IELTS™
English for International Opportunity

1. Interviewer Style and Candidate Performance in the IELTS Oral Interview

Annie Brown and Kathryn Hill
NLLIA Language Testing Research Centre
Department of Linguistics and Applied Linguistics
The University of Melbourne

Publishing details

International English Language Testing System (IELTS)

Research Reports 1998

Volume 1

Editor: Sandra Wood

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

ELICOS Association Limited
ACN 003 959 037

National Library of Australia
Cataloguing-in-Publication Data
1999 ed
IELTS Research Reports 1999 Volume 2
ISBN 0 86403 021 5

1 Interviewer Style and Candidate Performance in the IELTS Oral Interview

Annie Brown and Kathryn Hill
NLLIA Language Testing Research Centre
Department of Linguistics and Applied Linguistics
The University of Melbourne

Abstract

Recent research into the validity of oral language interviews has extended the focus beyond that of statistical analysis to investigations of the structure of the interview discourse itself, and to the language produced by both candidate and interviewer. Research has indicated that, despite training, interviewer behaviour varies considerably in terms of the amount of support they give candidates, the amount of rapport raters consider them to have established with candidates and the extent to which they follow the instructions in terms of the type of discourse elicited from candidates. While several writers allude to the potential of such variable interviewer behaviour to affect the validity of tests, studies have not yet empirically investigated the relationship between interviewer behaviour and candidate performance.

The study aims firstly to investigate the extent to which differential behaviour by IELTS interviewers affects the scores awarded to candidates and to identify interviewers who consistently present a difficult or easy challenge to candidates. The second part of the study involves a discourse analysis of the contributions of 'difficult' and 'easy' interviewers, and aims to identify aspects of interviewer behaviour which contribute to the challenge they present.

The study is based on interviews undertaken with 32 candidates, each of whom was interviewed twice by two different interviewers. Six interviewers took part in the study. The interviews were audio-taped and multiple-rated.

The test data were analysed using the multifaceted Rasch analysis program FACETS (Linacre, 1989) in order to identify cases where candidates perform differentially in the two interviews, as well as identifying interviewers who consistently elicit poorer or better performance. A total of 10 interviews from the two most difficult and two easiest interviewers were transcribed and analysed.

It was found that the easier interviewers tended to shift topic more frequently and asked simpler questions, spending longer in Phase 2 of the interview. The more difficult interviewers tended to use a broader range of interactional behaviours, such as interruption and disagreement as well as asking more challenging questions.

While the intent in the development of the IELTS interview has not been to standardise interviewer behaviour to the extent that all candidates receive exactly the same prompts, there must be some concern to ensure that all candidates are treated equally in terms of the challenge presented by the interviewer. By making explicit those features of interviewer behaviour which have the potential to affect the quality of the candidates' performance, this study is of relevance to the training of raters in terms of increasing their understanding of the effect of their performance on that of the candidate and in ensuring the comparability of the challenge presented to different candidates.

1.0 Introduction

This paper reports on a study into the extent to which differential behaviour by IELTS interviewers can affect the scores awarded to candidates, and which features of interviewer behaviour might contribute to this. Until recently there has been little focus on interviewer variation and the effect this might have on candidates' scores, the assumption being that variability in interviewer behaviour is not a source of unreliability in the same way as rater, or even task, is. Test developers have long been aware of the variability inherent in *rater* behaviour. Steps are generally taken to minimise this variability through the provision of explicit band descriptors, through initial and follow-up rater training, through the use of multiple ratings and, in some cases, through the use of Item Response Theory to compensate for rater harshness. Again using Item Response Theory, test tasks may be equated or scores may be adjusted to compensate for variation. Little, however, is yet understood about the extent of interviewer variation and its implications. This study attempts to add some understandings to what is a growing area of concern amongst language testers.

Oral interviews, such as that forming part of the IELTS test, generally follow a prescribed format. Interviewer training introduces prospective interviewers to the format of the interview and to relevant interviewing techniques. Nevertheless, the intent is normally *not* to standardise interviewer behaviour to the extent that all candidates receive exactly the same prompts; however, it would seem that personality and background factors are likely to influence the interviewing style adopted by individuals (just as they have been found to affect the awarding of scores) so there must, nevertheless, be some concern to ensure that all candidates are treated equally in terms of the support and challenge offered by the interviewer. Research into the discourse produced in oral interviews and the effect of individual interviewers on candidate performance can inform interviewer training and contribute to fairness for candidates.

This study aims to explore interviewer differences in both quantitative and qualitative terms. It does this firstly, by identifying whether interviewer style does in fact have an effect on scores, and secondly by using discourse analysis to explore the features of interviewing style which characterise 'difficult' and 'easy' interviewers; 'difficult' interviewers being those with whom a candidate is more likely to receive a lower score than with an 'easy' one. It is hoped that the findings of this study will contribute to the understandings beginning to emerge from other research into interviewer behaviour, and inform the process of interviewer training.

2.0 Research into Interviewer Behaviour

In the last few years, research into oral language interviews has begun to investigate the discourse produced by the participants. This research indicates that, despite training, interviewer behaviour appears to vary considerably in terms of the amount of support given to candidates (Ross and Berwick, 1990; Ross, 1992; Lazaraton and Saville, 1994), the amount of rapport established with candidates (Lumley and McNamara, 1993), and the extent to which the interviewer guidelines are followed in terms of the type of discourse elicited from candidates (Lazaraton, 1993; Lumley and Brown, forthcoming).

Ross and Berwick (1990) demonstrated a relationship between the amount of accommodation (modification of the 'form and content of the discourse in order to facilitate communication') provided by an interviewer and the score awarded. However, there has been no research into whether different interviewers interviewing *the same candidate* vary in the amount of accommodation they make and whether this might have an effect on the score awarded; in other words, whether the candidate would get a different score depending on who the interviewer was.

Ross (1992) again investigated accommodation within oral interviews, this time identifying the causes of accommodation. Using variable rule analysis he identified four factors: interviewee response to previous question, structure of response to previous question, outcome of the interview, and use of accommodation in the previous question. Again, however, no comparison of the use of accommodation was made across interviewers.

Lazaraton and Saville's 1993 study reported on an investigation of interviewer difficulty in CASE. However, as candidates were not double tested, it is not clear how the measures of interviewer difficulty were arrived at. Nevertheless, the authors identify several aspects of interlocutor support, including supplying vocabulary, rephrasing questions, evaluating responses, echoing and correcting responses, using interview prompts that require only confirmation and drawing conclusions for candidates.

In another study Lumley and McNamara (1993) obtained multiple ratings of Occupational English Test (OET) interviews. In addition to providing ratings of the candidates using the normal test rating scale, raters were asked to provide an assessment of the rapport established between interviewer and candidate. They found that raters tended to compensate for what they perceived as poor rapport. In other words, candidates received higher scores where the interviewer was perceived by the rater as 'difficult'. This finding is relevant to the present study in that interviewer 'difficulty' may be masked because of compensation by the raters.

Lumley and Brown (forthcoming) investigated nurses' perceptions of interviewer performance in OET role plays. They found that a wide variety of behaviours were considered 'authentic' but that different challenges were set for candidates according to the extent to which interviewers performed the role play as instructed, ie. with some degree of conflict, rather than engaging in more 'teacher-like' behaviour and supporting and agreeing with the candidate. Again, no study was made of the effect different interviewers might have on perceptions of candidate ability.

Nevertheless, a discourse analysis did indicate that certain interviewers have entrenched patterns of behaviour, that is, they consistently provided more or less support than other interviewers.

In conclusion, despite the growing literature on observed interviewer variation in terms of the discourse they produce, there has to date been little empirical analysis of the relationship between this and candidate scores. This study combines a qualitative approach, involving the analysis of actual test interactions, with a quantitative study using multiple interviews conducted by trained IELTS interviewers and multiple ratings. The stages of the study are as follows:

(i) using multi-faceted Rasch analysis, determine whether different interviewers represent different 'hurdles' in terms of the difficulty of doing an IELTS interview;

(ii) identify cases where candidates perform differentially in each of the two interviews they undertake;

(iii) transcribe and analyse these interviews in order to identify whether there are particular interviewing styles which characterize 'easy' or 'difficult' interviewers and which may contribute to better or worse performance by candidates.

3.0 The IELTS Interview and Rating

IELTS Speaking Module[1] 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 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 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 interview is scored using a set of global bandscales with ten levels (0-9).

[1]This information is quoted from the IELTS handbook.

4.0 Methodology

Thirty-two students from IELTS preparation courses and six accredited interviewers participated in the study. Each of the thirty-two candidates was interviewed twice by two different interviewers. In order to ensure that candidates were not exposed to the same topic twice, and to avoid any practice effect, in this study the suggested interview topics for the Extended Discourse section (Phase 2) and Speculation and Attitudes section (Phase 4) were divided into two lists. Interviewers were instructed to draw either on List A or on List B for each interview. See Appendix 1.1 for the information given to the interviewers about the phases of the interview and their content focus.

The interviews were audio-taped and each tape was later rated four times by seven accredited IELTS raters.

The candidates were all ELICOS students who at the time of the interviews were preparing to take IELTS prior to submitting applications for tertiary study in Australia. Hence there was a high level of motivation on the part of the candidates to take part in the interviews so as to gauge their readiness to take the test. Candidates were informed that if they agreed to take part in the study, undertaking two IELTS interviews each, they would receive an informal assessment of their proficiency in the oral component of IELTS. This assessment was given at the end of the second interview rather than the first interview as this would potentially discourage the candidate from proceeding to the second interview.

The interviewers were all accredited and practising IELTS interviewers who responded to a request for assistance with an IELTS research project. In order not to affect their behaviour when interviewing, they were not given any information about the focus of the research other than that it was 'looking at' the IELTS interview; most assumed that the focus was on the candidates. They were informed after the interviews had been completed of the aims of the study.

Each of the 32 candidates was interviewed twice, each time by different interviewers. The interviews were carefully planned so that the interviewers were equally assigned to first and second interviews, and so that they overlapped in their pairings, ie. they were each paired with several of the other interviewers rather than being paired with just one in order to allow for calibration of the interviewers against each other. Where two interviewers interviewed several candidates in common, the number of first and second interviews each carried out by each interviewer was balanced. As has already been mentioned, the interviews were controlled to the extent that no candidate was subjected to the same Phase 2 and 4 topics in either interview in order to avoid a practice effect.

The interviews were audio-taped and each interview was later rated from the tape using accredited IELTS raters [2]. In order to take rater harshness into account (ie. to compensate for it in the estimate of candidate ability), each tape was rated four times using a patterned design of any four of the seven raters employed. This overlap between raters enables the program used to analyse the data to model 'rater' as a facet and hence compensate for the effect of rater harshness.

[2] The interviewers also gave a rating (as is normal practice in IELTS administrations) but this data was not used for the present study.

The analysis was done in two stages:
a) The multi-faceted Rasch analysis program FACETS (Linacre, 1989) was used to analyse the test data. Facets which are normally considered to contribute to a candidate's score are candidate ability and rater harshness[3]. In this study we are trying to determine whether interviewer 'difficulty' may be an additional factor. Specifically, we wanted to identify whether different interviewers represent different 'hurdles' for candidates in terms of the difficulty of doing an IELTS interview, in that they consistently elicit poorer or better performances from candidates.

Through the use of IRT analysis it is possible to compensate for rater harshness and derive candidates 'fair scores'[4]. We were able therefore to identify cases where, after compensating for the effect of the particular raters involved, a candidate's performance in the two interviews was judged to be at two different levels of ability, and also to identify the extent of the difference.

b) In the second part of the analysis, pairs of interviews were chosen where the same candidate performed at different levels and selected interviews were transcribed. An analysis was undertaken in order to identify whether there are particular patterns of interviewer behaviour which contribute to better or worse performance by candidates. While differential performance may be due to factors other than interviewer behaviour, such as choice of topic, motivation or other aspects of the interviewer-candidate relationship, this study attempts to isolate those features of interviewer behaviour which co-vary with candidate performance. The analysis focused on a range of potentially relevant aspects of interview technique. These were drawn to some extent from previous research into oral interview discourse and included aspects such as questioning technique and topic organisation.

[3] In cases where the tasks are substantially different, task difficulty may also be included; in this case task was not considered as it was felt that variability due to topic was considerably less likely to affect scores than variation in interviewer behaviour

[4] The fair score in a FACETS analysis represents a modification of the actual score(s), taking other variables (facets) into account. In this case, as *rater* is a facet of the analysis, it compensates for rater harshness.

5.0 The Analysis

Question 1: Are there significant differences in interviewer difficulty?

An analysis (Analysis 1) was carried out using FACETS, with four facets: *candidate*, *interviewer*, *occasion* and *rater*, in order to estimate interviewer difficulty.
The findings of this analysis are shown in Table 1.

	Interviewer ID	Interviewer Difficulty (logits)	Model SE	Model Fit			
				Infit		Outfit	
				MnSq	Std	MnSq	Std
most difficult	5	0.75	0.42	0.4	-2	0.3	-2
	6	0.48	0.45	1.1	0	1.1	0
	3	0.15	0.22	0.9	0	1.0	0
	1	0.01	0.24	1.0	0	1.0	0
	2	-0.52	0.33	1.4	1	1.4	1
easiest	4	-0.86	0.25	0.7	-1	0.7	-1
	RMSE 0.33 Adj S.D. 0.44 Separation 1.34 Reliability 0.64						
	Fixed (all same) chi-square: 17.9 d.f.: 5 significance: .00						
	Random (normal) chi-square: 4.9 d.f.: 4 significance: .30						

Table 1 *Interviewer difficulty*

The interviewer difficulty measures are presented in logits, the units of measurement used within Rasch analysis. As can be seen, these range from 0.75 logits (the most difficult interviewer) to -0.86 logits (the easiest interviewer). The separation information given within the FACETS analysis and reproduced in Table 1 above confirms that there are significant differences amongst this group of interviewers in terms of their difficulty: the interviewer separation index indicates 1.34 statistically distinct interviewer strata[5], separated with a reliability of 0.64. The low reliability (generally 0.8 is considered acceptable) is most likely a consequence of the small sample size. In addition, there is a 0.00 probability that the interviewers can be considered equally severe (the 'fixed' chi-square), although there is a 0.30 probability that they are not sampled at random from a normally distributed population (the 'random' chi-square). This latter statistic is again likely to be a consequence of the small n-size.

Turning to the fit of the interviewers to the model, as shown in Table 1, we can consider all the interviewers to be reasonably well fitting to the model. That is, none of the fit indices are unacceptably high (standardised scores ranging from +2 to -2 are generally considered acceptable).

[5] where these strata are defined by their centres being three measurement errors apart

In order to determine exactly which pairs of raters presented a significantly different level of difficulty for candidates, the following calculation was carried out:

Is the difference in difficulty measures greater than the square root of the sum of the two standard errors squared, ie.

$$\text{Is } d1-d2 > \surd(se^2+se^2) \text{ ?}$$

The result of this calculation is presented in Table 2.

To summarise Table 2, Interviewer 4 (the 'easiest') presents a significantly different level of difficulty from Interviewers 5, 6, 3 and 1 (the four most 'difficult' interviewers). In addition, Interviewer 2 (the second 'easiest') presents a significantly different level of difficulty from Interviewer 5 (the most 'difficult').

Pairs of Interviewers	Difference in Difficulty (d1-d2) (logits)	$\surd(se^2+se^2)$	Significant Difference
5 and 4	1.61	0.97	✓
5 and 2	1.27	1.07	✓
5 and 1	0.74	0.97	-
6 and 4	1.34	1.03	✓
6 and 2	1.00	1.12	-
3 and 4	1.01	0.67	✓
3 and 2	0.67	0.79	-
1 and 4	0.87	0.69	✓
2 and 4	0.34	0.83	-

Table 2 *Paired differences in interviewers*

It appears then, that interviewer difficulty may well affect a candidate's chances, in that the ability level construed for the candidate will be *not only* a result of his/her inherent ability, but *also* of the difficulty presented by the interviewer. This will be particularly the case where an interviewer at the extremes of the 'difficulty' continuum is used.

Question 2: **Can we identify pairs of interviews where the same candidate was judged as being of a different level of ability on each occasion, and to what extent are these differences consistent with interviewer difficulty?**

Before comparing scores across the two interviews it was necessary to ascertain the extent of any effect for 'occasion' (first or second interview). It was conceivable that any of a number of factors may come into play here to either increase or decrease the 'difficulty' of the second interview in relation to the first. It was, for example, possible that there may be a practice effect which would make it easier for candidates to gain a higher score on the second interview. While the topics had been carefully assigned to ensure that no candidate was exposed to exactly the same Phase 2 and 4 topics, there was still the likelihood that the format would be more

familiar and hence easier the second time around. On the other hand, it was also conceivable that fatigue or boredom might have the opposite effect, with candidates scoring lower on the second interview.

The FACETS analysis which included 'occasion' as a facet (Analysis 1) confirmed that occasion did indeed present a significant difficulty factor. The separation information on the facet 'occasion' was: Separation 1.99 ; Reliability 0.80 ; Fixed (all same) chi-square: 9.9 d.f.: 1 significance: .00

We were able to determine the extent of the effect of occasion by comparing the mean fair score (an average score adjusted for rater harshness but not converted to a logit) for all first interviews with the mean fair score for all second interviews. In order to do this a further FACETS analysis (Analysis 2) was set up with two facets, *candidate* and *rater*. In this analysis each interview was treated independently, resulting in two scores for each candidate, ie. one for each interview. A grouping facility was used to enable us to compare the mean of all occasion 1 scores with the mean of all occasion 2 scores. When the means of the fair scores on each occasion were compared, a difference of 0.2 of a band was found, with the first interview attracting the higher score.

Candidate	Occasion 1 Fair Average	Inter-viewer	Occasion 2 Fair Average	Occasion 2 Adjusted for Difficulty	Inter-viewer	Difference in Fair Score	Expected Direction of Difference
35	7.3	4	7.1	7.3	1	-	-
03	7.2	5	7.4	7.6	4	.4	3
25	5.9	6	6.9	7.1	2	.8	3
02	6.8	1	6.2	6.4	4	.4	5
21	6.8	4	6.4	6.6	5	.2	3
24	6.6	6	5.9	6.1	2	.5	5
06	6.5	2	5.4	5.6	6	.9	3
37	6.3	3	6.6	6.8	4	.5	3
14	6.3	3	6.2	6.4	4	.1	3
01	5.9	3	6.1	6.3	4	.4	3
18	5.9	4	4.9	5.1	5	.8	3
16	5.8	4	5.0	5.2	5	.6	3
15	5.4	3	6.2	6.4	4	1.0	3
38	5.2	2	5.0	5.2	3	-	-
19	4.3	5	4.3	4.5	3	.2	3

Table 3 *Interview pairs - score differences*

In order to make the first and second interview comparable 0.2 was added to the fair score of each candidate for the second interview. We then compared pairs of interviews involving the same candidate in order to identify firstly, cases where candidates received a different score on

each occasion, and secondly, whether these differences were consistent with what was known about the relative difficulty of the interviewers involved.

As not all interviewers were significantly different from each other, we only considered cases where the two interviewers were not adjacent in terms of difficulty rankings, a total of 15 pairs (Table 3). Of these, there were only two instances where there was no score difference and only two instances where the direction of the score difference was unexpected (ie. the candidate got a better score with the more difficult interviewer).

Six pairs of interviews, highlighted in Table 3, were selected for transcription: of these, 10 interviews were used in the analysis, two each from the two most difficult interviewers (Interviewers 5 and 6), two from the second easiest (Interviewer 2) and four from the easiest (Interviewer 4).

6.0 Discourse Analysis

6.1 Number and length of turns

A count was made of the total number of turns by each interviewer. These turns were classed either as 'interview' turns (turns aimed at eliciting information) or 'feedback' turns.
Types of feedback included:

i) minimal feedback (mm, yes, right, is it?, etc.)

ii) evaluative comment, eg.

57.47[6] *suits you*
32.14 *sounds lovely*

iii) summary comment, eg.

43.21 *and I am sure you have learnt a lot from that*
46.28 *even the women here are taller.*

iv) echo (repeating part of previous answer)

v) correction (repetition of part or whole of previous response, supplying correct grammar or more precise lexis)

vi) clarification questions (where the interviewer did not catch what the candidate said).

[6] Numbers refer to tape and interviewer turn

Interviewer (difficult to easy)	Tape	Turns Requiring Response	Feedback Turns	Unknown	Total Number of Turns
5	*46*	26	10	1	37
	50	42	11	3	56
6	*57*	38	9	1	48
	32	24	6	-	30
2	*66*	44	1	1	46
	8	40	1	-	41
4	*43*	41	26	-	67
	44	62	12	-	74
	45	59	16	-	75
	27	33	9	-	42

Table 4 *Interviewer turns*

From Table 4 we can see that the easiest interviewer, Interviewer 4, tends to conduct longer interviews than the others in terms of the total number of turns. This interviewer also tends to ask a larger number of information-seeking questions than the other interviewers, as well as a tendency towards more frequent use of feedback. The second easiest interviewer, in contrast with the other three, rarely provides feedback alone: on the few occasions when she does provide feedback she follows up immediately with a question:

66.03 *one and a half months, ah good, um, where do you come from?*
66.04 *Malaysia, and have you got a family in Australia?*
8.07 *your dog? Ah how lovely. You have a pet too, and who's looking after it now?*

The two most difficult interviewers both varied in the number of questions they asked in each of their two interviews. Given the variation shown by all four interviewers in this data, one cannot here infer any connection between length and difficulty. Further studies focusing specifically on length may, however, reveal some relationship between the amount of information supplied and the ability inferred by the assessor.

Table 5 presents information on the balance of talk in the interview between candidate and interviewer, and average length of turn. It shows that each interviewer is consistent in the length of their turns, and that with the exception of the second most difficult interviewer (Interviewer 6) this is around 10 words. Interviewer 6's turns are roughly double this length. Candidates, on the other hand, are more varied in the amount of speech they produce as would be expected (weaker candidates being more likely to produce shorter turns). The length of candidates' turns also tends to be similar in each of their two interviews.

Inter-viewer	Tape	Cand-idate	Number of Turns	Number of Words		% Inter-viewer Talk	Average Length of Turn (words)	Average Length of Response (words)
5	*46*	*3*	37	I	365	31	9.9	
				C	806			21.8
	50	*18*	56	I	560	60	10	
				C	381			6.8
6	*57*	*6*	48	I	945	64	19.7	
				C	527			11
	32	*25*	30	I	642	47	21.4	
				C	713			23.8
2	*66*	*6*	46	I	550	50	11.9	
				C	541			11.8
	8	*25*	41	I	424	34	10.3	
				C	807			19.7
4	*27*	*37*	42	I	495	50	11.8	
				C	1000			23.8
	44	*18*	74	I	786	56	10.6	
				C	623			8.4
	43	*3*	67	I	532	30	7.9	
				C	1263			18.8
	45	*15*	75	I	758	43	10.1	
				C	1021			13.6

Table 5 *Interviewer and candidate turns*

6.2 Question forms

The interviewers' questions were classified according to whether they were open or closed. Closed questions included those which

i) required a yes/no response:

44.0 *Do you live in a flat?*

27.05 *Is that near the university?*

ii) expected confirmation:

50.33 *but sometimes you'd eat Indian?*

44.71 *....you're generally quite happy here at the moment?*

iii) required the selection of one of two alternatives offered:

9.24 *and are the marriages arranged or do the young people meet each other by themselves?*

Table 6 presents the findings of this analysis. There does not appear to be any marked difference between easy and difficult interviewers in their choice of question form.

Inter-viewer	Tape	Candidate	Total Turns requiring response	Open Questions	Yes/No Questions	Confirm-ation Questions	Alter-native Questions
5	*46*	*3*	26	13	11	2	-
	50	*18*	42	22	12	2	6
6	*57*	*6*	38	15	16	7	3
	32	*25*	24	16	5	1	2
2	*66*	*6*	44	17	24	3	-
	8	*25*	40	24	12	2	2
4	*43*	*3*	41	23	16	1	1
	44	*18*	62	40	19	1	2
	45	*15*	59	31	23	5	-
	27	*37*	32	16	14	1	1

Table 6 *Question forms*

6.3 Question focus

The interviewers' turns were classified according to the question focus or content. It was hypothesised that easier interviewers would be characterised by more frequent use of simpler questions (those asking for simple factual information and description) rather than the more complex skills of speculating or presenting and justifying an opinion. Accordingly, questions were categorised as follows:

Type 1 Simple factual information - personal and general

9.05 *and how many, do you have brothers and sisters?*
57.11 *ten hours, so what time would they normally start?*
44.32 *What are your favourite kind of movies?*

Type 2 Feelings

45.25 *Oh dear, so you had to move did you? Are you happy at the moment?*
45.18 *and do you like living in Melbourne city?*

Type 3 Straightforward description

57.22 *no, so what happens to those people?*
43.16 *What do they do at midnight?*

Type 4 Personal plans

43-65 *So now you have this year to prepare for 1997, what are you going to do next year, X?*

Type 5 Considered response: requires judgement or analysis to select content

43-19 *...so when you think of ideal living conditions for yourself what would you choose next?*

50.38 *in commerce, right, why commerce?*

44.67 *did you, right, so you've just been here a short time, what are your first impressions of Australia?*

Type 6 Speculation

66.47 *Do you think it would be easy to earn a living? How far to engineering, would that be easy to get a living in Malaysia?*

Type 7 Confirmation of understanding

50.19 *That's in [name of city] they have those?*

Interviewer	Tape	Candidate	Total turns requiring response	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
5	*46*	*3*	26	16	1	1	3	2	2	1
	50	*18*	42	22	1	7	3	6	1	2
6	*57*	*6*	38	18	1	3	3	6	5	2
	32	*25*	24	6	1	1	4	8	4	-
2	*66*	*6*	44	25	-	2	6	4	3	3
	8	*25*	40	14	2	7	4	10	-	2
4	*43*	*3*	41	26	1	3	1	7	2	1
	44	*18*	62	34	1	16	2	9	-	-
	45	*15*	59	30	3	8	3	10	-	5
	27	*37*	33	23	-	2	1	6	-	1

Table 7 *Question focus*

What we find in Table 7 is that of the number of turns requiring a response from the candidate, there do not appear to be any significant patterns in the number of turns allocated to each question type. However, there are three tendencies apparent in this data:

1) The largest percentage of all interviewers' questions are of the simple factual type. Interviewer 4, the easiest interviewer, tends to ask more of these than the other interviewers.
2) Interviewer 4 failed to ask any speculative questions in three of the four interviews, as did Interviewer 2 in one of the two she carried out.
3) Interviewer 4 asked fewer questions about the candidate's personal plans.

6.4 Topic

Table 8 shows the number of topics introduced in each interview as well as the number of turns and subtopics within each topic. Examples of topic and subtopic include the topic *how the candidate lives* with subtopics *the flat* and *food* (Tape 45); the topic *studying in Singapore* with subtopics *language* and *exams* (Tape 43).

What we find is that the easiest interviewer, Interviewer 4, introduces many more topics than the other interviewers. For example, candidate 18 experienced 9 topic shifts with Interviewer 4 compared with 2 topic shifts with Interviewer 5. For the other three interviewers the smaller number of topics was accompanied by a larger number of turns within each topic. The number of subtopics within a topic does not seem to distinguish difficult and easy raters.

Table 8 also shows the number of turns in Phase 2 (Extended discourse) and Phase 4 (Speculation and attitudes) as well as the total number of turns for each interview. We find that the more difficult interviewers devoted roughly the same number of turns to each of Phases 2 and 4. In contrast, for Interviewer 4 the overwhelming majority of turns occur in Phase 2 (e.g. Tape 45 Phase 2 = 69 turns, Phase 4 = 4 turns). This finding is consistent with the earlier finding that Interviewer 4 tends to ask more simple factual questions with fewer questions about personal plans and no questions requiring speculation.

The fact that candidates assigned the easiest interviewer experienced more frequent topic shifts means that they were not required to talk about any topic in depth. It seems then that the interview is 'easier' (or candidates appear more competent) when several topics are touched on briefly rather than fewer topics explored in depth, and where questions are possibly less 'probing'. It may also be that the more questions there are on the one topic, the more complex they become referentially and the less complete grammatically due to the shared knowledge that is being built up. A further analysis will be required in order to investigate this question. It is also worth noting that Interviewer 4's interviews are typically much longer than the others, giving candidates the opportunity to produce more language and more information, either of which may lead raters to perceive a candidate as being more able.

The difficult interviewers not only require the candidate to go into greater depth about the chosen topic, but they also appear less inclined to accommodate their questions to the candidate's level. For example, the more 'difficult' interviewers are much more likely to persist with a topic or predetermined sequence of questioning when a candidate is obviously struggling.

50.21 *What about the types of architecture, what kinds of architecture, style of buildings do you see in X?*

50.C *Local?*

50.22 *Yes, any sort of traditional architecture?*

50.C *Yeah.*

50.23 *Tell me about that.*

50.C *Something like Malay style.*

50.24 *Yeah, what does that look like?*

50.C *There's a lot of um, um, ... the Malay .... they're like in Malay..*

50.25 *What does that mean though?*

50.C *The name of the language is Malay language.*

Inter-viewer	Tape	Cand-idate	Topics	Turns per Topic	Sub-topics	Turns per Phase 2	Turns per Phase 4	Total Turns
5	46	3	1	17	2	17	12	29
			2	12	2			
	50	18	1	29	2	29	22	51
			2	22	3			
6	57	6	1	24	7	24	18	42
			2	18	6			
	32	25	1	13	5	13	12	25
			2	12	7			
2	66	6	1	1		22	20	42
			2	21	2			
			3	20	5			
	8	25	1	22	3	24	8	32
			2	2				
			3	8	4			
4	45	15	1	12	2	69	4	73
			2	9	2			
			3	6				
			4	4				
			5	11	2			
			6	7				
			7	17	2			
			8	3				
			9	4				
	27	37	1	8	2	37	2	39
			2	2				
			3	7				
			4	2				
			5	9				
			6	9	3			
			7	2				
	43	3	1	17	1	59	4	63
			2	9	1			
			3	14	2			
			4	8				
			5	11	2			
			6	4	2			
	44	18	1	23	5	66	4	70
			2	2				
			3	5				
			4	5	1			
			5	17	3			
			6	6	1			
			7	8	1			
			8	2				
			9	2	1			

Table 8 *Topic*

Interviewer 4, on the other hand, rephrases and breaks the question down where the candidate has not produced the required response, either through lack of comprehension or where the interviewers intention was not clear, as in this extract:

44.41 *um, can you tell me about any special festivals that you have in Malaysia?*
44.C *oh, yeah.*
44.42 *any celebrations that everybody has at sometime during the year, can you think of one special one?*

The difficult interviewers are also more likely to challenge the candidate, for example, to justify a decision. Interviewer 5 in one interview challenged the candidate consistently in relation to his study plans, firstly in relation to his chosen subject:

50.43 *so why accounting, isn't it better to learn management than accounting if you want to be, and have your own company?*

Secondly, in relation to his chosen place of study:

50.45 *why aren't you studying in Malaysia?*
50.48 *but why did you come to Australia, why didn't you stay in Malaysia?*
50.49 *but why Australia, why not England or America?*

And thirdly in relation to the relevance of studying the chosen subject in the chosen country:

50.50 *now if you study commerce here, I imagine the course here is very much centred around Australian business, the Australian economy, how are you going to use that in Malaysia?*

Interviewer 5's questioning style could be characterised in two ways. Firstly, he tends to use many fragments rather than complete sentences:

46-09 *so the same state though?*
46-12 *for secondary school?*
50.34 *sometimes Malay?*
50.38 *in commerce, right, why commerce?*

Secondly, a number of his questions are somewhat ungrammatical and potentially confusing:

50.15 *ahm okay, I have a list of things to talk about here. Tell me, is Port Kelang not a big, it's a small city, if you go to KL for example, that's much bigger.*
50.26 *right, in Kelang is there many Malay or a lot of Chinese or what is it in Kelang?*
46.21 *how do you actually when speaking to the teacher how do you?*

Interviewer 6 (the second most difficult) also appears to create difficulty through the syntactic complexity of her questions. Her turns, as was noted earlier, tend to be much longer than those of the other interviewers. This seems to be a consequence of a large percentage of her questions consisting of multiple formulations, any of which might be incomplete, resulting in potential confusion for candidates.

57.48 *Now if you could have a career path, we are talking about after you finish your study here, if you could choose a career path that led anyway you wanted, what would you choose to do with your career, if you could work anywhere you wanted, do anything you*

32.17 I mean especially for an Australian to go to Japan, especially to Tokyo. Is there any way that I can overcome that, is there some way that I can live in Tokyo and be able to afford it? Do you have any advice?

57.10 Okay, can you tell me a little bit about perhaps work, I know you probably don't work in Malaysia, you look probably a bit too, you are obviously a student still, um but you probably know about work in Malaysia, generally what's what are the conditions like. Do people, you know do they work long hours? is the pay good?

Another noteworthy aspect of interviewer 6's behaviour is that she frequently interrupts the candidates with another question before they have completed what they want to say in their previous response.

In contrast, the easier interviewers (4 and 2) consistently use economical, complete and grammatically correct questions. While the amount of backchannelling (ie. mm, right, oh, aha, etc.) taking place while the candidate is still talking does not appear to distinguish easy and more difficult interviewers, feedback at the beginning of a next turn or as a stand-alone turn is a characteristic of the two easier interviewers. This could be read as both acceptance of the previous answer and encouragement to elaborate, in other words a positive evaluation of the candidate's contribution, possibly contributing to increased confidence on the candidate's part or, alternatively, presenting to the raters a sense that the candidate is able to participate adequately in an interaction with a native speaker.

7.0 Conclusions

In this study we set out to investigate firstly whether different interviewers could be said to present significantly different hurdles for candidates, and secondly what features of interviewer behaviour might contribute to this. Through a research design using multiple interviews and ratings, analysed using multi-faceted Rasch, we were able to demonstrate that there are indeed significant differences. Of six randomly selected interviewers, one was significantly easier than all but the second easiest, and the second easiest was significantly easier than the most difficult. In other words there is no doubt that candidates can be disadvantaged or advantaged by 'the luck of the draw' in interviewer allocation.

An initial analysis of interviewer styles showed some differences. While it is not possible from this limited study to draw any firm conclusions about which interviewer behaviours could be said to contribute to difficulty, certain tendencies were identified here which warrant further investigation.

In particular, the easier interviewers tended to shift topic more frequently, with fewer turns per topic, asked more questions of a simpler nature and spent considerably longer in Phase 2 than in Phase 4. Furthermore, it seems that the more structured the interview is as a straightforward question-and-answer routine, the easier it appears to be (or the more competent the candidate appears). Those interviewers identified as the most difficult in this study were, in fact, more likely to engage in more 'natural' conversational techniques such as interruption and disagreement. They were more likely to produce sentence fragments or complex ungrammatical utterances. Moreover, they were also more likely to push the candidate into a range of harder linguistic behaviours including speculating and justifying opinions.

For IELTS then, as for any other oral interview, the challenge is to decide what behaviour is appropriate and to ensure that it occurs. Is the aim to replicate authentic interaction (which would imply a lack of simplification and accommodation) or simply to elicit information

(which would imply limiting the interview to a question/answer format and making allowances for weaker candidates)? There appear to be two types of interviewer, one (the most difficult) which makes fewer allowances and provides less support, uses more complex language, and pushes the candidates into more complex interactional skills such as speculation and justification, and the other which uses simple language and more straightforward questions and which provides more support and feedback. These findings support those of Lumley and Brown (forthcoming), where two types of interviewer were identified, those who took on the role prescribed in the role play and acted it out in the spirit intended, and those who exhibited more 'teacher-like', or supportive, behaviour. Whatever the intention of the test developers, interviewers need to be trained accordingly as to what is and is not suitable behaviour. This could include monitoring of their own performance, discussion of how they should deal with particular situations (for example where they do not feel the candidate will cope with the speculative phase), even comparison of various interview techniques and behaviours - all these, while naturally contributing to additional expense in training, are necessary to ensure equivalence across interviews and interviewers, and hence fairness to candidates. It is after all, only as much as is done in the training of raters. Why should interviewer training warrant less attention? The findings of studies such as this demonstrate that interviewer talk is not neutral and indicate that the time is ripe to re-evaluate the emphasis we place on training for oral tests.

Bibliography

Lazaraton, A. (1993) 'A qualitative approach to monitoring examiner conduct in the Cambridge Assessment of Spoken English (CASE)'. Paper presented at 15th Language Testing Research Colloquium. Cambridge, England.

Lazaraton, A. and Saville, N. (1994) 'Processes and outcomes in oral assessment'. Paper presented at 16th Language Testing Research Colloquium. Washington DC.

Linacre, J.M. (1989) *Many-Facet Rasch Measurement.* Chicago: Mesa Press.

Lumley, T. and McNamara, T. (1993) 'The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings'. Paper presented at 15th Language Testing Research Colloquium. Cambridge, England (ERIC Document Reproduction Service ED 364 066).

Lumley, T and Brown, A. (1996) 'Specific-purpose language performance tests: task and interaction'. In G. Wigglesworth and C. Elder (eds). *The Testing Cycle: New Perspectives, Australian Review Of Applied Linguistics Series S.*

Ross, S. and Berwick, R. (1990) 'The discourse of accommodation in oral proficiency interviews'. *Studies in Second Language Acquisition,* 14: 159-176.

Ross, S. (1992) 'Accommodative questions in oral proficiency interviews'. *Language Testing,* 9, 2: 173-186.