4. The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test

Author
Barry O'Sullivan
University of Roehampton, UK

Yang Lu
University of Reading, UK

Grant awarded Round 8, 2002

This paper shows that the deviations examiners make from the interlocutor frame in the IELTS Speaking Test have little significant impact on the language produced by candidates.

ABSTRACT

The Interlocutor Frame (IF) was introduced by Cambridge ESOL in the early 1990s to ensure that all test events conform to the original test design so that all test-takers participate in essentially the same event. While essentially successful, Lazaraton (1992, 2002) demonstrated that examiners sometimes deviate from the IF under test conditions. This study of the IELTS Speaking Test set out to locate specific sources of deviation, the nature of these deviations and their effect on the language of the candidates.

Sixty recordings of test events were analysed. The methodology involved the identification of deviations from the IF, and then the transcription of the candidates' pre- and post-deviation output. The deviations were classified and the test-takers' pre- and post-deviation oral production compared in terms of *elaborating and expanding* in discourse, *linguistic accuracy* and *complexity* as well as *fluency*.

Results indicate that the first two parts of the Speaking Test are quite stable in terms of deviations, with relatively few noted, and the impact of these deviations on the language of the candidates was essentially negligible in practical terms. However, in the final part of the Test, there appears to have been a somewhat different pattern of behaviour, particularly in relation to the number of paraphrased questions used by the examiners. The impact on candidate language again appears to have been minimal.

One implication of these findings is that it may be possible to allow for some flexibility in the Interlocutor Frame, though this should be limited to allowing for examiner paraphrasing of questions.

AUTHOR BIODATA

BARRY O'SULLIVAN

Barry O'Sullivan has a PhD in language testing, and is particularly interested in issues related to performance testing, test validation and test-data management and analysis. He has lectured for many years on various aspects of language testing, and is currently Director of the Centre for Language Assessment Research (CLARe) at Roehampton University, London.

Barry's publications have appeared in a number of international journals and he has presented his work at international conferences around the world. His book *Issues in Business English Testing: the BEC revision project* was published in 2006 by Cambridge University Press in the Studies in Language Testing series; and his next book is due to appear later this year. Barry is very active in language testing around the world and currently works with government ministries, universities and test developers in Europe, Asia, the Middle East and Central America. In addition to his work in the area of language testing, Barry taught in Ireland, England, Peru and Japan before taking up his current post.

YANG LU

Dr Yang Lu has a BA in English and English Literature from Jilin University, China. She obtained both her MA and doctorate degrees from the University of Reading. Her PhD investigates the nature of EFL test-takers' spoken discourse competence. Dr Yang Lu has 18 years' experience of language teaching and testing. She worked first as a classroom teacher and later as Director of the ESP Faculty and Deputy Coordinator of a British Council project based at Qingdao University, where she also worked as Associate Professor of English. Her academic interests are spoken discourse analysis and its applications in classroom and oral assessment contexts.

Dr Yang Lu's publications include papers on: EFL learners' interlanguage pragmatics; application of the Birmingham School approach; the roles of fuzziness in English language oral communication; and task-based grammar teaching. She has presented different aspects of her work at a number of international conferences. Dr Yang Lu was a Spaan Fellow for a validation study on the impact of examiners' conversational styles.

IELTS RESEARCH REPORTS, VOLUME 6, 2006

Published by: IELTS Australia and British Council

Project Managers: Jenny Osborne, IELTS Australia, Uyen Tran, British Council

Editors: Petronella McGovern, Dr Steve Walsh

British Council
Bridgewater House
© British Council 2006

IELTS Australia Pty Limited
ABN 84 008 664 766 (incorporated in the ACT)
© IELTS Australia Pty Limited 2006

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

CONTENTS

1 INTRODUCTION

While research into various aspects of speaking tests has become more common and more varied over the past decade, there is still great scope for researchers in the area, as the fractured nature of research to date betrays the lack of a systematic research agenda in the field.

O'Sullivan (2000) called for a focus on a more clearly defined socio-cognitive perspective on speaking, and this is reflected in the framework for validating speaking tests outlined by Weir (2005). This is of particular relevance in tests of speaking where candidates are asked to interact either with other candidates and an examiner or, in the case of IELTS, with an examiner only. The co-constructive nature of spoken language means that the role played by the examiner-as-interlocutor in the test event is central to that event. One source of construct irrelevant variance in face-to-face speaking tests lies in the potential for examiners to misrepresent the developer's construct either by consciously or subconsciously changing the way in which individual candidates are examined. There is considerable anecdotal evidence to suggest that examiners have a tendency to deviate from planned patterns of discourse during face-to-face speaking tests, and to some extent we might want this to happen, for example to allow for the interaction to develop in an authentic way. However, the dangers inherent in examining speaking by using what is sometimes called a conversational interview (Brown 2003:1) are far more likely to result in test events that are essentially unique, though this is something that can be said of any truly free conversation – see also van Lier's (1989) criticism of this type of test in which he convincingly argues that true conversation is not necessarily reflected in interactions performed under test conditions. These dangers, which include unpredictability in terms of topic, linguistic input and expected output, all of which can have an impact on test-taker performance, have long been noted in the language testing literature (see Wilds 1975; Shohamy 1983; Bachman 1988; 1990; Stansfield 1991; Stansfield & Kenyon 1992; McNamara 1996; Lazaraton 1996a).

There have been a number of studies in which rater linguistic behaviour has been explored in terms of its impact on candidate performance (see Brown & Hill 1998; Brown & Lumley 1997; Young & Milanovic 1992), and others in which the focus was on linguistic behaviour without an overt focus on the impact on candidate performance (Lazaraton 1996a; Lazaraton 1996b; Ross 1992; Ross & Berwick 1992). Other studies have looked at the broader context of examiner behaviour (Brown 1995; Chalhoub-Deville 1995; Halleck 1996; Hasselgren 1997; Lumley 1998; Lumley & O'Sullivan 2000; Thompson 1995; Upshur & Turner 1999). The results of these studies suggest that there is likely to be systematic variation in how examiners behave during speaking test events, in relation both to their language and to their rating.

These studies have tended to look either at the scores achieved by candidates or at the identification of specific variations in rater behaviour and have not focused so much on how the language of the candidates might be affected as a result of particular examiner linguistic behaviour (with the exception perhaps of Brown & Hill 1998). Another limitation of these studies (at least in terms of the study reported here) is the fact that they were almost all conducted on so-called conversational interviews (with the exception of the work of Lazaraton 2002). Since the 1990s, many tests have moved away from this format, to a more tightly controlled model of spoken test using an Interlocutor Frame.

2 THE INTERLOCUTOR FRAME

An Interlocutor Frame (IF) is essentially a script. The idea of using such a device is to ensure that all test events conform to the original test design so that all test-takers participate in essentially the same event. Of course, the very nature of *live* interaction means that no two are ever likely to be *exactly*

the same but some measure of standardisation is essential if test-takers are to be treated fairly and equitably. Such frames were first introduced by Cambridge ESOL in the early 1990s (Saville & Hargreaves 1999) to increase standardisation of examiner behaviour in the test event – though it was demonstrated by Lazaraton (1992) that there might still be deviations from the Interlocutor Frame even after examiner training. This may have been at least partly a response by the examiners to the extreme rigidity of the early frames, where all responses (verbal, paraverbal and non-verbal) were scripted. Later work by Lazaraton (2002) provided evidence of the effect of examiner language and behaviour on ratings, and contributed to the development of the less rigid Interlocutor Frames used in subsequent speaking tests.

As we have pointed out above, the IF was originally introduced to give the test developer more control of the test event. However, Lazaraton has demonstrated that, when it comes to the actual event itself, examiners still have the potential to deviate from any frame.

The questions that emerge from this are:

1. Are there identifiable *positions* in the IELTS Speaking Test in which examiners tend to deviate from the Interlocutor Frame?
2. Where a deviation occurs, what is the *nature* of the deviation?
3. Where a deviation occurs, what is the *effect* on the linguistic performance of the candidate?

To investigate these questions, it was decided to revisit the IELTS Speaking Test following earlier work. Brown & Hill (1998) and Brown (2003) reported a study based on a version of the IELTS Speaking Test which was operational between 1989 and 2001. Findings from this work, together with outcomes from other studies on the IELTS Speaking Test, informed a major revision of the test in the late 1990s; from July 2001 the revised test incorporated an Interlocutor Frame for the first time to reduce rater variability (see Taylor, in press). (The structure of the current test is described briefly below in 3.1.) Since its introduction, the functioning of the Interlocutor Frame in the IELTS Speaking Test has been the focus of ongoing research and validation work; the study reported here forms part of that agenda and is intended to help shape future changes to the IF and to inform procedures for IELTS examiner training and standardisation.

3 METHODOLOGY

Previous studies into the use by examiners of Interlocutor Frames used time-consuming, and therefore, extremely expensive research methodologies, particularly conversation analysis (see the work of Lazaraton 1992, 1996a, 1996b, 2002). Here, an alternative methodology is applied. In this methodology, audio-recorded examination events were first studied for deviations from the specified IF. These deviations were then coded and the area of discourse around them transcribed and analysed.

The methodology involved the identification of deviations from the existing IF (in 'real time'). The deviations identified were then transcribed to identify the test-takers' pre- and post-deviation oral output. A total of approximately 60 recorded live IELTS Speaking Tests undertaken by a range of different examiners were analysed. The deviations were classified and the test-takers' pre- and post-deviation oral production compared in terms of *elaborating and expanding* in discourse, *linguistic accuracy* and *complexity* as well as *fluency*.

3.1 The IELTS Speaking Test

The Speaking Test is one of four skills-focused components which make up the IELTS examination administered by the IELTS partners – Cambridge ESOL, British Council and IELTS Australia.

The Test consists of a one-to-one, face-to-face oral interview with a single examiner and candidate. All IELTS interviews are audio-taped for purposes of quality assurance and monitoring. The test has three parts (see Figure 1), each of which is designed to elicit different profiles of a candidate's language. This has been shown to be the case in speaking tests for the Cambridge ESOL Main Suite examinations by O'Sullivan, Weir & Saville (2002) and O'Sullivan & Saville (2000) through use of an observation checklist. Brooks (2003) reports how a similar methodology was developed for and applied to IELTS; an internal Cambridge ESOL study (Brooks 2002) demonstrated that the different IELTS test parts were capable of fulfilling a specific function in terms of interaction pattern, task input and candidate output.

Part	Nature of interaction	Timing
Part 1 Introduction and interview	Examiner introduces him/herself and confirms candidate's identity. Examiner interviews candidate using verbal questions selected from familiar topic frames	 4-5 minutes
Part 2 Individual long turn	Examiner asks candidate to speak for 1-2 minutes on a particular topic based on written input in the form of a candidate task card and content-focused prompts. Examiner asks one or two questions to round off the long turn.	3-4 minutes (incl. 1 minute preparation time)
Part 3 Two-way discussion	Examiner invites candidate to participate in discussion of a more abstract nature, based on verbal questions thematically linked to Part 2 topic.	4-5 minutes

Figure 1: IELTS Speaking Test format

The examiner interacts with the candidate and awards scores on four analytical criteria which contribute to an overall band score for speaking on a nine-point scale (further details of test format and scoring are available on the IELTS website: www.ielts.org). Since this study is concerned with the language of the test event as opposed to the outcome (ie score awarded) no further discussion of the scoring will be entered into at this point except to say that the band scores were used to assist the researchers in selecting a range of test events in which candidates of different levels were represented.

The test version selected for use in this study is Version 88, a version that was in use after July 2001, but that was later retired.

3.2 Test-takers

A total of 85 audio-taped live IELTS Speaking Test events using Test Version 88 were selected from administrations of the test conducted during 2002. Of these, 70 were selected for the study after consideration of test-takers' nationality and first language. This was done to reflect the composition of the general IELTS candidature worldwide. Band scores awarded to candidates were also looked at to avoid a situation where one nationality might be over-represented at the different overall score levels. However, this was not always successful as it is clear from the overall patterns of IELTS scores that there are differences in performance levels across the many different nationalities represented in the test-taking population.

After an initial listening, a further eight performances were excluded because of poor quality of recording (previous experience has shown that this makes accurate transcription almost impossible), leaving 62 speaking performances for inclusion in the analysis. There were 21 female test-takers and 41 males. The language and nationality profile is shown in Table 1. From this table we can see that the population represents a wide range of first languages (17) and nationalities (18). This sample allows for some level of generalisation to the main IELTS population. More detailed information about the test-takers can be found in Appendix 1.

Language	Nationality	Number	Language	Nationality	Number
Arabic	Iraq	1	Portuguese	Brazil	1
Arabic	Oman	5	Portuguese	Portugal	1
Arabic	UAE	3	Punjabi	India	3
Bengali	Bangladesh	3	Pushtu	Pakistan	1
Chinese	China	17	Spanish	Colombia	1
Chinese	Taiwan	1	Spanish	Mexico	1
Farsi	Iran	1	Swedish	Sweden	5
German	Switzerland	1	Telugu	India	1
Hindi	India	5	Urdu	Pakistan	4
Japanese	Japan	1	Other	India	1
Korean	S Korea	1	Other	Malawi	1

Table 1: Language and nationality profile

3.3 The examiners

A total of 52 examiners conducted the 62 tests included in the matrix. The intention was to include as large a number of examiners as possible in order to minimise any impact on the data of non-standard behaviour by particular judges. For this reason, care was also taken to ensure that no one examiner would conduct the test on more than three occasions.

As all of the test events used in this study were 'live' (ie recordings of actual examinations), the conditions under which the tests were administered were controlled. This meant that all of the examiners were fully trained and standardised and had experience working with this test.

4 THE STUDY

4.1 The coding process

The first listening was undertaken to identify the *nature* and *location* of the obvious and recurring deviations from the Interlocutor Frame by examiners. The more frequent deviations were first identified, then categorised, and finally coded. Efforts were made to be consistent with the coding according to a set of definitions given to these deviations which was generated gradually during the listening. As is usual with this kind of work, definitions were very sketchy at the outset but became more clearly defined when the first careful listening was finished. Table 2 presents the findings of this first listening.

Types of deviations	Coding	Definitions
interrupting question	**itr**	question asked that stops the test-taker's answer
hesitated question	**hes**	question asked hesitatingly – possibly because of unfamiliarity with the interlocutor frame
paraphrased question	**para**	question that is rephrased without test-taker's request – appears to be based on examiner's judgement of the candidate's listening comprehension ability
paraphrased and explained question	**parax**	question that is both paraphrased and explained with example with or without test-taker's request
comments after replies	**com**	comment made after test-taker's reply that is more than the acknowledgement or acceptance the examiner is supposed to give; it tends to make the discourse more interactive
improvised question	**imp**	question that is not part of the interlocutor frame but asked based on test-taker's reply – very often about their personal interests or background
informal chatting	**chat**	informal discussion mainly held by examiner who is interested in test-taker's experience or background
loud laughing	**la**	examiner's loud laughing caused by test-taker's reply or answer
offer of clues	**cl**	examiner's utterance made to offer a hint and/or to facilitate candidate reply

Table 2: Development of coding for deviations (Listening 1)

A second careful listening was undertaken to confirm the identification of deviations, to check the coding for each case and to decide on a final list of the deviations to be examined. As can be seen from Table 2, there were two distinct types of deviation related to paraphrasing. While this coding appeared at first a useful distinction, it became quite difficult to operationalise, as the study was based on audio tapes, a medium which does not allow the researcher to observe the body language and facial expressions of the parties involved. This made it practically impossible to know whether paraphrasing was performed in response to test-takers' requests (verbal or non-verbal) or volunteered by the examiner. Therefore, the decision was made to collapse the two 'paraphrasing' categories and to report only the single category 'paraphrase'.

A list of occurrences of the deviations resulted as shown in Table 3:

Types of deviations	Coding	Occurrences
interrupting question	Itr	34
hesitated question	Hes	7
paraphrased question	Para	47
comments after reply	Com	12
improvised question	Imp	28
informal chatting	Chat	9
Laughing	La	5
Clues	Cl	2

Table 3: Occurrences of deviations

Two decisions were made after the second listening:

1. The four types of deviations that were found to be most frequent in the tests were selected for investigation. They are: *interrupting question, paraphrased question, comment after replies* and *improvised question*. We also believe that these four types of deviations can be established because in the *Instructions to IELTS Examiners* (Cambridge ESOL 2001) it is made very clear to the examiners that:

    - The Interlocutor Frame is used for the purpose of standardisation in order that all candidates are treated fairly and equally. Deviations from the script may introduce easier or more difficult language or change the focus of a task.
    - In Part 1 the exact words in the Frame should be used. Reformulating and explaining the questions in the examiner's own words are not allowed.
    - In Part 2 examiners must use the words provided in the Frame to introduce the long turn task.
    - In Part 3 the Frame is less controlled so that the examiner's language can be accommodated to the level of the candidate being examined.
    - In all parts of the test, examiners should refrain from making unscripted comments or asides.

Explanation needs to be given at this point about the rationale for including the *interrupting questions* and *paraphrased questions* in Part 3 as deviation types. Although, understandably, examiners sometimes cannot help stopping test-takers whose replies in Part 1 and 3 are lengthy and slow down the procession of the Speaking Test, this should be done in a more subtle way with body language as suggested in IELTS Speaking Test-FAQs and Feedback document (Cambridge ESOL 2001) or by using more tentative verbal hints. These strategies are suggested so as to limit any potential impact on future candidate linguistic performance. The interrupting questions we have coded as deviations occur neither after lengthy replies by test-takers nor are they made in a non-threatening (ie tentative) manner.

In Part 1, as the *Instructions to IELTS Examiners* states, 'examiners should not explain any vocabulary in the frame'. Therefore, any reformulating of the questions is regarded here as a deviation and coded as such. However, in Part 3 examiners have more independence and flexibility within the Frame and are even encouraged 'to develop the topic in a variety of directions according to the responses from the candidates' (Cambridge ESOL 2001). The examiners' decisions to reformulate, rephrase, exemplify or paraphrase the questions in Part 3 were noticed in the first listening of the tapes. For most of the cases this was done without a specific request from the test-takers and appears to have been based on examiner judgements of the individual test-taker's level of proficiency and ability to discuss the comparatively more abstract topics contained in this section of the Test. However, it should be noted that this part differs from Parts 1 and 2 in that the prompts are just that – indicative prompts designed for them to articulate in a way that is appropriate to the level of the candidate, but *not* fully scripted questions for them to 'read off the page' as in Parts 1 and 2.

2. The second decision concerned the amount of speech to be transcribed on either side of the deviation. Since it was believed that we needed a significant amount of language for transcription so that realistic observations could be made, and that all language chunks transcribed should be of similar length, we decided that 30 seconds of pre- and post-deviations should be transcribed and analysed to provide reliable data for investigation. Details of the transcription conventions used are given below. Pre-deviations that were found to be overlapping with the post-deviation of a previous question could not be transcribed. As a

result, the number of pre- and post-deviation sections from the oral production by the candidates in each category was reduced, the final numbers being:

- 33 paraphrased questions
- 26 interrupting questions
- 17 improvised questions
- 9 comments after replies.

4.2 Locating deviations

The reason for looking at the points of deviation was to identify places in the Interlocutor Frame that might be prone to lead to unintended breakdowns or deviations. It was thought that locating these 'weak' points in the Frame would offer valuable insights into why the breakdown occurred and lead to a series of practical recommendations for the improvement of the IF as well as guidance for examiner training. Two procedures were undertaken for this purpose:

1. Occurrences of each deviation in the three test parts were identified to highlight where they were most likely to occur.
2. Occurrences of the questions where examiners deviated most were counted in order to discover where certain deviations would be most likely to occur within each test part.

4.3 Transcribing

Transcribing was conducted after the second, more detailed listening. The maximum amount of time for each pre- or post-deviation chunk was 30 seconds.

Conventions for transcriptions are as below:

- er ---- filled pauses
- x ---- one syllable of a non-transcribed word
- …… ---- not transcribed pre- or post-deviation oral production.

A total of over 10,000 were transcribed in the pre- and post-deviation data. This dataset was then divided into nine files:

- Part 1. com (comments after replies in Part 1)
- Part 2. com (comments after replies in Part 2)
- Part 3. com (comments after replies in Part 3)
- Part 1. itr (interrupting questions in Part 1)
- Part 3. itr (interrupting questions in Part 3)
- Part 1. imp (improvised questions in Part 1)
- Part 3. imp (improvised questions in Part 3)
- Part 1. para (paraphrased questions in Part 1)
- Part 3. para (paraphrased questions in Part 3)

5 ANALYSIS

To realise the aim of the study (to compare the quality of the candidates' oral production in the pre and post deviation sections), four categories of measure were used; these are presented in Table 4 along with the sub-categories.

Category of measures	Sub-category of measures
Fluency	1. filled pauses per AS-unit 2. words per second (excluding repetitions, self-corrections and filled pauses)
Grammatical Accuracy	1. number of errors of plural or singular forms per word 2. number of errors of subject and verb agreement per word
Linguistic Complexity	Average number of clauses per AS-unit
Discoursal Performance	1. number of *expanding* moves per T-unit 2. number of *elaborating* moves per T-unit 3. number of *enhancing* moves per T-unit

Table 4: Categories of measures used in transcription analysis

The Analysis of Speech Unit, or AS-unit (Foster, Tonkyn & Wigglesworth 2000) was used for calculating filled pauses and investigating linguistic complexity; for comparing the discoursal performance before and after deviations, the T-unit (Hunt 1970) was chosen as the unit in which changes were examined. The rationale for this approach is:

1. According to Foster et al (2000: 365), the AS-unit is 'a mainly syntactic unit…consisting of *an independent clause, or sub-clausal unit*, together with *any subordinate clause(s)* associated with either'. This allows us to analyse speech at different clausal units such as the non-finite clauses, so that the complexity of linguistic features can be measured.
2. Since studies of pausing in native-speaker speech have shown that pauses often occur at syntactic unit boundaries, especially at clausal boundaries (Raupach 1980; Garman 1990), the AS-unit was selected as the most appropriate unit for calculating filled pauses.
3. The T-unit is the 'shortest unit into which a piece of discourse can be cut without leaving any sentence fragments as residue' (Hunt 1970:189). The T-unit enables us to include in the analysis all acts, some of which can be coordinate clauses or fragments of clauses. This is beyond the scope of the AS-unit which regards these structures as separate units.

6 RESULTS

6.1 Overall

The results are presented in relation to the three research questions posed in section one. We will look at the overall evidence of deviation and at any apparent impact on test-taker language of these deviations. In addition, we will look at the location of the deviations for evidence of systematicity which may point to inherent weaknesses in the interlocutor frame method. The overall results are presented so as to reflect the four areas identified as the most common deviation type above.

6.1.1 Paraphrasing

The results suggest that there is a very limited impact on fluency, while in the other areas there are mixed results. There appears to be a reduction in accuracy immediately following the deviation in terms of plural/singular errors, though this is counteracted by the post-deviation increase in subject/verb agreement accuracy. It is in the area of complexity that the most obvious change occurs, with both the number of AS-units and the number of clauses per AS-unit appearing to significantly drop following the deviation. The discourse indicators also appear to show a mixed reaction. The results are grouped together as Table 5.

Fluency	Filled pauses per T-unit		Words per second	
	pre	post	pre	post
Average	1.021	1.346	1.77	1.67
Total	31.993	36.933	58.33	55.26

Accuracy	Plural/Single Error per word		Subject/Verb agreement Error per word	
	pre	post	pre	post
Average	0.01	0.01	0.02	0.03
Total	0.47	0.17	0.64	0.92

Complexity	Clauses per AS-unit	
	pre	post
Average	0.01	0.01
Total	0.47	0.17

Discourse	Expanding per T-Unit		Elaborating per T-Unit		Enhancing per T-Unit	
	Pre	Post	pre	post	pre	Post
Average	0.43	0.31	0.16	0.22	0.23	0.17
Total	14.28	10.28	5.41	7.12	7.75	5.57

Table 5: The impact of paraphrasing questions on candidate language

6.1.2 Interrupting

In Table 6 we can see that there is quite a large reduction in filled pauses per T-unit, though there is little change as regards the number of words spoken per second. Like the results from the paraphrasing analysis, there seems to be a reduction in accuracy immediately following the deviation in terms of plural/singular errors, though this is again reversed with the post-deviation increase in subject/verb agreement accuracy. The pattern found for complexity is not repeated here, and is instead seen to be much more inconsistent. The discourse indicators are the most consistent, with a slight drop in the post-deviation position, though this does not appear to be great enough to suggest a significant reaction.

Fluency	Filled pauses per T-unit		Words per second	
	Pre	post	pre	post
Average	1.035	0.558	1.832	1.857
Total	26.919	14.500	47.63	48.28

Accuracy	Plural/Single Error per word		Subject/Verb agreement Error per word	
	Pre	post	pre	post
Average	0.009	0.005	0.008	0.016
Total	0.222	0.142	0.207	0.428

Complexity	Clauses per AS-unit	
	Pre	post
Average	0.89	1.01
Total	23.05	26.13

Discourse	Expanding per T-Unit		Elaborating per T-Unit		Enhancing per T-Unit	
	pre	post	pre	post	pre	post
Average	0.356	0.340	0.118	0.058	0.147	0.125
Total	9.255	8.833	3.060	1.500	3.833	3.250

Table 6: The impact of interrupting questions on candidate language

6.1.3 Improvising

As far as the results for fluency are concerned (Table 7), there seems to be a significant reduction in the number of filled pauses following the deviation, though a corresponding reduction in the number of words spoken per second does not appear great. As for accuracy, there seems to be a very slight increase in the measures over the two sections, though the numbers are probably too small to draw any definite conclusions. With complexity, the picture is once again mixed, while the discourse indicators also appear to show little reaction apart from the amount of expanding carried out.

Fluency	Filled pauses per T-unit		Words per second	
	pre	Post	pre	post
Average	0.666	0.373	2.159	2.023
Total	11.328	6.333	36.710	34.390

Accuracy	Plural/Single Error per word		Subject/Verb agreement Error per word	
	pre	Post	pre	post
Average	0.005	0.008	0.012	0.026
Total	0.093	0.137	0.212	0.449

Complexity	Clauses per AS-unit	
	pre	Post
Average	1.217	1.431
Total	20.692	24.333

Discourse	Expanding per T-Unit		Elaborating per T-Unit		Enhancing per T-Unit	
	pre	post	Pre	post	pre	post
Average	0.340	0.152	0.156	0.153	0.198	0.229
Total	5.787	2.583	2.660	2.600	3.368	3.892

Table 7: The impact of improvising questions on candidate language

6.1.4 Commenting

In the results from the analysis of the language bordering the deviations which were identified as being related to unscripted comments made by the examiners, we can see that there is a drop in the number of filled pauses, while there is little significant change in the number of words spoken per second (Table 8). The figures for accuracy are so small that there seems little point in attempting to make any meaningful comment on them, while for complexity there is quite a large increase in the number of clauses per AS-unit. Finally, the discourse indicators seem to indicate a systematic decrease right across the board.

Fluency	Filled pauses per T-unit		Words per second	
	pre	Post	pre	post
Average	0.666	0.473	2.137	2.353
Total	4.983	4.386	19.230	21.180

Accuracy	Plural/Single Error per word		Subject/Verb agreement Error per word	
	pre	Post	pre	post
Average	0.000	0.002	0.008	0.015
Total	0.000	0.017	0.069	0.137

Complexity	Clauses per AS-unit	
	pre	post
Average	0.609	0.816
Total	5.483	7.343

Discourse	Expanding per T-Unit		Elaborating per T-Unit		Enhancing per T-Unit	
	pre	post	pre	post	pre	post
Average	0.372	0.257	0.206	0.083	0.307	0.254
Total	3.345	2.317	1.852	0.750	2.760	2.283

Table 8: The impact of commenting on responses on candidate language

6.2 Impact on test-takers' language of each deviation type

If we then review these results in terms of each of the four language areas, we can see that of the four deviation types, paraphrasing seems to result in relatively little change to the language performance of the candidates, while all other deviation types seem to be having a negative impact on fluency (see Table 9). However, the rate of speed does not appear to be affected to any great extent by the deviations.

The negative direction of interrupting/improvising/commenting' suggested by Table 9 could imply that examiners should really avoid doing any of these, while the positive direction of the impact of 'paraphrasing' suggests that examiners need not be so concerned about doing this because it may even have a positive impact?

Fluency	Filled pauses per T-unit		Words per second	
	pre	post	pre	Post
Paraphrasing	1.021	1.346	1.77	1.67
Interrupting	1.035	0.558	1.832	1.857
Improvising	0.666	0.373	2.159	2.023
Commenting	0.554	0.487	2.137	2.353

Table 9: The impact on fluency of each deviation type

In terms of the accuracy of the output, we can see that there does not appear to be any significant impact as a result of the deviations recorded here – though the numbers recorded may in any case be too small to make any meaningful difference (see Table 10).

Accuracy	Plural/Single Error per word		Subject/Verb agreement Error per word	
	pre	Post	pre	post
Paraphrasing	0.01	0.01	0.02	0.03
Interrupting	0.009	0.005	0.008	0.016
Improvising	0.005	0.008	0.012	0.026
Commenting	0.000	0.002	0.008	0.015

Table 10: The impact on accuracy of each deviation type

The complexity of the language is affected in different ways (Table 11). If anything, there is a slight increase in the complexity of the language used following each of the deviations with the exception of paraphrasing.

Complexity	Clauses per AS-unit	
	Pre	Post
Paraphrasing	0.01	0.01
Interrupting	0.89	1.01
Improvising	1.217	1.431
Commenting	0.609	0.816

Table 11: The impact on complexity of each deviation type

Finally, we can see from Table 12 that the amount of expanding undertaken by candidates is systematically reduced following all four deviation types, though the picture for elaborating and enhancing is quite mixed.

Discourse	Expanding per T-Unit		Elaborating per T- Unit		Enhancing per T- Unit	
	Pre	post	pre	post	Pre	post
Paraphrasing	0.43	0.31	0.16	0.22	0.23	0.17
Interrupting	0.356	0.340	0.118	0.058	0.147	0.125
Improvising	0.340	0.152	0.156	0.153	0.198	0.229
Commenting	0.372	0.257	0.206	0.083	0.307	0.254

Table 12: The impact on discourse of each deviation type

6.3 Location of deviations

The other aim of the research is to investigate where the deviations occur to identify a pattern of the possible or likely situations or conditions for the deviations to occur. Two kinds of deviation location were studied: deviations across the three test parts and deviation within each test part.

6.3.1 Deviations by test part

Table 13 shows the numbers of occurrences of both the transcribed and non-transcribed (ie where the amount of language on either side of the deviation was too small to make meaningful inferences from the analyses) deviations in the tasks used in the three parts of the test. The non-transcribed deviations are added here to give a more complete picture of the amount of deviation from the IF that actually took place during these test events.

	Deviation Type											
	Paraphrased Questions			Improvised Questions			Comments after Replies			Interrupting Questions		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Deviations analysed for this study	4	0	29	8	0	9	2	4	4	14	0	12
Total number of Deviations	4	0	43	10	0	18	2	4	6	19	0	15

Table 13: Number of deviations by test part

There are a number of clear tendencies implied by Table 13:

- Interrupting questions spread more or less evenly in Part 1 and Part 3. This is possibly due to the two-way nature of these parts both of which involve questions and answers. When the test-taker gives a longer reply than necessary from the point of view of the examiner, the examiner may ask the next question to stop the candidate's reply to the previous question in the middle of a sentence or even a word. The table also suggests that about 30% of interrupting questions do not result in an extended turn (at least 30 seconds) from the candidate. This may be due to the fact that the questions are rhetorical (and do not require a response); or they may be yes/no questions or questions that elicit only very short responses; or it may be that the questions are either not clearly heard or understood by the candidates (and are either ignored or poorly answered). Since these possibilities can have potentially different impacts on candidate performance, it is clear that this aspect of examiner behaviour deserves more detailed examination.
- There are more improvising questions in Part 3 than in Part 1, though the discourse patterns are the same. It is possible that the improvising questions in Part 3 result from the more abstract nature of the questions, and is most likely related to the way Part 3 is designed from the examiner's perspective – see the above discussion. However, under what conditions the examiners tend to ask questions which are not in the Frame but are spontaneously raised by the examiners according to information given by test-takers can only be disclosed by examining the location of deviations within tasks. We can also see that in only half of the instances was there enough language resulting from the improvised question to merit inclusion in this study. This implies that this question type did not tend to result in the elicitation of a meaningful response (in terms of length of utterance) and as such may not always impact on candidate performance – though any

lack of response may result in a lowering of the examiner's opinion of the proficiency level of the candidate. Again, more detailed study of this phenomenon is required.

- The only type of deviation observed in Part 2 (the individual long turn) was where the examiners made comments following the candidate responses. This is not really surprising when we consider that the nature of the task reduces the potential for paraphrasing and improvising questions. Also, since the candidates are told before they start the task that they will be stopped when time is up, interruptions are not expected to occur.
- Comments after test-takers' replies seem to occur most often in the Individual long turn task, if we bear in mind that in this part of the test examiners are only required to ask one or two rounding-off questions. Where and when these commenting deviations happen is certainly an interesting revelation, which will be discussed in the next part of this study.
- 91% of the paraphrasing questions occurred in Part 3, the two-way discussion task, where examiners invite the candidates to discuss the abstract aspect of the topic linked to Part 2 using unscripted questions. There is a suggestion here that in this part of the test the test-takers may have more difficulty answering the questions. Because of this, the examiners offered (based on their assessment of the candidates' levels of proficiency and ability to answer abstract questions) to rephrase or explain the questions without examinees' requests in most of the cases. The nature of the questions seems to be the cause, as there are far fewer paraphrasing questions in Part 1 where the purpose of the questions is to access factual information. When we consider the overall number of paraphrased questions to those analysed here, we can see that there is no difference for Part 1, suggesting that the paraphrasing was successful – in that it always resulted in a long response (at least 30 seconds). The picture in Part 3 is different; here one in three of the paraphrased questions failed to elicit a long enough turn to be included in this analysis. This suggests that the paraphrases failed to enlighten the candidates, perhaps not surprisingly, since the concepts in Part 3 tend to be more abstract, and therefore more difficult to paraphrase than in Part 1. The implication here is that examiner training, in this particular examination and in other tests in which this approach is used, should focus specifically on developing noticing, questioning and paraphrasing skills. It is also clear that this element of the test should be closely monitored in future administrations to ensure that candidate performances are not significantly affected by features of examiner behaviour that are not relevant to the skill being tested.

6.3.2 Details of the deviations

We will now examine each part of the test separately in order to identify which of the scripted questions were most likely to lead to or result in deviations from the Interlocutor Frame.

In Part 1 we can see that there is an even spread of deviations across the various questions (see Table 14). All of these questions are scripted for the examiner, who makes decisions on which ones to ask during the course of the test. It should be mentioned that there are more questions than listed in the table. They are not included here either because they were not asked by the examiners or there were no deviations associated with them.

PART 1	Paraphrased Questions	Improvised Questions	Comments after Replies	Interrupting Questions	Total Deviations
Introductory	Not analysed as this section is not assessed				
Place of origin	0	0	0	3	3
Work/study	0	0	1	2	3
Accommodation in UK	0	0	0	1	1
Everyday habits	0	1	0	0	1
Likes and personality	0	1	0	1	2
Favourite clothing	0	1	0	1	2
Language & other learning	0	1	0	0	1
Mode of learning	0	1	1	0	2
Cooking	0	0	0	1	1
New experiences	0	0	0	1	1
Museums & galleries	1	0	0	1	2
Most loved festivals	1	0	0	2	3
Festival games	1	0	0	0	1
Festival general	0	0	0	1	1
Sports	0	1	0	0	1
Sporting addictions	0	1	0	0	1
Most loved sports	1	1	0	0	2
Total	4	8	2	14	28

Table 14: Spread of deviations in Part 1

There are a number of observations that can be made at this juncture:

1. One examiner was responsible for five of the interrupting questions, suggesting that this is more of a test monitoring issue than a training issue (if it were a training issue we would expect to find a greater spread of occurrences).
2. The majority of the interrupting questions served to bring a candidate turn to an end, and as such do not appear to impact on candidate performance on the task.
3. We might need to think further about improvised questions. These are unscripted, and represented a real threat to the integrity of the test. It may well be that this type of question can be eliminated to a great extent by training and by the inclusion of a statement on the Frame specifically referring to the problem.
4. There does not appear to be a systematic pattern of deviation in relation to specific questions or question types (direct or slightly more abstract).

PART 2	Paraphrased Questions	Improvised Questions	Comments after Replies	Interrupting Questions	Total Deviations
Instructions	0	0	0	0	0
During long turn	0	0	0	0	0
Anyone with job?	0	0	2	0	2
Will you have the job?	0	0	2	0	2
Total	0	0	4	0	4

Table 15: Spread of deviations in Part 2

Table 15 shows that in Part 2, the Individual long turn, the examiners stayed very clearly with the Frame both during the introductory section of the task (when they were giving instructions) and while the candidate was involved in the long turn itself. There were four commenting responses by the examiners out of a total of 10 analysed for Part 2. A further probing of the data shows that they all happened when the examiners were rounding off this part by asking one or two questions. It also seems that at this point they tend to make comments about the candidates' answers to the questions, thus giving more acknowledgement and/or acceptance than required by the IF. This is an interesting finding, in that it suggests that examiners sense some need to 'backchannel'; although the original purpose of the rounding-off questions appears to have been to help examiners form a bridge from Part 2 to Part 3, they still seem to need to say something else. This is yet another area in which further exploration is likely to significantly add to our understanding of the Speaking Test event in general and examiner behaviour in particular.

In Part 3 (Table 16) we can see that the stable patterns observed in the first two parts are not repeated. Instead, there are a far greater number of deviations from the IF, though this is not unexpected as examiners are offered a choice of prompts from which to select and fashion their questions, depending on how the interaction evolves and are likely make unscripted contributions in this final part of the test. As we have seen above, Parts 1 and 3 are somewhat similar in design, with both designed to result in interactive communication. We would therefore expect to see similar patterns of behaviour from the examiners in the two parts. In fact, it is true that the patterns are strikingly similar in most areas – there are similar levels of occurrence of improvised questions, comments and interruptions. However, it is clear that there are far more instances of paraphrasing in this last part than in any of the others (in fact there are almost as many paraphrased questions in Part 3 as there are deviations in total for the other two parts). This may well be due to the less rigid nature of this final part, with the examiner offered a broad range of prompts to choose from when continuing the interaction, but is more likely due to the nature of the questions asked. Even if we take a less rigid view of paraphrasing (where scripted questions are asked using alternative wording or emphasis) and view this final part as being more loosely controlled, there is an issue with the degree of variation here. Examiners must regularly make 'real-time' decisions as to the value or relevance of questions. The fact that they are likely to make changes to the alternatives offered in this part of the test implied that they may not be totally comfortable with the alternatives offered, at least in terms of language.

PART 3	Paraphrased Questions	Improvised Questions	Comments after Replies	Interrupting Questions	Total Deviations
Factors for choice of career	3	2	1	3	9
Different factors for men/women	1	1	0	0	2
More important factors	5	2	0	3	10
Career structure important?	7	1	1	0	9
(±) of job for life and change of jobs	2	1	2	2	7
Future working patterns?	6	0	0	1	7
Being a boss (±)	1	1	0	2	4
Qualities of a good employer?	4	0	0	1	5
Future boss/employee relationship?	0	1	0	0	1
Total	29	9	4	12	45

Table 16: Spread of deviations in Part 3

We can see from Table 16 that some of the prompts appear to be more likely to result in paraphrasing than others (though the number of times each question was asked varied); it is possible that they potentially place a greater demand on the resources of the candidate in terms of background knowledge and understanding or awareness of European/Western working habits. The inability of candidates to respond to the questions may well result in the greater resort to paraphrasing seen in this part of the test. As with the other findings here, this raises as many questions as it answers, particularly in relation to examiner decision making, and the impact on overall score awarded of these deviations appearing so late in the test event.

7 CONCLUSIONS

In this study, we set out to explore the way in which IELTS examiners deviated from the relatively new Interlocutor Frame in the revised IELTS Speaking Test introduced in July 2001. We were interested to identify the nature and location of any deviations and to establish evidence of their impact on the language of the candidates who participated in the test events.

Our analyses appear to show that the first two parts of the Speaking Test are quite stable in terms of deviations, with relatively few noted; where these were found they were either associated with a single examiner or were unsystematically spread across the tasks. It was also clear that the examiners seemed to adhere very closely to the IF, and that the deviations that did occur came at natural interactional boundaries, such as at the end of medium or long turns from candidates. The impact of these deviations on the language of the candidates was essentially negligible in practical terms.

In the final part of the Test, there appears to have been a somewhat different pattern of behaviour, particularly in relation to the number of paraphrased questions used by the examiners. While Part 3 mirrors the other interactive task in terms of the number of improvised questions, comments on candidate responses and interrupting questions, there are seven times more paraphrased questions in the final task. The reasons for this difference appears to be related to the alternative format of the task which offers the examiner greater flexibility than in Parts 1 or 2: while the candidate was

basically asked information-based questions in the first part (typically of a personal nature), in the final part the questions asked the candidate to conjecture, offer opinions and reflect on often abstract topics. The other possible explanation is that the question types may have been beyond the typical candidate in terms of cognitive load or of their cultural or background knowledge. Whatever the cause of the deviations, the impact on candidate language appears to have been minimal, though it remains unclear if there was any impact on the final score awarded to candidates.

The use of an Interlocutor Frame is based on the rationale that without a scripted guide, examiners are likely to treat each test event as unique and that candidates risk being unfairly advantaged or disadvantaged as a result. Anecdotal evidence from some stakeholders, principally teachers and examiners, suggests that there is some concern that very tight Interlocutor Frames might cause examiners to become too stilted and unnatural in their language during a test event and that this has a negative impact on the face validity of the test. Test developers therefore have to balance the need to standardise the test event as much as possible (to ensure that all test-takers are examined under the same conditions and that an appropriate sample of language is elicited) against the need to give examiners some degree of flexibility so that they (and the more directly affected stakeholders) feel that the language of the event is natural and free flowing.

The results of our analyses suggest that examiners in the revised IELTS Speaking Test essentially adhere to the Interlocutor Frame they are given. The absence of systematicity in the location of deviations implies that the Frames are working as the test developers intended, and that there are no obvious points in the test in which deviation is likely to occur, particularly for the first two tasks. There is some slight cause for concern with the final part. It may well be that it is not possible to create a Frame that can adequately cope with the requirements of less controlled interaction, though the evidence from this study suggests that the extensive paraphrasing that resulted in the less controlled final section did not seriously impact on candidate performance; indeed, if anything it resulted in slightly improved performance. However, the evidence from this study implies that greater care with the creation of question options may result in a more successful implementation of the Frame. The most relevant implication of the findings of this study is that it may be possible to allow for some flexibility in the Interlocutor Frame, though this flexibility might be best confined to allowing for examiner paraphrasing of questions. That this might be achieved without negatively impacting on the language of the candidate is of particular interest.

ACKNOWLEDGEMENT

The authors would like to acknowledge the valuable input provided by Dr Lynda Taylor in preparing the report of the study that appears here.

REFERENCES

Bachman, LF, 1988, 'Problems in examining the validity of the ACTFL oral proficiency interview', *Studies in Second Language Acquisition*, vol 10, pp 149-64

Bachman, LF, 1990, *Fundamental considerations in language testing*, Oxford University Press, Oxford

Brooks, L, 2002, Report on functions observed in the old IELTS Speaking Test versus those in the revised Speaking Test, Internal Cambridge ESOL Report, Cambridge

Brooks, L, 2003, 'Converting an observation checklist for use with the IELTS Speaking Test', *Research Notes Issue 11*, University of Cambridge ESOL Examinations, Cambridge, pp 20-21

Brown, A, 1995, 'The effect of rater variables in the development of an occupation specific language performance test', *Language Testing*, vol 12, pp 1-15

Brown, A and Hill, K, 1998, 'Interviewer style and candidate performance in the IELTS oral interview', *IELTS Research Reports*, vol 1, IELTS Australia, Canberra, pp 1-19

Brown, A, 2003, 'Interviewer variation and the co-construction of speaking proficiency' *Language Testing*, vol 20, pp 1-25

Brown, A, and Lumley, T, 1997, 'Interviewer variability in specific-purpose language performance tests' in *Current Developments and Alternatives in Language Assessment*, eds A Huhta, V Kohonen, L Kurki-Suonio and S Luoma, University of Jyväskylä and University of Tampere, Jyväskylä, pp 137-150

Cambridge ESOL, 2001, *IELTS Speaking Test-FAQs and feedback*, Cambridge ESOL, Cambridge

Chalhoub-Deville, M, 1995, 'A contextualized approach to describing oral language proficiency', *Language Learning*, vol 45, pp 251-281

Foster, P, Tonkyn, A, and Wigglesworth, G, 2000, 'Measuring spoken language: a unit for all reasons' *Applied Linguistics*, vol 21, pp 354-375

Garman, M, 1990, *Psycholinguistics*, Cambridge University Press, Cambridge

Halleck, G, 1996, 'Interrater reliability of the OPI: using academic trainee raters', *Foreign Language Annals*, vol 29, pp 223-238

Hasselgren, A, 1997, 'Oral test subskill scores: what they tell us about raters and pupils' in *Current Developments and Alternatives in Language Assessment*, eds A Huhta, V Kohonen, L Kurki-Suonio and S Luoma, University of Jyväskylä and University of Tampere, Jyväskylä, pp 241-256

Hunt, K, 1970, *Syntactic maturity in school-children and adults*, Monograph of the Society for Research into Child Development

Lazaraton, A, 1992, 'The structural organisation of a language interview: a conversational analytic perspective', *System*, vol 20, pp 373-386

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

Lazaraton, A, 1996b, 'A qualitative approach to monitoring examiner conduct in the Cambridge Assessment of Spoken English (CASE)', *Performance, Testing and Cognition: Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, eds M Milanovic and N Saville, UCLES/Cambridge University Press, Cambridge, pp 18-33

Lazaraton, A, 2002, *A qualitative approach to the validation of oral language tests*, Cambridge University Press, Cambridge

Lumley, T, 1998, 'Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency', *English for Specific Purposes*, vol 17, pp 347-367

Lumley, T and O'Sullivan, B, 2000, 'The effect of speaker and topic variables on task performance in a tape-mediated assessment of speaking', Paper presented at the *2nd Annual Asian Language Assessment Research Forum*, The Hong Kong Polytechnic University, January 2000

McNamara, T, 1996, *Measuring second language performance*, Addison Wesley Longman, Harlow

O'Sullivan, B, 2000, 'Towards a model of performance in oral language testing', unpublished PhD dissertation, The University of Reading

O'Sullivan, B and Saville, N, 2000, 'Developing observation checklists for speaking tests', *Research Notes*, vol 3, pp 6-10

O'Sullivan, B, Weir, C and Saville, N, 2002, 'Using observation checklists to validate speaking-test tasks', *Language Testing*, vol 19, pp 33-56

Raupach, M, 1980, 'Temporal variables in first and second language production' in *Temporal Variables in Speech: Studies in Honor of Freida Goldman-Eissler*, eds HW Dechert and M Raupach, Mouton, The Hague

Ross, S, 1992, 'Accommodative questions in oral proficiency interviews', *Language Testing*, vol 9, pp 173-186

Ross, S and Berwick, R, 1992, 'The discourse of accommodation in oral proficiency interviews', *Studies in Second Language Acquisition*, vol 14, pp 159-176

Saville, N and Hargreaves, P, 1999, 'Assessing speaking in the revised FCE', *ELT Journal*, vol 53, pp 42-51

Shohamy, E, 1983, 'The stability of oral proficiency assessment on the oral interview testing procedures', *Language Learning*, vol 33, pp 527-40

Stansfield, CW, 1991, 'A comparative analysis of simulated oral proficiency interviews' in *Current Developments in Language Testing*, ed S Anivan, SEAMEO Regional Language Centre, Singapore, pp 199-209

Stansfield, CW and Kenyon, DM, 1992, 'Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview' *System* vol 20, pp 347-64

Taylor, L (in press), 'Introduction' in *IELTS Collected Papers: Research in Speaking and Writing Assessment. Studies in Language Testing Volume 19*, eds L Taylor and P Falvey, Cambridge ESOL/Cambridge University Press, Cambridge

Thompson, I, 1995, 'A study of interrater reliability of the ACTFL oral proficiency interview in five European Languages: Data from ESL, French, German, Russia, and Spanish', *Foreign Language Annals*, vol 28, pp 407-422

Upshur, JA and Turner, C, 1999, 'Systematic effects in the rating of second-language speaking ability: test method and learner discourse', *Language Testing*, vol 16, pp 82-111

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

Weir, C, 2005, *Language testing and validation: an evidence-based approach*, Palgrave, Oxford

Wilds, C, 1975, 'The oral interview test' in *Testing Language Proficiency*, eds RL Jones and B Spolsky, Center for Applied Linguistics, Arlington, VA, pp 29-44

Young, R and Milanovic, M, 1992, 'Discourse variation in oral proficiency interviews', *Studies in Second Language Acquisition*, vol 14, pp 403-424

APPENDIX 1: PROFILES OF THE TEST-TAKERS INCLUDED IN THE STUDY

Cand. No.	Gender	Score (speaking)	Nationality	L1	Examiner
1188	M	6	UAE	Arabic	9
0214	M	7	Jordan	Arabic	23
0105	F	6	UAE	Arabic	28
0397	M	7	Iraq	Arabic	22
0385	M	6	UAE	Arabic	22
0801	M	6	Oman	Arabic	12
0803	F	9	Oman	Arabic	48
0810	M	8	Oman	Arabic	48
0890	M	6	Oman	Arabic	53
0971	F	4	Oman	Arabic	50
0190	M	6	Bangladesh	Bengali	1
0403	M	6	Bangladesh	Bengali	22
0386	F	8	Bangladesh	Bengali	38
0931	M	5	China	Chinese	26
1089	M	6	China	Chinese	41
1119	M	5	China	Chinese	35
1383	F	6	China	Chinese	43
1427	M	5	China	Chinese	34
1436	F	6	China	Chinese	41
1487	F	4	Taiwan	Chinese	27
0437	F	6	China	Chinese	40
0466	F	4	China	Chinese	31
0478	M	6	China	Chinese	40
0439	M	5	China	Chinese	20
0515	M	7	China	Chinese	21
0549	M	6	China	Chinese	17
0702	M	6	China	Chinese	24
0717	M	5	China	Chinese	15
0727	M	5	China	Chinese	51
0752	F	5	China	Chinese	29
0168	M	6	China	Chinese	36
1396	M	6	Iran	Farsi	41

Cand. No.	Gender	Score (speaking)	Nationality	L1	Examiner
0767	M	7	Switzerland	German	18
3526	M	9	India	Hindi	37
3527	M	6	India	Hindi	37
5372	F	8	India	Hindi	39
5375	M	7	India	Hindi	39
6060	M	7	India	Hindi	11
0941	M	8	Japan	Japanese	32
1015	F	5	Japan	Japanese	6
0078	F	6	Japan	Japanese	45
0466	F	4	S Korea	Korean	30
1002	M	8	Malawi	Other	44
5371	M	6	India	Other	39
1423	F	7	Brazil	Portuguese	9
1494	M	7	Portugal	Portuguese	34
3880	M	8	India	Punjabi	33
4292	M	6	India	Punjabi	3
5415	M	6	India	Punjabi	4
1235	M	8	Pakistan	Pushtu	49
1236	F	7	Colombia	Spanish	32
0354	F	8	Mexico	Spanish	31
0996	M	8	Sweden	Swedish	9
0381	F	9	Sweden	Swedish	31
0128	M	8	Sweden	Swedish	10
0137	M	8	Sweden	Swedish	13
0152	F	7	Sweden	Swedish	14
6351	F	7	India	Telugu	25
0229	M	7	Pakistan	Urdu	8
0420	M	6	Pakistan	Urdu	52
0371	F	8	Pakistan	Urdu	42
0449	M	5	Pakistan	Urdu	42