The impact of gender in oral proficiency testing

Kieran O'Loughlin *The University of Melbourne*

To date, the role of gender in speaking tests has received limited attention in language testing research. It is possible in oral interviews, for instance, that both interviewing and rating may be highly gendered processes. In tests like the IELTS interview, where the interviewer also acts as the rater, this poses the question of whether a gender effect, if it exists, stems from the interview itself, the rating decision or a combination of both these 'events'. The data collected for this study consisted of the audio-taped performances of 8 female and 8 male test-takers who undertook a practice IELTS interview on two different occasions, once with a female interviewer and once with a male interviewer. The interviews were transcribed and analysed in relation to previously identified features of gendered language use, namely overlaps, interruptions and minimal responses. The scores later assigned by 4 raters (2 males and 2 females) to each of the 32 interviews were also examined in relation to the gender of both raters and test-takers using multi-faceted Rasch bias analyses. The results from both the discourse and test score analyses indicated that gender did not have a significant impact on the IELTS interview. These findings are interpreted in relation to more recent thinking about gender in language use.

I Introduction

Recent research into oral language interviews has indicated that interviewers vary considerably from each other in relation to their test behaviour. Such variability includes the amount of support they give to candidates, the amount of rapport they establish with candidates and the extent to which they follow the instructions relevant to their role (e.g., Young and Milanovic, 1992; Brown and Hill, 1996; Lazaraton, 1996; McNamara and Lumley, 1997; Morton *et al*., 1997).

A key issue arising from such findings is why interviewers vary from each other. One possibility is that such variability stems, at least partly, from gendered differences in communicative style. There is a large body of research in the field of language and gender (see, for example, Maltz and Borker, 1982; Tannen, 1990; Coates, 1993; Thwaite, 1993) which suggests that male and female conversational

Address for correspondence: Kieran O'Loughlin, Department of LLAE, Doug McDonell Building, The University of Melbourne, Victoria, Australia 3010; email: kjo@unimelb.edu.au

Language Testing 2002 19 (2) 169–192 10.1191/0265532202lt226oa

styles are quite distinct. These studies broadly characterize the female conversational style as collaborative, co-operative, symmetrical and supportive whereas its male equivalent is portrayed as controlling, unco-operative, asymmetrical and unsupportive.

In her book *Women, men and language* Jennifer Coates (1993: 140), for instance, argues that women and men seem to differ in terms of their communicative competence in so far as they 'have different sets of norms for conversational interaction'. Therefore, she concludes 'women and men may constitute distinct speech communities'. Such claims may have serious implications for language testing since they imply that the construct of communicative competence is not gender neutral. Is it reasonable, for instance, to assess female and male speakers against the same set of norms? Equally, we might ask, is it fair for test-takers, especially females, to be interviewed and rated by members of the opposite gender if they belong to different speech communities? On the other hand, it could be argued that a language test need not reflect all aspects of 'real-life' communication (including gendered differences) in order to still be valid.

More recently, the research which has found clear gender differences in spoken interaction has been strongly criticized for its tendency to overgeneralize its findings to all men and all women irrespective of other social identity factors (such as their age, ethnicity, occupation and sexual identity) and situational factors such as the communicative context and the gender of their interlocutors. In recent studies men and women in fact show themselves capable of using a range of conversational styles in different speech contexts. Where men and women exhibit similar conversational behaviour it may be that other aspects of their social identity which override potential gender differences are brought into play. In other words, instead of being fixed, polarized and predictable, the language use of men and women is now seen as varying across cultural, social and situational contexts, sometimes exhibiting difference and other times similarity (see, for example, Freed, 1995; Freed and Greenwood, 1996; Freeman and McElhinny, 1996; Stubbe, 1998).

Notwithstanding such critiques of fully predictable and inevitable gendered differences in spoken interaction, the potential for such differences is clearly an important issue in the testing context. In the interests of test fairness systematic investigations into whether clearly distinct styles are consistently evident for male and female interviewers, for instance, need to be carried out together with what effects such gendered differences (if they are found to exist) have on candidate performance.

As Sunderland (1995) suggests differences in male and female interviewer styles *per se* can be viewed as one potential gender effect. Another possibility she identifies is that the behaviour of interviewers of either gender may vary according to whether they are paired with a male or female candidate. In both cases, it is feasible that the gendered behaviour of the interviewer will influence the outcome of the test by either strengthening or undermining the candidate's performance.

A further gender consideration in oral test interviews is that candidates' output may vary according to their own gender. As suggested above, the quantity and quality of their output may also be affected positively or negatively by the gender of the interviewer.

Finally, it is also worth considering whether there could be a gender effect in the rating of oral interviews. It is possible in oral interviews that male and female raters may assess differently. It is also possible that their assessments are influenced by the gender of the candidate. In the case of tests like the IELTS interview where the interviewer also acts as the rater, this poses the question of whether a gender effect, where it exists, stems from the interview itself, the rating decision or a combination of both these 'events'.

There have been a number of recent studies which have examined the possibility of a gender effect in the rating of candidates by their interviewers in oral interviews. Most of this research reveals some kind of gender effect on test scores although, interestingly, the effect is not always the same. Some studies report that test-takers scored more highly with male interviewers (e.g., Locke, 1984; Porter 1991a; 1991b) while others report higher scores with female interviewers (e.g., Porter and Shen, 1991; O'Sullivan, 2000). An interaction effect between the gender of the interviewer and interviewee has also been reported (Buckingham, 1997). In this case candidates achieved a higher scores when paired with an interviewer of the same gender. By virtue of their very inconsistency these findings appear to support more recent thinking about the shifting, unstable nature of gender in spoken interaction to which I have just alluded and to which I return at the conclusion of this article.

The study reported in this article addressed the following questions:

1) What impact does the gender of test-takers and interviewers have on the discourse produced in the IELTS oral interview?
2) What impact does the gender of test-takers and raters have on the rating of the IELTS oral interview?
3) If a gender effect is found in the course of interviewing and/or rating, how might its impact on test scores be managed?

II Methodology

1 The IELTS oral interview

The IELTS is a four-skill test employed in the selection of prospective students whose first language is not English to universities in such countries as Australia, Canada and the UK. The version of the Speaking sub-test used at the time this study was conducted (July 1998) lasted between 10 and 15 minutes. It was described by UCLES (1998: 11) as 'an oral interview, a conversation, between the candidate and an examiner' and consisted of five phases as outlined below:

- Phase 1 Introduction: The examiner and candidate introduce themselves. The candidate is made to feel comfortable and encouraged to talk briefly about their life, home, work and interests.
- Phase 2 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.
- Phase 3 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.
- Phase 4 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.
- Phase 5 Conclusion: The interview is concluded.

Interviewers also carried out the assessment of the candidate's proficiency using a global band scale with nine increments. The assessment took into account evidence of communicative strategies, and appropriate and flexible use of grammar and vocabulary. Note that the format of the IELTS Speaking sub-test has changed as of 1 July 2001 (UCLES, 2000: 15).

2 Interview design

Sixteen different students (8 male and 8 female) and 8 accredited IELTS interviewers (4 male and 4 female) participated in this stage of the study. Each of the candidates were interviewed on 2 different occasions by a male and a female interviewer yielding a total of 32 interviews.

The candidates were international students engaged in an IELTS preparation course with the aim of undertaking further study in Australia. Consequently, they volunteered for this project on the basis of experiencing the interview under exam-like conditions, gauging their readiness for the test and receiving feedback from the interviewers about possible areas to develop in their preparation for the official test. The candidates came from a range of language and cultural backgrounds. The women came from China, Indonesia, Japan and Thailand and ranged in age from 19–31. The men came from China, Indonesia, Japan, Korea and Thailand and ranged in age from 20–30. While female and male participants came from similar backgrounds, a possible limitation of the research is that candidates from other language and cultural backgrounds did not take part in the study.

The interviewers were all fully trained, current IELTS examiners ranging in age, workplace and length of time as an examiner. Both interviewers and candidates were not given any indication of the focus of the project beyond it being a study of the discourse and scores produced in the Speaking sub-test. As indicated above the interviewers were each asked to provide feedback to the candidates about their strengths and weaknesses in preparation for the official test. This was done immediately after each interview. All the students were interviewed twice on the same day, once by a female interviewer and once by a male interviewer. Half of the students were interviewed by a male interviewer first and the other half by a female interviewer first. The interviews were conducted at the same site on two consecutive days. Candidates were not exposed to the same topics in the two interviews so as to minimize any potential practice effect. Each of the interviews were audiotaped, as they are in the official IELTS Speaking sub-test. Of course, this restricted the discourse analysis of the interviews which followed to verbal behaviour only.

3 Rating design

Each candidate was subsequently assessed by 2 female and 2 male accredited IELTS interviewers who represented a range of ages, workplaces and experience as IELTS examiners, using the audio-recordings of the interviews. Like the interviewers and candidates, the raters were not given any indication of the focus of the project beyond it being a study of the discourse and scores produced in the Speaking sub-test. A mixed design was used for these additional ratings whereby each interview was assessed by different combinations of male and female raters drawn from a pool of 8 females and 8 males with each rater carrying out a total of 8 assessments. Although

it is the interviewer who normally carries out the assessment of the candidate, this strategy enabled multiple ratings of the same interview to be collected and raters to be calibrated against each other in the statistical analyses which followed.

Table 1 provides a detailed overview of the interviewing and rating design of the study in which there were 16 test candidates (numbered 1 to 16 in the table), 8 interviewers (numbered 1 to 8) and 16 raters (numbered 1 to 16).

III Findings

In order to address the research questions, analyses of both the discourse of the interviews and the test scores allocated to candidates were conducted.

Table 1 Interview and rating design

Candidate	Candidate gender	Interviewer	Interviewer gender	Female raters	Male raters
1	M	1	M	5, 6	9, 10
1	M	2	F	1, 2	11, 12
2	M	1	M	3, 4	13, 14
2	M	2	F	5, 6	15, 16
3	F	1	M	7, 8	10, 12
3	F	2	F	2, 4	9, 11
4	F	1	M	1, 3	14, 16
4	F	2	F	6, 8	13, 15
5	M	3	M	5, 7	9, 12
5	M	4	F	1, 4	10, 11
6	M	3	M	2, 3	13, 16
6	M	4	F	5, 8	14, 15
7	F	3	M	6, 7	12, 16
7	F	4	F	4, 8	9, 13
8	F	3	M	1, 5	10, 14
8	F	4	F	2, 6	11, 15
9	M	5	M	3, 7	9, 14
9	M	6	F	1, 6	10, 13
10	M	5	M	2, 5	11, 16
10	M	6	F	3, 8	12, 15
11	F	5	M	4, 7	10, 16
11	F	6	F	2, 8	9, 15
12	F	5	M	1, 7	12, 14
12	F	6	F	4, 6	11, 13
13	M	7	M	3, 5	9, 16
13	M	8	F	1, 8	10, 15
14	M	7	M	2, 7	11, 14
14	M	8	F	3, 8	12, 13
15	F	7	M	4, 5	11, 16
15	F	8	F	3, 4	9, 10
16	F	7	M	1, 2	15, 16
16	F	8	F	7, 8	13, 14

1 Interviews

Before transcribing the interviews it was decided to focus on participants' use of three features, namely overlaps, interruptions and minimal responses. These features were chosen on the basis that they seem to be highly 'gendered' in spoken interaction according to research reviewed by Coates (1993). Although it could be argued that pre-selecting these categories meant that the analysis ignored other ways in which gender may have been accomplished in the interviews, this strategy did allow direct comparisons to be made with Coates' (1993) specific claims about the use of these conversational features.

The audio-recordings of all 32 interviews in this study were transcribed by a research assistant using a broad notation system adapted from Tannen (1984). The transcription notation is shown in Appendix 1. This coding system was considered adequate for the purpose of capturing the discourse features chosen for the study. The coding was later checked by the researcher. The transcripts were then analysed in relation to the use of these three features by the different gender pairs. The frequencies of each discourse feature in all 32 interviews were independently calculated by both the research assistant and the researcher to ensure a high degree of reliability. Where their figures differed the interviews were re-examined until consensus was achieved. Although such quantification of language data is not without its problems (see, for example, Schegloff, 1993), it did allow overall comparisons to be made between the different gender pairings (female-female, female-male, male-female and male-male) for each of the discourse categories.

a Overlaps: Coates (1993: 109) defines overlaps as 'instances of slight over-anticipation by the next speaker: instead of beginning to speak immediately following current speaker's turn, the next speaker begins to speak at the very end of current speaker's turn, overlapping the last word (or part of it)'. Based on previous research Coates (1993) suggests that overlaps are likely to be equally distributed between participants in same-sex conversations, but that in mixed-sex conversations they are much more likely to be caused by male speakers. Coates' explanation for this finding is that women are concerned that the man does not feel that his turn is being violated and so wait until he has finished speaking.

Most of the overlaps used by interviewers (interviewer overlaps) seemed to be offering support for the person whose turn it was, both by confirming information and continuing the topic. The following example shows the interviewer confirming the candidate's idea that 'many people want to see this game'. As the candidate reformulated this utterance, the interviewer perhaps recognized the candidate's

need for support in this idea and thus joined in to confirm it. Note that in this example (and those which follow) the symbol • indicates the focus of analysis. Also, the identities of the interviewer and candidate are abbreviated below the example. Here 'I.3 (male)' refers to Interviewer 3 who is male, and 'C.5 (male)' refers to candidate 5 who is also male.

1) *C*: So many people want to see this game.
I: Sure.
C: Looking forward to [see this game.]
• *I*: [lots of students] want to go and see this game.

I.3 (male) / C.6 (male)

There were only a few overlaps observed which seemed to involve an attempt to close down the topic of discussion. In the following example the interviewer attempts to introduce the idea of unemployment. First, she refers to people losing jobs and then, when the student continues by talking about government protection of industry, the interviewer overlaps with a question about the unemployment rate. Following the overlap she then reiterates the question, thus reinforcing the direction she wants the discussion to go:

2) *C*: Um (0.5) because I want to I think it's OK to trade with (.) to trade with another country. Because I think labour in Thailand have trend have trend to ah expensive in the future, yeah. so if if we use (0.5) if we use ah: (.) not no (.) we can ah we can import something from the other countries?
I: Mm,
C: which will cheaper than in my country in the future=
I: =but will that help your country develop? If if people lose jobs? Because you traded from outside?
C: Um I think it doesn't matter because my government will have a policy to protect (.) um (.) some industrial in Thailand. Yes.
Same [Australia (.) in here,]
• *I*: [what's the unemployment] what's the unemployment rate in Thailand.
C: Unemployed?

I.4 (female) / C.5 (male)

The total number of interviewer overlaps across all 32 interviews was 79. Table 2 shows a breakdown of the results for each gender

Table 2 Interviewer overlaps: range of overlaps (total number in parentheses)

	Interviewers	
Candidates	Female	Male
Female	0–11 (23)	0–9 (24)
Male	2–6 (25)	0–3 (7)

pairing. Each cell shows, first, the range of overlaps (i.e., the minimum and maximum number of overlaps) and, secondly, in brackets, the total number of overlaps for that gender pairing across eight interviews. It is evident that the total number of overlaps used by male interviewers with male candidates is clearly lower than the other three gender pairings overall. It was not considered prudent to use a procedure such as the chi-squared statistic to determine whether the differences were statistically significant here (and in the tables which follow) because of the broad range of observed frequencies in some of the cells.

All of the observed candidate overlaps appeared to play a facilitative role in the interactions. In the following example the candidate (female) is supporting the interviewer's idea of shopping at the local supermarket and continues this by offering examples of names of supermarkets:

3) *I*: And and the supermarket um (0.5) the local supermarket a good idea to buy [buy food?]
- *C*: [Mm: like Coles?] Target,
I: Coles Target Safeway,

I.4 (female) / C.8 (female)

The total number of candidate overlaps was 77. The breakdown of figures in Table 3 indicates that male candidates used less overlaps

Table 3 Candidate overlaps: range of overlaps (total number in parentheses)

	Candidates	
Interviewers	Female	Male
Female	0–10 (20)	0–3 (11)
Male	0–17 (33)	0–4 (13)

than female candidates with both female and male interviewers. Of note was an interview where a female candidate used 17 overlaps with her male interviewer. Overall, the total numbers of overlaps for the different gender pairings in these tables do not reveal a clear gendered pattern of use. The range figures also indicate that there was a fairly high degree of variability in the use of overlaps within each gender pairing.

b Interruptions: Coates (1993: 109) states that:

> Interruptions on the other hand are violations of the turn-taking rules of conversation. The next speaker begins to speak while the current speaker is still speaking, at a point in the current speaker's turn which could not be defined as the last word. Interruptions break the symmetry of the conversational model; the interrupter prevents the speaker from finishing their turn, at the same time gaining a turn for themselves.

Based on previous research, Coates (1993) suggests that same-sex interlocutors are unlikely to interrupt each other. On the other hand, men frequently interrupt women in mixed-sex conversation while the reverse rarely occurs.

There were very few instances of interruptions by interviewers (interviewer interruptions) in these interviews. In the following example we see the interviewer intervening to take up and develop the first response given by the candidate:

4) *C*: Yeah. Firstly I would like to improve my English because I think it's important for me to (.) to study English [and also,]
- *I*: [Why?] Why do you think you need English?

I.1 (male) / C.2 (male)

Overall, there were only 7 instances of interviewer interruptions. Table 4 shows their distribution. Because interruptions were used by interviewers so infrequently there is no discernible gendered pattern of use.

Table 4 Interviewer interruptions: range of interruptions (total number in parentheses)

	Interviewers	
Candidates	Female	Male
Female	0–1 (1)	0–0 (0)
Male	0–1 (2)	0–4 (4)

Again, there were only a few instances of candidate interruptions in the data. In the example below the candidate continues the topic of Melbourne's weather referred to in the interviewer's previous turn:

5) *I*: Because you know what Melbourne's like? Huh?=
C: =Yeah. (laughs)
I: Always very unpredictable? Um so just listen carefully to the forecast,=
C: =Mm hm,=
I: =And then take the right stuff for this [kind of weather.]
• *C*: [Because sometimes] we can't believe them.
I: That's right.

I.3 (male) / C.8 (female)

There were 17 instances of candidate interruptions across the 32 interviews. Table 5 shows their distribution. Again, because they occurred so infrequently, there is no discernible gendered pattern of use.

c Minimal responses: Coates (1993: 109) describes minimal responses (MRs) such as *yeah* and *mhm* as not constituting a turn. Instead, 'they are a way of indicating the listener's positive attention to the speaker, and thus a way of supporting the speaker in their choice of topic' (Coates 1993: 109). Once again, based on previous studies of spoken interaction, Coates (1993) claims that women use minimal responses more than men and at more appropriate moments.

In the IELTS interview data analysed here, minimal responses appeared to serve a consistently supportive function in keeping with Coates' definition. That is to say, they encouraged the interlocutor to continue speaking by providing a signal to show active listening.

The example here shows a female interviewer using MRs (interviewer minimal responses) to encourage the candidate to continue the idea she is trying to express:

Table 5 Candidate interruptions: range of interruptions (total number in parentheses)

	Candidates	
Interviewers	Female	Male
Female	0–2 (3)	0–3 (5)
Male	0–4 (7)	0–2 (2)

6) *C*: Ah I'm marketing supervisor,
- *I*: Mm hm,
  C: Also I still work hard. Everyday busy,
- *I*: Mm hm,
  C: Because I have a analyst analyst team? And ah (.) investigate (?) marketing information and I should do I should start I should do project and ah supervise the project how the progress,
- *I*: Mm:,
  C: And how affect in this project,
- *I*: Mm:,
  C: And I feel stress and too busy and ah no too much time for holiday yeah so I cracked,
  I: Not too much free time.

I.6 (female) / C.11 (female)

There were many more instances of this feature throughout the 32 interviews than there were of either overlaps or interruptions. Interviewer MRs totalled 805. Looking at the distribution of their use in Table 6 it is noticeable that male interviewers used slightly more minimal responses than their female counterparts with candidates of both genders, a finding which conflicts with Coates' (1993) claim that women use more minimal responses. In one instance a male interviewer used 66 MRs in an interview with a female candidate.

In this example the candidate uses MRs (candidate minimal responses) to show that she is listening and to provide support for the information being given by the interviewer:

7) *I*: Well Japanese is usually expensive.
   C: Yeah I think so,
   I: Um there is a nice Japanese restaurant in the city,
- *C*: Yeah,
  I: At the top of mm: (1.0) off the main at the top of Bourke Street you might know it.

Table 6 Interviewer minimal responses: range of responses (total number in parentheses)

	Interviewers	
Candidates	Female	Male
Female	1–50 (169)	8–66 (233)
Male	11–36 (199)	13–45 (204)

C: No I don't know it,
I: Anyway that's about,

- *C*: Yeah,
  I: That's one of the you know for value,
- *C*: Oh:?
  I: That's probably the cheapest Japanese restaurant?
- *C*: Yeah,
  I: But if you want Japanese you have to pay in Australia.

I.8 (female) / C.15 (female)

Candidates' minimal responses totalled 301. The fact that this figure is much lower than the equivalent one for the interviewers (805) is perhaps not surprising given the respective roles of the two groups of speakers in this context, i.e., the interviewer's role is to facilitate and support the candidate's test performance. As such, the oral interview may differ from casual conversations where participants tend to have more equal rights and responsibilities.

Table 7 indicates that candidates being interviewed by a person of the same gender used more minimal responses overall than the mixed gender pairings. This finding is again at odds with Coates' claim that women always use more minimal responses. O'Sullivan (2000) suggests that minimal responses may be used in a more supportive way by female interviewers than male interviewers in speaking tests. However, the data examined in this study did not indicate such a qualitative gendered difference for either interviewers or candidates.

In general, as was the case for the use of overlaps, there was no clear gender pattern in the use of minimal responses for either interviewers or candidates. The range figures also show that there was a very high degree of variability in the use of minimal responses within each gender pairing for the two groups of speakers. These results are further discussed later in this article.

Having reported the results of the discourse analysis, we now turn to the analyses of the test scores.

Table 7 Candidate minimal responses: range of responses (total number in parentheses)

	Candidates	
Interviewers	Female	Male
Female	6–25 (125)	4–16 (57)
Male	1–10 (48)	1–32 (71)

2 Test scores

The focus of the analysis which follows is on the scores of the raters who later assessed the audio-recordings of the interviews since comparison could then be made between multiple scoring of each interview carried out under the same conditions.

The raw scores assigned by these raters were analysed using multi-faceted Rasch measurement (one-parameter, rating scale model). The computer program FACETS (Linacre and Wright, 1992) was employed to calculate ability estimates for each of the 16 candidates based on their scores in each of the two interviews. The unit of measurement of these ability estimates is the logit. A feature of the FACETS program is that it can compensate candidates for differences in severity between raters and other facets of the test situation (such as item difficulty) in calculating these ability estimates, thereby enhancing their validity and reliability. The facets included in the first analysis were candidate and rater. Table 8 shows these ability estimates together with their standard error (s.e.) and infit mean square values. In this table candidates are ordered in terms of their ability from highest to lowest, i.e., the higher the logit value the more able the candidate. Thus, Candidate 3, with the highest logit score, is the most able and Candidate 9, with the lowest logit score, the least able in this set of results.

Table 8 Candidate measurement report

Candidate ID	Candidate gender	Logit	s.e.	Infit mean square
3	Female	4.95	0.67	1.1
2	Male	2.46	0.60	0.9
7	Female	1.35	0.61	0.3
14	Male	1.15	0.62	0.6
16	Female	1.09	0.63	1.6
10	Male	0.97	0.63	0.9
15	Female	0.79	0.61	1.6
8	Female	–0.01	0.65	0.8
4	Female	–0.21	0.67	0.3
13	Male	–0.26	0.64	1.6
1	Male	–1.00	0.67	1.4
6	Male	–1.21	0.73	0.7
5	Male	–2.38	0.72	0.6
11	Female	–2.46	0.70	2.2
12	Female	–3.84	0.68	0.5
9	Male	–4.41	0.67	0.7
Mean		–0.19	0.66	1.0
s.d.		2.30	0.04	0.5

Table 9 Rater measurement report

Rater ID	Rater gender	Logit	s.e.	Infit mean square
11	Male	1.73	0.70	0.3
8	Female	1.71	0.70	0.8
2	Female	1.16	0.69	0.7
13	Male	1.11	0.67	0.4
7	Female	1.11	0.65	0.5
14	Male	0.50	0.66	0.2
16	Male	0.41	0.65	2.2
4	Female	0.15	0.65	1.1
5	Female	–0.32	0.65	0.4
6	Female	–0.66	0.62	0.8
3	Female	–0.67	0.62	1.1
12	Male	–0.72	0.64	2.0
15	Male	–0.99	0.63	0.8
9	Male	–1.00	0.66	1.3
10	Male	–1.63	0.65	1.2
1	Female	–1.90	0.64	1.8
Mean		0.00	0.65	1.0
s.d.		1.12	0.02	0.6

The program also routinely provides figures for each of the facets incorporated into the analysis. The other facet included here was rater severity. Table 9 shows the relevant logit scores together with their standard error and infit mean square values. In this table, raters are ordered from highest to lowest in terms of their severity. Thus, the most severe rater (Rater 11) has the highest logit score while the most lenient rater (Rater 1) has the lowest logit score.

In order to obtain an overview of these results in relation to gender a second FACETS analysis was then undertaken, this time incorporating the two facets candidate gender and rater gender. In Table 10 the higher logit score for men indicates that the test was more difficult for them than for the women in the study overall. This is consistent

Table 10 Candidate gender measurement report

Candidate gender	Logit	s.e.	Infit mean square
Male	0.12	0.14	0.9
Female	–0.12	0.14	1.1
Mean	0.00	0.14	1.0
s.d.	0.12	0.00	0.1

Table 11 Rater gender measurement report

Rater gender	Logit	s.e.	Infit mean square
Female	0.02	0.14	0.8
Male	–0.02	0.14	1.2
Mean	0.00	0.14	1.0
s.d.	0.02	0.00	0.2

with Table 8, which showed that 5 of the top 8 candidates were female. Table 11 shows the overall severity of raters in terms of gender. Here the higher logit value for female raters indicates that, as a group, they were slightly harsher than male raters.

Before moving on to the impact of gender on the rating of the interviews it is worth rounding off the investigation of the effect of gender on the discourse produced in the interviews by matching the findings of those analyses up with the ranking of candidates (using their ID numbers) based on their ability estimates. The issue here is whether the use of overlaps, interruptions and minimal responses varied according to both the gender and the ability of the candidate. For example, was there greater or lesser use of any of these features in interviews with more able female candidates? Table 12 provides, first, a list of candidates ranked from highest to lowest proficiency on the basis of their ability estimates (see Table 8). The table also shows their gender and the ID number and gender of their interviewer (Int) in each interview. The number of overlaps, interruptions and minimal responses used by both the candidate and the interviewer for each interview are then detailed. Overall, the figures in Table 12 do not appear to support the existence of an interaction effect between the gender and ability of candidates in relation to the use of the three discourse features examined in the study.

We move now to the second research question, that is, what impact does the gender of test-takers and raters have on test scores. In order to address this issue the scores of the raters who assessed the audio-recordings of the interviews were examined using an extension of the computer program FACETS known as bias analysis. Bias analysis in multi-faceted Rasch measurement identifies unexpected but consistent patterns of behaviour which may occur from an interaction of a particular rater or group of raters with respect to some component or 'facet' of the rating situation. Bias analysis was therefore used in this study to investigate the impact of candidate and rater gender on test scores.

Table 12 Breakdown of discourse analysis figures

Rank	Cand ID	Cand Gen	Int ID	Int Gen	Int OLs	Cand OLs	Int IRs	Cand IRs	Int MRs	Cand MRs
1	3	F	1	M	3	8	–	1	24	1
	3	F	2	F	11	–	–	–	19	13
2	2	M	1	M	1	–	4	–	14	2
	2	M	2	F	2	–	1	–	12	16
3	7	F	3	M	5	5	–	2	44	4
	7	F	4	F	1	2	1	2	8	8
4	14	M	7	M	1	3	–	–	38	4
	14	M	8	F	6	2	–	3	72	8
5	16	F	7	M	1	–	–	–	66	9
	16	F	8	F	2	1	–	–	50	25
6	10	M	5	M	–	–	–	–	13	32
	10	M	6	F	2	2	–	–	18	0
7	15	F	7	M	1	1	–	–	42	3
	15	F	8	F	–	1	–	1	23	22
8	8	F	3	M	3	17	–	4	28	12
	8	F	4	F	3	10	–	–	1	13
9	4	M	3	M	3	1	–	–	36	22
	4	M	4	F	2	1	–	1	15	11
10	13	M	7	M	1	–	–	–	40	2
	13	M	8	F	4	2	–	–	36	8
11	1	M	1	M	1	4	–	–	13	1
	1	M	2	F	–	1	–	–	24	7
12	6	F	1	M	9	2	–	–	10	1
	6	F	2	F	4	–	–	–	27	21
13	5	M	3	M	–	1	–	2	45	1
	5	M	4	F	7	0	1	1	11	3
14	11	F	5	M	2	–	–	–	11	8
	11	F	6	F	2	4	–	–	21	6
15	12	F	5	M	–	0	–	–	8	10
	12	F	6	F	–	2	–	–	20	17
16	9	M	5	M	–	4	–	–	5	7
	9	M	6	F	2	3	–	–	11	4

Notes: F = female; M = male; Cand = candidate; Int = interviewer; OLs = number of overlaps; IRs = interruptions.

Question 1: Do raters score candidates of one gender significantly more harshly than the other gender? The first issue to be examined here involves the interaction between raters' scores and candidate gender. The focus of the analysis is on the Z-score values as these figures provide a measure of rater bias. The FACETS program first calculates a bias logit which is then divided by its standard error to obtain the Z-score. Where the Z-score values fall between –2.0 and +2.0, the rater may be considered to be scoring candidates from the specified gender without significant bias. Where the Z-score value is greater than +2.0 the rater is scoring candidates of the specified gender significantly more harshly compared to the way that rater treats the other gender. On the other hand, where the Z-score value falls

Table 13 Bias calibration report, rater–candidate gender interaction

Rater ID	Candidate gender	Bias (logit)	Error	Z-score	Infit mn sq
1	F	0.34	0.59	0.6	0.4
1	M	–0.34	0.58	–0.6	1.6
2	F	–0.19	0.59	–0.3	0.4
2	M	0.23	0.64	0.4	0.0
3	F	–0.25	0.81	–0.3	1.3
3	M	0.09	0.49	0.2	0.8
4	F	0.33	0.50	0.7	0.3
4	M	–0.92	0.81	–1.1	1.3
5	F	0.25	0.84	0.3	0.4
5	M	–0.09	0.49	–0.2	0.4
6	F	0.54	0.62	0.9	1.0
6	M	0.51	0.58	–0.9	0.3
7	F	0.48	0.54	–0.9	1.5
7	M	0.93	0.76	1.2	0.4
8	F	0.28	0.54	–0.5	1.5
8	M	0.55	0.75	0.7	0.4
9	F	0.33	0.57	–0.6	0.6
9	M	0.37	0.62	0.6	0.3
10	F	0.82	0.58	–1.4	1.7
10	M	0.93	0.64	1.5	0.0
11	F	0.21	0.62	–0.3	1.8
11	M	0.25	0.65	0.4	0.3
12	F	0.49	0.57	–0.9	0.9
12	M	0.54	0.62	0.9	1.8
13	F	0.18	0.64	0.3	0.8
13	M	0.16	0.64	–0.2	0.8
14	F	0.38	0.64	0.6	0.8
14	M	0.36	0.62	–0.6	1.0
15	F	0.17	0.58	0.3	0.3
15	M	0.17	0.58	–0.3	0.9
16	F	0.17	0.65	1.8	2.0
16	M	0.02	0.57	–1.8	0.6

below –2.0, the rater is marking candidates from the specified gender significantly more leniently than the other gender. Table 13 presents the results of this analysis. Since all the Z-scores are within the range of −2 to +2 it can be concluded that none of the raters were significantly biased in favour of candidates of either gender.

Question 2: Do raters of one gender score candidates significantly more harshly than raters of the other gender? The second bias analysis examined whether there was a significant interaction between candidate scores and rater gender. Table 14 presents the results of this analysis. Again, since all of the Z-scores are within the range of −2 to +2, it can be concluded that none of the candidates were treated significantly more harshly by raters of either gender.

Table 14 Bias calibration report, candidate–rater gender interaction

Candidate ID	Rater gender	Bias (logit)	Error	Z-score	Infit mn sq
1	F	–0.71	0.79	–0.9	0.6
2	F	–0.05	0.74	–0.1	0.4
3	F	0.91	0.77	1.2	0.6
4	F	–0.37	0.79	–0.5	0.6
5	F	–0.43	0.90	–0.5	1.6
6	F	–0.03	0.90	0.0	1.6
7	F	–0.05	0.75	–0.1	0.4
8	F	0.62	0.85	0.7	0.5
9	F	0.32	0.82	0.4	0.5
10	F	0.25	0.79	0.3	0.6
11	F	–0.02	0.87	0.0	0.6
12	F	–0.36	0.83	–0.4	0.7
13	F	–0.63	0.75	–0.8	1.5
14	F	0.25	0.79	0.3	0.6
15	F	–0.05	0.75	–0.1	1.5
16	F	0.25	0.79	0.3	0.6
1	M	0.81	0.90	0.9	1.6
2	M	0.03	0.75	0.0	0.4
3	M	–1.10	0.88	–1.3	0.6
4	M	0.38	0.85	0.4	0.5
5	M	0.45	0.87	0.5	0.6
6	M	0.05	0.90	0.1	0.0
7	M	0.03	0.75	0.0	0.4
8	M	–0.55	0.75	0.7	1.5
9	M	–0.28	0.83	0.3	0.7
10	M	–0.25	0.75	0.3	1.5
11	M	0.06	0.87	0.1	3.6
12	M	0.40	0.82	0.5	0.5
13	M	0.70	0.85	0.8	0.5
14	M	–0.25	0.75	0.3	1.5
15	M	0.03	0.75	0.0	1.5
16	M	–0.25	0.75	–0.3	0.4

Question 3: Do raters score candidates of the same gender as their own significantly differently than they score candidates of the opposite gender? The third bias analysis examined whether there is a significant interaction between candidate gender and rater gender. Table 15 presents the results of this analysis. The results showed that the interaction between candidate gender and rater gender was not significant, i.e., candidate scores were not significantly affected by whether their rater is of the same or opposite gender.

From the above analyses it appears that the impact of both candidate and rater gender on test scores in the IELTS oral interview is not significant. However, this conclusion should be regarded with some caution given the relatively small data set available for analysis. Furthermore, the findings here do not imply that the measurement

Table 15 Bias calibration report, candidate gender – rater gender interaction

Candidate gender	Rater gender	Bias (logit)	Error	Z-score	Infit mn sq
F	F	0.06	0.20	0.3	0.8
M	F	–0.06	0.20	–0.3	1.9
F	M	–0.06	0.20	–0.3	1.4
M	M	0.07	0.21	0.3	0.9

process can be considered flawless: it could still be true that certain candidates are rated significantly more harshly or leniently by individual raters compared to the way that rater treats other candidates irrespective of candidate or rater gender. This possibility leads us to the fourth question.

Question 4: Do individual raters score individual candidates significantly more harshly or leniently compared to the way they treat other candidates irrespective of candidate or rater gender? The analysis here revealed that there were two such occurrences out of a total of 128 ratings (see Table 16). In both instances, the raters marked the specified candidate significantly more harshly than they did other candidates. Considering this very low figure it can be concluded that there was a high degree of intra-rater reliability in this study and, therefore, that the overwhelming majority of candidates were treated fairly in the scoring process.

Table 16 Bias calibration report, biased interactions between individual candidates and raters

Candidate ID	Rater ID	Bias (logit)	Error	Z-score	Infit mn sq
1	12	5.07	1.92	2.6	0.0
11	16	4.47	1.93	2.3	0.7

IV Conclusion

To sum up the findings, therefore, the results from both the discourse and test score analyses suggested that gender did not have a significant impact on the IELTS oral interview in this study. The discourse analysis indicated, first, in relative terms, that there was limited use of overlaps, negligible use of interruptions and widespread use of

minimal responses in the interviews. Secondly, the use of these features did not appear to follow any clear gendered pattern. Thirdly, there was a high degree of variability in the use of overlaps and especially minimal responses within the different gender pairings. Most importantly, perhaps, both female and male participants indicated their ability to make supportive contributions to the interviews through their use of positive overlaps and minimal responses in particular. A collaborative style is therefore clearly not exclusively the province of female speakers in the testing context.

The test scores analyses also revealed that the gender of candidates and raters did not have a significant impact on the rating process. This finding, in particular, conflicts with other recent studies which have reported a significant gender effect in the rating of test-takers, although, as noted earlier in this article, the direction of this effect has not been consistent.

Why there was little or no discernible gender effect in either the interviews or subsequent ratings in this study is difficult to determine. Some of the possible reasons will now be examined. In terms of the interview process, perhaps the test tasks used and/or the roles of interviewer and candidate are particularly gender neutral in the IELTS test. Might a clearer gender effect emerge in oral tests where candidates are paired? Alternatively, in terms of methodology, is it possible that pre-selecting the discourse features used to examine the interviews in this study meant that the analysis ignored other ways in which gender may have been accomplished?

In terms of the rating process, could it be that the global band scale used in the test is not sensitive enough to register a gender effect amongst raters where it does exist? Or else, does focusing on the scores of raters who were not the original interviewers in this study mask a gender effect that results from the interaction between the interviewing and rating processes under normal conditions? Would there have been evidence of a gender effect in the ratings if the test performances had been video-taped rather than audio-taped? Any one or combination of these factors may account for the observed lack of gender effect in this study.

However, another way to understand why this and other studies into the impact of gender in speaking tests seem to contradict each other is to speculate from a broader social perspective about characteristics of the context and participants which might bring gender differences into play rather than simply on the test instrument itself. It is highly possible that aspects of the testing context itself, such as the purpose of the test, the language being tested, the country where it is administered as well as the social identities of the interviewer and test-taker (including their gender, age, ethnicity and perceived

status), may determine whether significant gender differences emerge in both the interviewing and rating processes. For instance, in Australia the IELTS oral interview is conducted by experienced ESL teachers of the host country who often work with international students on a regular basis. Their behaviour in the interviews may be most strongly influenced by how they view their task. If they consider it to be closely aligned to their teaching role then it is possible they will adopt a supportive, facilitative interviewer style. If they view it as more distant from their teaching role – more in terms of say impartial judge or gatekeeper – they may use a much less supportive style. This, in turn, could affect the way the candidate responds to them. In other words, the professional orientation of the teacher-as-interviewer may influence their behaviour more strongly then gender differences.

Furthermore, the fact that gendered differences amongst interviewers and candidates were not clearly evident in the interviews may have reduced the salience of gender to the raters who subsequently scored the audio-taped performances without significant gender bias. However, in other test settings where interviewers are not trained language teachers, then perhaps both the interviewing and rating processes may be more significantly affected by gender differences. Further research on these issues needs to undertaken.

It would appear, therefore, that gendered differences are not inevitable in the testing context. This is consistent with recent thinking in the fields of both gender studies and applied linguistics suggesting that gender competes with other aspects of an individual's social identity in a fluid and dynamic fashion. In one situation it may be strongly foregrounded, in another much less so. In short, we cannot always easily predict when gender will have a significant impact on speaking tests, and this seems to be equally true for both the interviewing and rating processes.

Acknowledgements

I am grateful to IELTS Australia for the research grant which enabled me to conduct this study and to Jeanette Carter for her assistance with many aspects of the project. This article is a revised version of a report published in 2000 by IELTS Australia entitled 'The impact of gender in the IELTS oral interview'.

IV References

Brown, A. and **Hill, K.** 1996: Interviewer style and candidate performance in the IELTS oral interview. IELTS Australia Reports Round 1.

Buckingham, A. 1997: Oral language testing: do the age, status and gender of the interlocutor make a difference? Unpublished MA dissertation, University of Reading.

Coates, J. 1993: *Women, men and language*. 2nd edition. London: Longman.

Freed, A.F. 1995: Language and gender. *Annual Review of Applied Linguistics* 15, 3–22.

Freed, A.F. and **Greenwood, A.** 1996: Women, men, and type of talk: what makes the difference? *Language in Society* 25, 1–26.

Freeman, R. and **McElhinny, B.** 1996: Language and gender. In McKay, S.L. and Hornberger, N.H., editors, *Sociolinguistics and language teaching*. Cambridge: Cambridge University Press, 218–80.

Lazaraton, A. 1996: Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13, 151–72.

Linacre, J.M. and **Wright, B.** 1992. *FACETS: Rasch measurement computer program, version 2.6.* Chicago, IL: Mesa Press.

Locke, C. 1984: The influence of the interviewer on student performance in tests of foreign language oral/aural skills. Unpublished MA project, University of Reading.

Maltz, D. and **Borker, R.** 1982: A cultural approach to male–female miscommunication. In Gumperz , J., editor, *Language and social identity*. Cambridge: Cambridge University Press, 196–216.

McNamara, T.F. and **Lumley, T.** 1997: The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing* 14, 142–51.

Morton, J., Wigglesworth, G. and **Williams, D.** 1997: Approaches to the evaluation of interviewer behaviour in oral tests. In Brindley, G. and Wigglesworth, G., editors, *Access: issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, 175–96.

O'Sullivan, B. 2000: Exploring gender and oral proficiency interview performance. *System* 28, 373–86.

Porter, D. 1991a: Affective factors in language testing. In Alderson, C.J. and North, B., editors, *Language testing in the 1990s*. London: Modern English Publications, 32–40.

—— 1991b: Affective factors in the assessment of oral interaction: gender and status. In Arnivan, S., editor, *Current developments in language testing*. Anthology series 25. Singapore: SEAMEO Regional Language Centre, 92–102.

Porter, D. and **Shen Shu-Hung** 1991: Sex, status and style in the interview. *The Dolphin* 21, 117–28.

Schegloff, E.A. 1993 Reflections on quantification in the study of conversation. *Research on Language and Social Interaction* 26, 99–128.

Stubbe, M. 1998: Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics* 29, 257–89.

Sunderland, J. 1995: Gender and language testing. *Language Testing Update* 17, 24–35.

Tannen, D. 1984: *Conversational style: analysing talk among friends*. Norwood, NJ: Ablex.
—— 1990: *You just don't understand: women and men in conversation*. New York: William Morrow.
Thwaite, A. 1993: Gender differences in spoken interaction in same dyadic conversations in Australian English. In Winter, J. and Wigglesworth, G., editors, *Language and gender in the Australian context. Australian Review of Applied Linguistics Series S No.10*, 149–79.
UCLES 1998: *The IELTS Handbook*. Cambridge: University of Cambridge Local Examinations Syndicate.
—— 2000: *The IELTS Handbook*. Cambridge: University of Cambridge Local Examinations Syndicate.
Young, R. and **Milanovic, M.** 1992: Discourse validation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 403–24.

Appendix 1 Transcription notation

1 Unfilled pauses and gaps: periods of silence, timed in tenths of a second by counting 'beats' of elapsed time in accordance with the rhythm of the preceding speech. Micropauses, those of less than 0.2 seconds are symbolised (.); longer pauses appear as time within parentheses: e.g., (0.8) = 0.8 seconds. Where 'real' time is indicated (e.g., in between the end of task instructions and the beginning of the candidate's response brackets { } are used.
2 Repair phenomena: reformulations are indicated by a hyphen -.
3 Intonation: a period . indicates a falling intonation, a question mark ? marks a rising intonation and a comma , is used for continuing intonation.
4 Overlapping talk: brackets [ ] are used to indicate overlaps i.e., where utterances start and/or end simultaneously.
5 Transcription doubt or uncertainty: these are marked by a question mark within parenthesis (?).
6 Quiet talk: percent signs %% are used to mark the boundaries of quiet talk.
7 Latched utterances: i.e., where there is no interval between utterances: equal signs = are used at the end of the first utterance and at the beginning of the second utterance.
8 Lengthened sounds or syllables: a colon : is used; more colons prolong the stretch.
9 Speakers: The interviewer is indicated by I and the candidate by C.