14: 140-156.

McNamara, T.F (1990) Item response theory and the validation of an ESP test for health professionals. *Language Testing* 7: 52-76.

Meiron, B.E. (1998b) Rating oral proficiency tests: a triangulated study of rater thought processes. Unpublished MA thesis: UCLA.

Milanovic, M. and Saville, N. (1994) An investigation of marking strategies using verbal protocols. Paper presented at 16th Language Testing Research Colloquium, Washington, DC, March 1994.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Studies in Language Testing 3* (pp. 92-114). Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

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

Munby, J. (1978) *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

Neeson, S. (1985) An exploratory study of the discourse structure of the Australian Second Language Proficiency Ratings test of oral proficiency. Unpublished MA thesis, University of Birmingham.

Perrett, G. (1990) The language testing interview: A reappraisal. In de Jong, J.H.A.L. & Stevenson, D.K. (Eds.) *Individualising the Assessment of Language Abilities* (pp. 225-237). Clevedon, UK: Multilingual Matters.

Pollitt, A. and Murray, N.L. (1993) What raters really pay attention to. In Milanovic, M. and N. Saville (Eds.) *Performance Testing, Cognition and Assessment. Studies in Language Testing 3*. Cambridge: Cambridge University Press.

Powers, D. and Stansfield, C.W. (1983) *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions. TOEFL Research Report 13.* Princeton, NJ: Educational Testing Service.

Raffaldini, T. (1988) The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition* 10: 197-216.

Ross, S. (1992) Accommodative questions in oral proficiency interviews. *Language Testing* 9: 173-186.

Ross, S. (1996) Formulae and inter-interviewer variation in oral proficiency interview discourse. *Prospect* 11, 3: 3-16.

Ross, S. and Berwick, R (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14: 159-176.

Shohamy, E. and Walton, A.R. (eds.) (1992) *Language assessment for feedback: Testing and other strategies.* Dubuque, Iowa: Kendall/Hunt.

Smagorinsky, P. (1994) *Speaking about writing: Reflections on research methodology.* California: Sage.

Van Lier, L. (1989) Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversations. *TESOL Quarterly* 23: 480-508.

Vaughan, C. (1991) Holistic assessment: what goes on in the rater's mind? In Hamp-Lyons, L. (ed.) *Assessing second language writing in academic contexts* (pp 111-125). Norwood, NJ: Ablex Publishing Corporation.

Wall, D., Clapham, C. and Alderson, J.C. (1994) Evaluating a placement test. *Language Testing* 11: 321-344.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing* 10: 197-223.

Young, R. & Milanovic, M. (1992) Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14: 403-424.