1 Interviewer style and candidate performance in the IELTS oral interview

Annie Brown and Kathryn Hill

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.

This study aims first 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 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 variability of rater behaviour or even of task are. 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. 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 understanding to what is a growing area of concern among language testers.

Oral interviews, such as those 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 can-

didate 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 first, by identifying whether interviewer style does in fact have an effect on scores, and second 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 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 (Lazaraton and Saville 1994, Ross 1992, Ross and Berwick 1990), 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 1996).

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 (1996) 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, i.e. 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:

1. Using multifaceted Rasch analysis, determine whether different interviewers represent different 'hurdles' in terms of the difficulty of doing an IELTS interview.
2. Identify cases where candidates perform differentially in each of the two interviews they undertake.
3. Transcribe and analyse these interviews in order to identify whether there are particular interviewing styles which characterise 'easy' or 'difficult' interviewers and which may contribute to better or worse performance by candidates.

3 The IELTS interview and rating

The 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 band scales with 10 levels (0–9). (IELTS Handbook 1997, Cambridge: UCLES.)

4 Methodology

Thirty-two students from IELTS preparation courses and six accredited interviewers participated in this study. Each of the 32 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 by four 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. After the interviews had been completed, they were informed 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, i.e. 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 by accredited IELTS raters.[2] In order to take rater harshness into account (i.e. 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.

The analysis was done in two stages:

(a) The multifaceted 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.

5 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 1.1.

The interviewer difficulty measures are presented in logits, the units of measurement used within Rasch analysis (see Appendix 1.2). 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 1.1 confirms that there are significant differences amongst this group of interviewers in terms of their

Table 1.1 Interviewer difficulty

	Interviewer ID	Interviewer difficulty (logits)	Model SE	Model fit Infit MnSq	Infit Std	Outfit MnSq	Outfit 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

difficulty: the interviewer separation index indicates 1.34 statistically distinct interviewer strata,[5] separated with a reliability of 0.64. This means that the probability that the differences between interviewers are due to chance is low. There is a greater possibility that the differences are significant. 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). This means that the chances that the interviewers are equally severe are very low (0.00 probability), although this likelihood is slightly lessened by the fact that 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 also likely to be a consequence of the small sample size.

Turning to the fit of the interviewers to the model, as shown in Table 1.1, 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). The highest is interviewer 2, one of the easier interviewers, at 1 and the lowest and most severe interviewer 5 at – 2.

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?

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

To take an example, the difference between the difficulty measures of Interviewer 5 (the most difficult) and Interviewer 4 (the easiest) is 1.61 logits. The square root of the sum of the squared standard errors of these two difficulty measures is 0.97. Therefore, as 1.61 is greater than 0.97, the two interviewers can be considered to be significantly different in difficulty.

The result of this calculation is presented in Table 1.2. Here, 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').

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.

Table 1.2 Paired differences in interviewers

Pairs of Interviewers	Difference in Difficulty (d1–d2) (logits)	✓($se^2 + se^2$)	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	–

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: 0.00.

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, i.e. 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.

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 first, cases where candidates received a different score on each occasion, and second, 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 1.3). 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 (i.e. the candidate got a better score with the more difficult interviewer).

Six pairs of interviews, highlighted in Table 1.3, 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).

Table 1.3 Interview pairs: score differences

Candidate	Occasion 1 fair average	Interviewer	Occasion 2 fair average	Occasion 2 Adjusted for difficulty	Interviewer	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	✓
25	5.9	6	6.9	7.1	2	.8	✓
02	6.8	1	6.2	6.4	4	.4	✗
21	6.8	4	6.4	6.6	5	.2	✓
24	6.6	6	5.9	6.1	2	.5	✗
06	6.5	2	5.4	5.6	6	.9	✓
37	6.3	3	6.6	6.8	4	.5	✓
14	6.3	3	6.2	6.4	4	.1	✓
01	5.9	3	6.1	6.3	4	.4	✓
18	5.9	4	4.9	5.1	5	.8	✓
16	5.8	4	5.0	5.2	5	.6	✓
15	5.4	3	6.2	6.4	4	1.0	✓
38	5.2	2	5.0	5.2	3	–	–
19	4.3	5	4.3	4.5	3	.2	✓