Interaction in paired oral proficiency assessment in Spanish

Ana María Ducasse

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

November 2008

Faculty of Arts, School of Languages and Linguistics

The University of Melbourne

ABSTRACT

Speaking tasks involving peer-to-peer candidate interaction are increasingly being incorporated into language proficiency assessments, in both large-scale international testing contexts, and in smaller-scale, e.g. course-related, ones. This growth in the popularity and use of paired and group orals has stimulated research, particularly into the types of discourse produced and the possible impact of candidate background factors on performance. The validation of tests of speaking involving paired candidate formats is increasingly focusing on ways in which interaction between candidates is sustained, as 'interaction' emerges as an important criterion for assessing candidate performance. However, despite the fact that the strongest argument for the validity of peer-to-peer assessment lies in the claim that such tasks allow for the assessment of a broader range of interactional skills than the more traditional interview-format tests do, there is surprisingly little research into the judgments that are made of such performances, the development of scales to rate interactional skills, and candidates' awareness of features of interaction in such tests.

The thesis reports on the findings of a verbal protocol study of teacher-raters viewing the paired test discourse of seventeen beginner dyads in a university-based Spanish as a foreign language course. The fact that raters, and rating criteria, are in a crucial mediating position between output and outcomes warrants investigation into how raters construe the interaction in these tasks. The thesis also reports on the development of an evidence-based rating method to score peer L2 communicative interaction, based on experienced judges' comments on videoed student samples filmed during operational paired candidate tests of beginner level Spanish. Six trained and experienced raters generated criteria for communicative interaction, which were incorporated into a tool for developing a discourse sample based rating procedure, the Empirically-based, Binary-choice, Boundary-definition (EBB) method (Turner & Upshur, 1996). The thesis examines the features of paired candidate interaction which raters used to define

the boundary between performance levels. Three main criteria emerged as the boundaries used to define levels of interaction: non-verbal interpersonal communication, interactive listening and interactional management. These new notions are evidence of how peer interaction can be rated and they advance our understanding of the significant features of interaction in the rating context.

Finally, this thesis also focuses on the previously unexplored area of candidates' awareness of features of interaction in such tests. It uses a retrospective Stimulated Verbal Recall methodology with video-taped test performances from 25 participants in the same beginners' level test of spoken Spanish. The participant reports are analysed in terms of student orientation to paired interaction. The analysis addresses non-verbal interpersonal communication, interactive listening and interactional management, the features of interaction identified a priori in the rater Verbal Protocol Study. Examples from the data are presented and discussed. The findings confirm that candidates are significantly oriented to the features of interaction which are salient to raters and which have been included in the devised rating procedure.

The findings from the three parts of the thesis (the rater orientation study, the scoring procedure and the candidate awareness study) have implications for our understanding of the construct of effective interaction in paired candidate speaking tests, and for the development of appropriate rating procedures.

DECLARATION

This is to certify that

i. *the thesis comprises only my original work towards the PhD*
ii. *due acknowledgement has been made in the text to all other material used,*
iii. *the thesis is less than 100,000 words in length, exclusive of Tables, maps, bibliographies and appendices*

Signed

Ana Maria Ducasse

PREFACE

The following material has been published from the research undertaken for the thesis.

Ducasse, A. 'An empirically based rating scale for interaction in a Paired Test.' In A. Brown and K. Hill (Eds.) ¨Tasks and Criteria in Performance Assessment¨ Peter Lang (2009)

Ducasse, A. 'How do candidates view interaction in a paired oral?' In C. Gitsaki (Ed.), Language and Languages: Global and Local Tensions, (pp. 184-200). Newcastle, UK: Cambridge (2007)

ACKNOWLEDGEMENTS

I would like to acknowledge and thank three groups: my student/academic community at The University of Melbourne, my work community at La Trobe University and my home community of family and friends.

I am indebted to my exemplary supervisor Professor Tim McNamara who has supported my research and has maintained interest in the project until the end. His infinite professional insight and breadth of knowledge have encouraged me to complete. He has suffered the tedium of many manuscripts and has tirelessly provided in valuable guidance even when on leave.

I am most grateful to Dr Annie Brown who got me started by referring to a conference abstract in 2001 as a "thesis topic not a paper". She proceeded to generously offer guidance when she was co-supervisor and has had continued interest in the project beyond the call of friendship or duty. Thanks also to Dr Paul Gruba and Dr Mary Stevens who read late drafts and offered practical, timely advice. The moral support provided by other post graduate students, in particular Dr Lyn May and Dr Kathryn Hill was much appreciated.

Without the Spanish Program at La Trobe University there would be no thesis. Fortunately, I have supportive, interested colleagues who participated in animated discussion, task trials, devised scoring procedures and are currently using the EBB tool. I have been granted leave and have been funded to attend national and international conferences, thanks to the school and the faculty. I thank current and former staff who took part as raters: Sarah Bignell, Nidia Castrillón, Analía Diego, Marta Eppel, Dr Ana Fernández, Martha Flórez, Luis González, Dr Viqui Gras, Borja Ibaseta, Filipina Maki, Dr Isabel Moutinho and Dr Carlos Uxo. The other participants were the students: this work is for them in the name of fairness.

Heartfelt thanks go to my mother, sisters and friends in respecting the enormity of the task and the time involved. My most constant support through the daily grind came from Gilbert, my ever understanding tolerant husband and children, Dominique and Pascal: you have been patient indeed in your infinite love.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

ASLPR	Australian Second Language Proficiency Rating
CA	Conversation Analysis
CASE	Cambridge Assessment of Spoken English
CLT	Communicative Language teaching
CPE	Certificate of Proficiency in English
EBB	Empirically-derived Binary-choice Boundary-definition
ELTS	English Language Testing System
ESL	English as a Second Language
ETS	Educational Testing Service
ESOL	English for Speakers of other Languages
FCE	First Certificate in English
FSI	Foreign Service Interview
IELTS	International English Language Testing System
L2	Second Language
NNS	Non Native Speaker
NS	Native Speaker
OPI	Oral Proficiency Interview
PT	Paired Test
SLA	Second Language Acquisition
UCLES	University of Cambridge Examination Syndicate
VPA	Verbal Protocol Analysis

Chapter 1

1.1 INTRODUCTION

1.1.1 Chapter overview

The introductory chapter is divided into five sections. The first section (§1.1) includes the rationale for the thesis, the context of the study and the statement of the problem dealt with in the thesis. The second section (§1.2) defines the aim and scope of the research, establishes the research questions and outlines the two studies that address them. Section §1.3 summarizes the research design; §1.4 contains an overview of the thesis; and §1.5 summarizes the introductory chapter.

1.1.2 Rationale

The focus of this thesis is on the interaction between peers in a paired oral proficiency test, and uses an examination of videoed test discourse as the basis for the development of criteria for rating that interaction. Interaction in this test broadly means two peer-interlocutor student candidates speaking to each other while they complete a task. The term is not used in the sense of interaction between variables.

Not enough is known about what takes place in the interaction between students in a paired oral proficiency test, and as a consequence, no rating scales have been developed based directly on empirical data from observed performances of such interactions. The lack of detailed empirical information available that describes paired interaction makes it difficult to develop rating scales that adequately reflect paired candidate performance at different levels (bands), thus making it difficult for raters to 'rate' paired interaction.

Oral interviews have been the standard type of oral assessment of foreign and second languages since the inception of oral proficiency testing through the Cambridge English Examinations in the first part of the 20th century (Spolsky, 1995). There is interest in the facets of oral interviews that might contribute to features of discourse in that setting because features of discourse in interview tasks have been shown to ultimately affect test outcomes (Brown 2003). In contrast, peer interaction in pair or group tasks was only introduced as an optional part of oral proficiency testing in the

revised Cambridge First Certificate of English (FCE) relatively recently, in the late 1980s (Saville and Hargreaves 1999), making it a relatively new and fertile area for research.

This leads us to the overall purpose of this thesis: to examine peer discourse in an oral proficiency test. Research into oral proficiency testing has long investigated interview discourse (Clark, 1978). From the late 1980s it became increasingly common to include pair and group tasks in oral proficiency testing batteries, which prompted research into the new task formats. From the mid 1990s, testers were challenged to acknowledge that spoken language display in a test required a social view of performance, taking account of the bearing interlocutors had on each other during co-constructed interaction (Jacoby & Ochs, 1995). The impact of the social view of performance on testing was a fundamental theoretical challenge (McNamara & Roever, 2006), impacting on tests and research on both interviewer-candidate and peer discourse. The paired format has been made a compulsory part of a large scale international testing suite as a gradual extension of the practice introduced in the late 1980's (Weir & Milanovic, 2003), by the Cambridge ESOL exams. This resulted in more research into the construct being measured in that format, that .is peer interaction. Test developers were encouraged by research into interaction to take into account co-construction. A by-product of accepting co-construction in peer interaction resulted in the development and the validation of improved scales - to measure and report on it.

What is known about the processes involved in language testing is changing in part due to the contribution to the field made by qualitative research methods through discourse studies (Lazaraton, 2002; McNamara, Hill, & May, 2002; May, 2006). Differences in discourse pattern have been of particular interest because of their potential effect on test scores (Brown, 2005). This has led to an increase in the amount of discourse based studies being carried out to validate oral tests which use qualitative research methods (Lewkowicz, 2000). As interest in interaction in peer interaction test processes increases as the target of research (versus the more conventional interview speaking tests), the number of qualitative research methodologies reflecting this interest also increases. Qualitative research is able to

aptly describe discourse that forms part of processes such as test taking, test rating and criteria development. Those three types of processes make up this thesis. This type of research informs test validation from a discourse perspective that includes candidates and raters.

The processes involved in test taking, particularly during oral proficiency tests, are difficult to investigate. In order to make rating of these test formats fair, it is interesting for researchers to know the focus of candidates during the performance with their partner. If candidates' focus is different to that which is salient to raters of paired performance, knowing what candidates focus on would allow for better preparation of pairs for successful communication and improved test outcomes.

The role of the rater is to intentionally overhear, but not participate in the paired task. The absence of the interviewer as a face-to-face participant in a paired or collaborative test task provides a contrast to the large amount of research carried out into the test discourse of oral interviews where the interviewer *is* a direct participant. Lazaraton (2002) conducted an in-depth study taking a qualitative approach to validating the Cambridge ESOL language test. It recognized that the grouping or pairing of candidates is a test method facet plays an important role in test performance. The empirical research shows that peer interaction enables the display of a greater number of speech functions and of conversation management skills. Conversation Analysis (CA) of peer test-taker dyads in the FCE (Galazci2004) has shown conversation management and collaborative behaviour.

Aside from Lazaraton (2002) and Galazci (2004), the lack of detailed research into the peer interaction construct is due in part to the fact that researchers have insufficient knowledge about the manner in which raters or candidates construe 'interaction'. This is despite the underlying assumption that paired interaction, as a format, can be described and can be rated - given the inclusion of paired tasks in the Cambridge ESOL tests.

Rating scales can be developed intuitively, which has been the norm in education, or empirically, which has been called for but is relatively new for rating second language proficiency (Fulcher, 1997; North & Schneider, 1998). If rating scales are developed intuitively, there is a strong chance that the features of interaction included in the rating criteria may be those features considered important by experienced raters. However, whilst the criteria are considered important, intuitively speaking, they may not be the same ones that raters actually attend to while rating. Raters have been reported to arrive at the same score by attending to different features of discourse (Douglas, 1994). So, if it can be shown that raters attend to interaction when rating test discourse, then rater participation in scale development is a way to validate empirical rating scales as Brown, Iwashita, & McNamara 2005) did in their study. Features attended to by raters while scoring can then be used to validate scales for a relatively new type of test such as the paired oral.

1.1.3 Context

The rationale for the introduction of the Paired Test in the Spanish program investigated here is connected to the slow introduction of applied linguistics specialists to language departments in university settings in Australia. There is a great need to provide professional development for lecturers and tutors in language departments in the area of language testing and in particular oral proficiency testing. Employing applied linguistics specialists in language programs fills this demand for professional development because their presence stimulates discussion, curriculum renewal and reflection on assessment practices.

The introduction of a speaking test in pairs for Spanish beginners by the researcher in 2001 in the setting for this thesis was a natural progression from the Communicative Language Teaching (CLT) style already used on site. Students were consistently taught in the target language within the framework of CLT. Tasks that required pair and group work made up a high proportion of the available class time. The test reflected the tasks and the type of interaction students were accustomed to participate in, in their university language classroom. This resulted in a paired test task being developed and trialled in 2001, details of which are reported in Chapter 3. The key issue here is that experienced teachers on the Spanish language program intuitively developed the rating scale for the trial of the task. It had criteria for communicative

interaction operationalised as 'communication' and 'comprehension', both of which the teachers/raters found difficult to apply to this assessment context. Study 1 Part A, reported in Chapter 4 of this thesis, is devoted to extending and validating the meaning of the word 'interaction' by the teachers who taught and assessed beginner level students on the Spanish program.

1.1.4 Statement of the problem

Discourse analysts have studied the interactional nature of face-to-face communication in test settings (e.g. Lazaraton, 2002; Brown, 2003). Within the social dimension of oral proficiency testing all speakers are responsible for contributing to what is said, which makes it difficult to report on an individual candidate's performance (McNamara & Roever, 2006) whether in an interview or in the peer setting. Fine-grained discourse analysis of peer interaction assessment highlights performance features that result from what happens between speakers when switching from speaker to listener in a paired speaking task (Galazci2004) and in interviews (Lazaraton 2002). The difference with the peer setting is that as interlocutors, both parties are responsible for managing the discourse. Candidates are presumed to have an equal status when talking peer-to-peer, but not candidate-to-interviewer.

The problem in this context is that research thus far has not provided sufficient information about the performance features for pairs in face-to-face interaction. Without this information it is difficult to know what is being rated when candidates work together on a paired task in an oral test (Orr, 2002). Where there is insufficient empirical research into the performance features and organization of a paired format speech event as is pointed out by Fulcher (1996), rating scales and criteria cannot be expected to accurately reflect performance. Rating scales and rating criteria devised with qualitative input from candidates and raters will broaden our knowledge of the paired speech event, thereby adding to the criteria for communicative interaction currently in use.

1.2 AIMS AND RESEARCH QUESTIONS

1.2.1 Aim of study

Starting from 'the product' of paired test discourse, the aim of this thesis is twofold: 1) to examine how 'interaction' is attended to and characterized by raters and

candidates and 2) to devise an empirically-based scale for face-to-face peer interaction. The first objective reflects the fact that it is important to include the features that raters conceptualize as 'interaction' in the scale development process. If it is found that raters attend to particular aspects of interaction when rating candidate discourse, the aim will be to include those aspects. In order to fulfil the second objective, an empirical data based scale development methodology needs to be identified that is suitable for rating 'interaction'. There are no current evidence-based rating scales for peer interaction based on actual live performances, so additionally, as a result of the intended scale development procedure, it remains to be seen whether peer interaction in a task setting is in fact scalable.

Firstly, this thesis focuses on describing the construct of paired oral interaction examined from two complementary perspectives: that of the raters and of the candidates. Analysing these two positions adds to the current body of knowledge through new evidence on the manner in which 'interaction' is defined and operationalized by raters and candidates. Without such a definition and operationalization, it would be difficult to develop a scale empirically. The definition of 'interaction' will allow identification of the scope of the features that need to be covered by the scale.

Secondly, after gathering the two complementary perspectives described above, the thesis reports on an innovative manner of adapting and applying existing techniques of data based scale development. The use of paired candidate test discourse samples in the development of an empirical scale enables a different type of scale to be created for rating peer interaction. The scale will be different from those normally used in rating oral proficiency in two ways: the manner in which it is developed (directly devised by raters, from their viewings of recordings of live tests) and in its appearance and content.

In the interests of test fairness, anything that is not entirely understood could favour the scores of some candidates over others, reducing validity and affecting fairness. We do not as yet understand the full range of the performance components of peers in the pair format and while that is so, the test outcomes continue to require a judgement

from raters based on a range of criteria that may not cover all that there is to measure. In the words of Willingham and Cole (1997:228), "[v]alidity is the all encompassing technical standard for judging the quality of the assessment process. Validity includes, for example, the accuracy with which a test measures what it purports to measure, […] and comparability of the test process for different examinees". With this in mind, the peer interaction construct, which we must identify and define (Bachman 1990) before attempting to measure its display in performance, will be explored and validated from live test discourse samples within the confines of empirical data-based scale development.

In sum, the first part of the current study, in Chapters 4 and 5, describes the paired oral interaction in detail from candidate and rater perspectives in order to define the construct. The definition arrived at will contribute to developing an empirically based rating scale in the second part of the study, in Chapter 6. The scale will focus on interaction, using the discourse sample drawn from the candidate pairs as part of the development process. The features that candidates and raters attend to while observing the interaction will be used to validate the features included in the scale if those features are found to correspond.

1.2.2 Originality and scope of the study

The innovation underlying the research design of this study drives the inquiry. The main difference between this study and others is the use of mixed methods research to collect and analyse the data in order to validate an evidence-based scale (to be devised in Chapter 6) using rater orientation to features of interaction in candidate collaborative tasks. 'Orientation' in this context means the features that the raters notice about paired interaction. In this way the study avoids the limitations of other studies where scale criteria for paired interaction have been validated from researcher analysis of paired candidate discourse, not features salient to raters. Post hoc validations of paired tasks and their rating criteria have primarily used transcriptions and discourse analysis of candidate performance. Instead, in the study reported here the rating criteria are validated concurrently with the scale development, as part of the scale development process. Rating scale development and validation of the particular

criterion 'paired interaction' takes place before it is implemented on the task. Scale development and validation are enabled by the development of a particular evidence-based scale. Evidence is found in the data for a particular feature before the criteria are developed around that part of the speaking construct. By using observed interaction in videoed live test performances as its data, the scale is not developed or validated merely from transcriptions. The method employed here is innovative and crucial to the study.

There are three other differences between this study and any precursors in addition to those set out above on the type of data and the timing of the validation. Firstly, the setting was an in-house university course achievement test called the Paired Test (PT). It was not a global high-stakes test such as the Cambridge ESOL on which most of the pair task research has been carried out. Secondly, the language being studied for scale development and validation was not English, but Spanish. The third distinguishing feature of this study is that an evidence-based rating scale for interaction was developed, adding further empirical knowledge about peer interaction in tests.

The scope of the study is limited by time, language, level and location. The study took place in the last semester of a beginners' Spanish language undergraduate course in two English medium universities. The data sample was derived from a cohort of adult candidates who had taken 104 hours of language instruction in the target language at 4 hours per week over 13 weeks when the test was videoed.

1.2.3 Research approach

Qualitative research has recently been extensively used in language testing, particularly to investigate joint construction in candidate-candidate and candidate-interviewer performance (McNamara, Hill, & May, 2002). Following in that vein, the research reported on in this thesis offers an interpretation of data gathered for Verbal Protocol Analysis (VPA) (Gass & Mackey, 2000) which is then used for test validation. To do so the qualitative data is combined with the quantitative. As mentioned in 1.2.2 above the use of mixed methods research is one of the innovations

of this thesis. For the purposes of this thesis the definition of mixed methods research from Creswell &Plano Clark (2007:.5) is set out below

> "Mixed methods research is a research design with philosophical assumptions as well as methods of inquiry. As a methodology it involves philosophical assumptions that guide the direction of the collection and analysis of data and the mixture of qualitative and quantitative approaches in many phases of the research process. As a method, it focuses on collecting, analysing, and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone."

Different interpretive accounts of the data from the paired task testing context are drawn from three groups: the candidates, with retrospective Verbal Protocols; the raters, with think alouds and scale development; and the researcher categorizing and interpreting the data.

Each qualitative step is ultimately related to a quantitative result. The verbal protocols transcriptions after coding, result in a taxonomy for interaction as viewed from two perspectives: that of the candidate and the rater. Both of these taxonomies are built on sequentially to result in a rating scale, an instrument devised for rating paired interaction. This type of design in which a qualitative phase is used to build a quantitative instrument is known as an 'Exploratory design' in mixed methods research (Creswell & Plano Clark 2007:79).

In testing research Fulcher (1996) used qualitative research to build theory for the development of evidence-based fluency scales. He employed an iterative process of interviewing, transcribing, coding and interpreting in order to obtain a 'thick description' within the Grounded Theory framework (Glaser & Strauss, 1967). The evidence-based scales reported on in this thesis derive input from and are validated by

VPA (Ericsson & Simon, 1993), which, although different to Grounded Theory, remains an iterative process, which also culminates in a coding system. The coding system used to validate a scale for paired interaction also adds to theoretical knowledge about the construct of paired peer interaction. The results will offer a better understanding of the nature of the interaction between paired peer candidates in this type of testing context.

1.3 THE STUDIES

Before directly focusing on the actual practical questions directing the studies two overarching questions that guide the research in a global sense need to be considered. Firstly, whether the process of 'communicative interaction' (as it is called by UCLES), is a construct that can be adequately operationalized for raters to "understand the model of communicative ability on which rating scales are based" (Orr 2002:153). Secondly, whether the communicative interaction construct is scalable in the same manner that linguistic abilities have traditionally been scaled, into band levels with accompanying descriptors. These two questions indicate the context of the rationale for the research questions stated for each study below.

1.3.1 Perceptions of peer interaction in a paired test (Study 1)

Study 1 is an examination of what L2 Spanish tutors and candidates themselves claim to notice as interaction on paired tasks. It examines the nature of language teacher raters' and candidates' orientation to 'interactional features' of test discourse. It also examines raters' and candidates' perceptions of what interaction consists of, by focusing on what both parties attend to while observing pairs performing a paired oral task in a beginner achievement test in Spanish.

Study 1 aims to define the construct of interaction as a necessary step before the development of a scale for rating interaction. The two research questions guiding this study are as follows:

> **Research question 1:** What features of peer interaction do raters attend to in paired task test performance?

Research question 2: How do candidates view interaction in a paired oral?

The study will result in a framework for describing candidate and rater ways of conceptualizing interaction in this setting. It will also offer empirical evidence of the raters' perception of interaction, a necessary step for the resulting rating scale that they will use in operational settings.

1.3.2 Developing a data based empirical rating tool for peer interaction (Study 2)

Study 2 reports on the development by raters of a data based empirical rating scale for peer interaction. The findings for the research questions in Study 1 will provide an initial framework of raters' conceptualization of performing in a paired oral. Material drawn from this initial framework will be used to confirm or otherwise, the findings of the data based scale in which the raters themselves are scale developers.

The type of rating procedure used is derived and adapted from the method known as Empirically-derived Binary-choice Boundary definition (EBB) (Upshur & Turner, 1995). This methodology uses samples of candidate performance to elicit judgements of the difference *between* levels. The levels are defined in terms of a yes/no question - known as a criterial question because the criterion used to separate levels is embedded within it. The results from Study 2 will generate an empirically developed rating procedure, relevant to the performance of tasks in pairs. This EBB procedure will reflect the interactional features attended to by trained raters of beginner oral Spanish in their observation of peer test performance. The EBB will be developed from the raters' direct observations of live test performance. In Study 2 of the thesis the research question that forms the basis for the scale development is:

> **Research question 3:** Can candidate peer performance samples from a paired test form the basis for developing a rating procedure for interaction?

If both candidates and raters focus on similar features within the paired interaction, and these are also found in the steps or hierarchical levels of the EBB scale, it will be argued that the empirically developed scale is valid. It should be pointed out that while Turner and Upshur (1995) call their procedure a scale, it works quite differently

to band scales. In this thesis 'rating procedure' and 'EBB scale' are used interchangeably.

1.3.3 Research design

Table 1 sets out the range of data, the size of the data set and the type of analysis conducted for each study. The stimulus data used in each study is gathered from a university language program where paired candidate oral tests are videoed at the end of a teaching semester. The videos are observed by raters and by candidates who produce Verbal Protocols. The discourse from both sets of Verbal Protocols is transcribed for qualitative content analysis.

Table 1. Methodology Study 1: orientation to paired interaction

Study 1	Stimulus materials	Resulting data	Analysis of data
Part1a	17 x 10 min. videos of operational paired orals	12 Spanish L2 specialists, working in pairs, observe and comment on 3 pairs each	Transcription of VPs
Part 1b	17 x 10 min. videos of operational paired orals	12 x 3 reports, as above	Content Analysis of Rater Verbal Protocols
Part 2a	17 x 10 min. videos of operational paired orals	25 individual L2 learner candidates perform retrospective Stimulated Verbal Recall on their own paired performance	Transcription of reports
Part 2b	17 x 10 min. videos of operational paired orals	25 reports, as above	Content Analysis of candidate Stimulated Verbal Protocols

The site for Study 1 is a university language program. Study 1 utilizes stimulated Verbal Protocols (Ericsson & Simon, 1993; Green, 1998) to explore the orientation towards interaction of the two groups of participants, raters and students. The 12 raters in this study are specialists in teaching and rating Spanish as a Foreign Language (Table 1 part 1a and 1b). In the Stimulated Verbal Protocols, the rater participants verbalize what they attend to when observing videos of a subset of 17 pairs of students in a paired oral proficiency test. The aim is to uncover rater (and subsequently candidate) orientation to the construct 'paired peer interaction'. The resulting insights will highlight the key features that reflect the view of the participants (the 12 raters). Then, in Study 2, through a set of processes from

individual, team and consensus building, a rating procedure will be developed to reflect the responses to the paired performances.

In the second part of Study 1, a Stimulated Retrospective Verbal Recall (Ericsson & Simon, 1993; Gass & Mackey, 2000) is used. This task involves 25 candidate participants who have already taken a similar test once before the previous semester. Considered 'experts' (Brindley, 1998), they are asked to individually comment on their own performance within the paired interaction in the Paired Test (see Table 1, part 2a and 2b). The intention of the VP is to note the salience of any particular issues candidates attended to while taking part in the interaction. By exploring candidates' comments, this second part of Study 1 provides a counterbalance to the features that raters noticed when observing interaction from the same set of pairs. The comments will be compared to see if raters and candidates find the same features to be salient regarding interaction in the paired oral (Table 1 part 2).

Study 2, in which a rating procedure for interaction is developed, has two parts: a data reduction and selection procedure, and the scale development process, both shown in Table 2 below.

Table 2. Methodology Study 2: developing an EBB rating procedure for paired interaction

Study	Stimulus materials	Resulting data	Analysis of data
Part 2a	17 x 10 min. videos of operational paired orals	Transcriptions of rater VPs of operational paired oral	Data reduction of VP key features presented as Protocol Feature sets for pairs used in procedure
Part 2b	8 selected 10 min. videos of operational paired orals and 8 Protocol Feature sets	3 teams of L2 specialists empirically develop EBB rating procedure	Integration of EBB procedures developed by each team of raters

In terms of Part 2a, the key features from the VP transcriptions are summarized to reduce the data. These sets of reduced data, called Protocol Feature sets, make the key

features of the VP accessible for raters, if necessary, during the scale development procedure. In the interests of a broad representation of performance types, without an unmanageable amount of data, neither all the Protocol Feature sets, nor all the paired candidate performances are used to develop the scale. Instead, eight pairs (16 candidates) and eight corresponding reduced Protocol Feature sets are selected. This methodological step is further discussed in Chapter 6.

1.4 THESIS OVERVIEW

This section provides an overview of the eight chapters that make up the thesis. The first three chapters set up the thesis context. The middle three present the studies. The last two connect the findings and present the conclusions that are drawn.

Following the present chapter, Chapter 2, a literature review, presents the relevant research on testing speaking in pairs and groups, on issues in rating discourse and issues regarding scale development. Chapter 3, the last chapter in the introductory part of the thesis, includes information about the site for the study, details about the Paired Test on which the study is based, the research agenda for the thesis and an overview of the methodology.

The middle three chapters in the thesis contain the empirical studies on the orientation to the peer interaction construct, by raters and by candidates, and the development of the EBB evidence-based rating procedure. Chapter 4 describes Study 1 Part A, which focuses on the orientation by raters to peer interaction through an elicited Verbal Protocol. After a content analysis, the findings on rater orientation to peer interaction are presented. In Chapter 5, Study 1 Part B describes orientation to peer interaction through candidate Stimulated Verbal Recall and presents findings on candidate orientation to peer interaction. Chapter 6 contains Study 2, which describes the development of an EBB evidence-based rating procedure for paired interaction. Firstly, there is an adaptation of the EBB methodology for the context of Study 2; secondly, an EBB rating procedure is developed before the findings are presented.

Chapter 7 discusses the findings, linking those that emerged from the two different studies. It reviews the studies and research questions before presenting a synthesis of findings and implications. Chapter 8, the final chapter, concludes the thesis.

1.5 CHAPTER SUMMARY

The overarching aim of this research is to describe those 'things' that take place in a paired interaction (Fulcher, 2000), and to gain insight and understanding of them. This chapter outlined how for this thesis, the construct of interaction will be empirically defined, and, based on this information, an empirically based rating scale will then be developed and validated. Such a scale could potentially be used to rate the effectiveness of peer paired communicative interaction during speaking performances in Paired Tests in beginner Spanish.

Chapter 2: ISSUES RELEVANT TO ASSESSING SPEAKING IN PAIRS

2.1 INTRODUCTION

Rating scale development for peer discourse needs to acknowledge findings on oral interview discourse. The reliability of discourse in interviews has been criticised (Lazaraton, 1996). For example, oral proficiency interviews are often highly dependent on the interviewer and thus the interviewer can affect the result. Tension remains between the need to elicit a variety of functions during either interviewer/candidate interaction or peer interaction, and the need to validate the inferences made from a test. Here the tension is caused by the different tasks which elicit different discourse and the way in which the differences can be acknowledged in the rating procedures.

The introduction of the paired format, and subsequent research into that format follows a long history of empirical research on oral interviews. When peer tasks became part of oral proficiency tests they reflected Communicative Language Teaching in the classroom and in part they also compensated for:

- the power differential between candidate and interviewer in the interview situation (Young & Milanovic, 1992; Skehan, 2001);
- the native/ non-native speaker's influence on discourse in the interview (Ross & Berwick, 1992); and
- the short fall in interactional functions elicited by interviews (Perret, 1990; Hatch, 1992; Johnson, 2001), which means that candidate and interviewer had different roles in the interaction and thus displayed different functions.

Despite the power of the interviewer, the native speaker influence on the discourse and the narrow range of functions, the interview was maintained as an assessment tool because it was efficient and practical. Nonetheless, it was occasionally accompanied by a paired or group task.

The introduction of the pair format addressed issues raised by research into oral interviews. Including 'communicative interaction' as a criterion for both interviews and paired tasks in an oral test acknowledges the direction that research on interview discourse has taken. Critics of the paired format show concern for mismatched proficiency levels (e.g. Norton, 2005) and for the lack of expertise of a peer interlocutor (van Lier, 1989), both of which raise rating issues for the interactive communication that takes place. Ratings mechanisms developed specifically for the paired task should reflect the implication of both parties in interaction.

Empirical scales have been called for, specifically for speaking scales because "most holistic and other analytic scales lack firm empirical substantiation in respect to evidence about L2" (Cumming, Kantor, & Powers, 2002:68). In addition, taking into account that it is claimed that peer tasks permit testers to test "complex constructs" (Fulcher, 2003:189) then necessarily, this complex construct needs to be identified and described. How does the construct compare to that already tested in interviews? And if it is 'complex', in what way is it so? There are suggestions in the literature that the speaking construct in the paired format involves non-verbal communication (Orr, 2002), listening comprehension (Pollitt & Murray, 1996) or the ability as peers to change position from listener to speaker (Galaczi, 2004). This is not to say that these do not also apply to interviews. The issue is whether listening, non-verbal communication and the changing of roles apply differently in the paired format. If the construct *is* different this should be reflected in the criteria for the paired format.

Rater cognition studies (e.g. Pollitt & Murray, 1996; Brown, Iwashita, & McNamara, 2005), and other related scale validation studies (e.g. Meiron, 1998; Brown, 2000), have so far looked at interviews but not at peer interaction. Galazci(2004:265) describes paired discourse in the First Certificate in English (FCE) and explicitly recommends that further research could come from raters and test-takers themselves in the form of "think aloud protocols [which] would provide valuable insights into understanding the issues at play in test-taker interaction." This thesis represents such a project.

2.1.1 Chapter overview

This chapter reviews relevant literature on speaking tests focusing in particular on the paired test format. It is made up of five main parts.

Part one (§2.1) focuses on testing speaking in pairs, and is presented in six sections: the background to testing in pairs, the origins of testing in pairs; group oral speaking tests; affective support of paired and group oral interaction; paired and groups tests in use and finally issues to consider with the paired format.

Part 2 of the literature review (§2.2) covers the interlocutor effect on scores in speaking in four sections: the proficiency effect, the familiarity effect, the personality effect and other effects.

In §2.3 of the literature review the interlocutor effects on discourse are presented in sections covering peer task effect on language sampling, increased functions in paired or group orals and conversation management in peer tasks.

§2.4 focuses on studies on rating scales for paired or group interaction, while §2.5 focuses on developing rating scales for communicative interaction in pairs.

§2.6 evaluates the significance of the literature reviewed to the studies to be presented in this thesis, before the chapter summary in §2.7.

2.1.2 Background to testing speaking in pairs

The background on testing speaking in pairs begins with a survey of the changes in views on what 'speaking' in tests consists of, from 1980 to the present. These changes in views are most apparent in the assessment of speaking in pairs in international tests and are connected to the increase of paired speaking tasks in the language classroom.

The shift from the unenlightened view that speaking in a second language (L2) generally meant information transfer, to the acknowledgement that speaking involved negotiating meaning (Savignon, 1983) entered Second Language Acquisition (SLA), teaching and testing in the 1980s. In SLA, speaking had broadened to include communicative competence (Canale & Swain, 1980), which, apart from the grammar rules of linguistic competence also made room for sociolinguistic competence, discourse competence and strategic competence. In second language teaching, changes in the classroom meant that pair and group work reflected the move towards developing communicative competence within the teaching approach known as Communicative Language Teaching (CLT) which was a more enlightened view of teaching 'speaking' as a communicative skill. In the model representing second language proficiency used for second language testing, Bachman (1990) presented communicative competence in a framework with two overarching categories: organizational competence and pragmatic competence. The changes in L2 language learning, teaching and testing referred to above all have an impact on research into spoken interaction.

Communicative competence for a second language learner is the ability to speak correctly *and* to be situationally appropriate in spoken interaction. This differs from plain information transfer. Spoken interaction is a key term in this study, and an inherent part of the communicative paradigm in second language learning, teaching and testing. The point is that language educators and testers have had up till now second language acquisition models that include communicative competence but not 'two-way interaction'. Communicative competence was defined in 1980 by Canale and Swain as being made up of four parts:

1. Grammatical competence
2. Sociolinguistic competence
3. Discourse competence
4. Strategic competence

These four divisions were realigned into three groupings by Bachman (1990): organizational, pragmatic and illocutionary competence. Other theories of second language performance also had varying models of communicative competence, which

included oral discourse (Hymes, 1972; Canale & Swain, 1980; Canale, 1983; Celce-Murcia, Dornyei, & Durrell, 1995; Bachman & Palmer, 1996), but not two-way interaction (McNamara, 1996), as a component of second language learners' communicative ability.

In the 1990s when research into speaking test discourse between interviewers and candidates started to emerge, it showed that two-way interaction had implications for testing speaking. In particular, interlocutors affected each other's performance while speaking, and this needed to be taken into account. This research in the 1990s was a result of "the increasing influence on applied linguistics of Conversation Analysis, triggered by the intellectual synergy within the applied linguistics program at UCLA" (McNamara, Hill, & May, 2002:222). What is important to note is that at that time, Bachman's (1990) model for language testing theory included communicative language ability but not 'two-way interaction'. As pointed out by McNamara (1997), Bachman's model lacked the socially constructed nature of performance, which was under scrutiny. Beyond the inclusion of communicative ability in models of second language acquisition (the first step towards the recognition of context and situated speech for the co-construction of dialogue) any further discussion of the differences between these models falls beyond the scope of this thesis.

With 'two-way interaction' recognized by language testing researchers, we move now to one strand of the current literature on discourse in oral proficiency testing which focuses on paired peer discourse in particular. Our literature review focuses on these areas:

- The origin of tests in pairs
- Interlocutor effects on scores
- Interlocutor effects on discourse
- Developing rating scales
- Assessing communicative interaction

The situation in essence with the paired format is that if only an interview is used to test oral proficiency the interviewer/interlocutor effect prevails, but if pairs are used to

remedy the first problem then a different issue arises: that of rating peer discourse. It was demonstrated by Brown (2005) that the interviewer/interlocutor effect arises in Oral Proficiency Interviews (OPIs). The issue remains the same for pairs, as raters must disentangle the contribution of the two partners - peers or interviewer/candidate - in either test format. It's perhaps more practically relevant in peer tests as both interlocutors have to be rated, and the issue arises more consciously or explicitly. This conflict begs the question: why use the paired format at all? To demonstrate that researching paired discourse is worthwhile, the origin of tests in pairs is first presented, in §2.1.3, before taking up the two points that lie at the heart of the issue: the many interlocutor effects that impact scores (§2.2) and discourse (§2.3) in tests, and the resulting difficulties in rating pairs (§2.4).

2.1.3 The origin of tests in pairs

The critical question 'What is speaking?'(Fulcher, 2003), as in 'speaking' in a test, was seriously considered as a research issue for oral proficiency tests at a time when the only oral test that had been extensively researched was the Foreign Service Institute (FSI) Oral Interview. It was claimed by its proponents that this interview involved 'natural context' and 'real life' tasks, which together produced 'natural conversation' (Wilds, 1979). In response to such claims, the early debate in oral proficiency testing was whether the dialogue between interviewer and candidates in an interview, loosely called 'the conversation', had any parallels with actual non-test casual conversation (Young & Milanovic 1992). Further research identified many types of interlocutor effects that were a result of interviewers and candidates producing discourse together. The motivation for paired testing is based on the fact that when two candidates speak to each other in a peer test task, the result is 'test discourse' or institutional talk, (just as when an interviewer speaks with a candidate in a test) and although this pairing does not represent a casual conversation (van Lier, 1989) it allows for different interactional moves than in an interview.

A survey of the literature on issues arising from oral language testing is presented in Spolsky (1990, 1995), where the first face to face language tests held in the UK - the Certificate of Proficiency in English (CPE) - are described. Introduced in 1913, the CPE was the first test that had a sub-test of spoken English (Fulcher 2003:5). It was

followed by the First Certificate in English (FCE) in 1939, a requirement for foreign students for entry into English universities.

It was in the 1980s that UCLES gave candidates the option of taking a paired oral, and in December 1996 that task became compulsory (Saville & Hargreaves, 1999), as part of a revised test format of the Cambridge First Certificate of English (FCE) and the Cambridge Proficiency in English (CPE). The test revision was "the outcome of a rational process of test development" and the aim of the new design was to "provide improvement in the assessment of speaking" (Saville & Hargreaves, 1999:42). At that time there were other speaking tests with a group format (taken up in §2.1.4).

The improved oral included a paired format with two candidates and two examiners. As a result, the main processes of oral testing, elicitation and rating were both affected. Rating was affected because of the presence of two examiners: one as the assessor, providing an analytical assessment, and the other as the interlocutor, providing a global assessment. Elicitation was affected because during the oral there are different phases where the interlocutor, the assessor or the other candidate, provide different patterns of interaction between them, as in candidate/candidate or interviewer/candidate interaction. In fact, this was the principal reason for revising the FCE and for the inclusion of the pair format in the Cambridge suite of tests: it "allows more varied patterns of interaction during the examination" (Saville & Hargreaves, 1999:46).

Progressively this format was made compulsory so that between 1991 and 2002 all the tests from the Preliminary to the Proficiency had the compulsory paired/group format. Although the modifications to examinations that resulted took place in an "evolutionary way" (Saville & Hargreaves, 1999:42), these changes were validated by research (Lazaraton 1996, 2002; Milanovic, Saville, Pollitt and Cook 1996; Young and Milanovic 1992; Weir 2003). As the pair format was more widely used, it was also more extensively researched, though until recently very little of the research was on the discourse that ensued from Cambridge paired interaction (Galaczi, 2004; Nakatsuhara, 2004, 2007).

Studies on the pair format can be compared with those on research into discourse in interviews where turn taking, topic organization, sequence and the overall structure are predetermined or controlled by the interviewer. Drawing on the results of her comparative study of paired and individual high school English orals in Hungary, Csepes (2002), refers to the differences between the interaction styles of interlocutors in interviews and in the paired modes in terms of who has control over the interaction. The interviewer control in the interview and lack of it in the peer task significantly influences discourse in terms of communicative interaction and discourse management.

2.1.4 Group oral speaking tests

As noted above, the group orals were experimented with and introduced in school and university settings in the late 1970s and early 1980s. There exists, in fact, little research into group orals because they have not been widely used as a measure of oral proficiency. Nonetheless, the group oral for English was experimented with at secondary school level nationally in both Zambia (Hildson, 1991) and Israel (Shohamy, Reves, & Bejarano, 1986) and at university level in both Finland (Folland & Robertson, 1976) and Hong Kong (Morrison & Lee, 1985) and for university entrance (Swain, 2001). The Cambridge suite has the possibility of a paired or group format depending on the number of candidates that need to be examined.

A survey of the research shows that group orals, as a precursor to the paired format used by Cambridge, were not new to language testing. A summary of the research into secondary and university level group orals shows the rationale behind their use claims that group orals:

- are cost effective (Hildson, 1991),
- are linked to claims of positive washback effects on the classroom (Hildson, 1991) (where washback is the effect of a test on teacher and learner behaviour (Messick, 1996),
- as a test task, are representative of what is required in class (Morrison & Lee, 1985), and

- elicit a broader range of language and discriminate better between levels (Shohamy, Reves & Bejarano, 1986).

The group oral was embraced for all the practical and affective reasons put forward by the research listed above. Later research findings from discourse studies on interviews and the paired format made it possible to start to gain a better understanding of paired interaction, as opposed to early studies that focused on affective factors and the effect the peer format had on scores.

The two previous points of broader speaking styles and problems in scoring groups were taken up by larger scale research into the effect of grouping on test outcomes. Shortly after the group test was implemented in Zambia, Shohamy, Reves and Bejarano (1986) were trying to find an alternative to the existing oral interviews for English matriculation in Israel. Because of the lack of research on tests apart from the Foreign Service Institute (FSI) oral interviews mentioned above, they trialled students on four different oral tasks and compared the results

One of the four tasks, the group oral, with 4 students per discussion group, had previously been trialled (Reves, 1981). As part of the trial, between 1980 and 1981 in 18 schools all over Israel, students had been given the option of a repeat examination - with an interview. The findings were that there was only a 50% consistency between the interview and the group oral, with 30% of participants raising their marks by one grade and 20% of them deserving lower marks. This inconsistency meant more conclusive evidence was required following this trial.

In the study that followed (Shohamy, Reves and Bejarano, 1986), a role-play and a reporting test were included alongside the group discussion and the oral interview. The study consisted of 103 matriculation students completing all four tasks before taking the interview for the matriculation exam. The marks were subsequently compared. The study concluded that students completing a variety of tasks were tested for a broader range of oral speech styles and that the test with four tasks "discriminated properly between the various levels of oral proficiency" (Shohamy, Reves, & Bejarano, 1986: 217). They arrived at these conclusions by reporting a

comparison between all four tasks, and an overall comparison with the matriculation exam results. The relationship between the four experimental tasks and the matriculation examination was found to be relatively weak. Separate reporting on the group discussion showed rater reliability to be 0.73. Of the four tasks, the group discussion had the lowest mutual variance with the other tests. In the relationship between the experimental test and the final exam the group oral had the lowest mutual variance at 0.20, (which meant it would be hard to then compare the data from this test with other test formats). This weak relationship between the four experimental tasks would "statistically justify the need for separate tests to tap different aspects of oral proficiency" (Shohamy, Reves, & Bejarano, 1986:217). The matriculation test was changed as a result of the research.

So far in this section we have seen the historical background to the introduction of testing in pairs. What follows is a survey of the literature that supports (§2.1.5), describes (§2.1.6) and questions (§2.1.7) this format.

2.1.5 Support for paired and group interaction

Early research on paired orals supported the use of paired orals because they were positively appraised by students (Scott, 1986; Fulcher, 1996; Humphry-Baker, 2000; Egyud & Glover, 2001). Later research on paired orals moved on to validate the paired format, through an increasing number of empirically based studies - some of them discourse studies.

Tests naturally cause anxiety and are not pleasant experiences for the candidates involved. Early research into pairs made many claims as to why paired orals should be used, based on the tests' face validity. Face validity or test appeal (Bachman 1990) refers to how well test takers or apposite stakeholders perceive the test relates to the skill being tested. Such affective reactions were usually collected by questionnaire or by interview. Studies on the affective reaction of candidates towards paired and group orals reported positive attitudes and reduced anxiety. In a paired or group format, it is reported that there is:

- Increased motivation for students to speak in a group in English (Folland & Robertson, 1976)
- A reflection of best practice in the classroom and a greater fit with teaching using a group oral (Morrison & Lee, 1985) and using paired orals (Taylor 2001, Egyud and Glover 2001)
- A belief by students that the pair tests are an accurate representation of their level (Scott, 1986)
- A willingness and a positive attitude towards collaborating in the paired or group test (Scott, 1986; Fulcher, 1996; Humphry-Baker, 2000; Egyud & Glover, 2001)
- Cost efficiency (Swain, 2001)
- Positive washback on classroom teaching (Hildson, 1991; Taylor 2005)
- Reduced anxiety for the test takers (Iwashita, 1999)

Increased use of pair or group work in language classrooms enabled a smooth "move from learning exercises to test exercises" (Messick, 1996: 241). The early group test studies in the eighties reflected the recognition of the importance of interaction and negotiation of meaning in SLA research (Long, 1983; Gass & Varonis, 1985). However, the studies on pairs tell us little about the type of interaction or how the tests were scored, calling for empirical research to validate the tests.

2.1.6 Paired and group tests in use

Much of the published research on paired speaking tests is on part 3 of the First Certificate of English (FCE) of the Cambridge ESOL speaking tests. To put the research into the paired format in context, we summarize the current practice in Cambridge ESOL (English for Speakers of Other Languages) speaking tests, using the FCE as an example.

In these speaking tests there are typically two assessors and two candidates, although sometimes there are three candidates, for practical test administration reasons. The task format consists of four parts. In part 1 the interlocutor interviews candidates. In part 2 candidates produce a long extended discourse turn from pictures, which they compare and contrast. Part 3 is the collaborative or pair task, which is only a four-

minute component of the longer test. In it candidates use visual material to produce negotiation and turn taking. Part 4 is a discussion task in which the pair format is extended into a discussion with the interlocutor. The assessors give two kinds of ratings: the interlocutor provides a holistic score and the observing rater provides analytic ratings which include five assessment criteria: grammar, vocabulary, discourse management, pronunciation and interactive communication.

All the literature discussed in this section has positively appraised the paired format, whereas studies that have looked at the discourse have revealed issues with this format, as discussed below.

2.1.7 Issues with paired format

Interview discourse studies, for example Brown (2003), report on the affect speakers have on each other during an interaction. The manner in which speakers affect each other may cause variability in test results because, depending on whom a candidate speaks to, the performance could be rated as better or worse. Inconsistent results due to this variability would not allow tests, whether interview or peer, to make reliable claims about proficiency. This issue causes a tension between the desirability for 'interactive communication' in the paired format, and the ability to validate the inferences from test scores (Taylor & Jones, 2001).

Unpredictability should be acknowledged as a side effect of interaction. In interviews there are solutions (such as scripting interviews) so that where necessary, unpredictability in the discourse between two interlocutors can be restricted. Control is gained by requesting all rater interviewers to follow prompts in order to use exactly the same wording with each candidate. This has been the case with the IELTS since 2001 (Taylor and Jones, 2001), which was redesigned from a semi-structured conversation-style interview to an almost scripted, much more structured format. These changes were made in response to concerns about the lack of consistency in behaviour between interviewers, which could advantage or disadvantage candidates (Taylor 2000). Those changes were generally well accepted by the raters (Taylor and Brown 2006).

The pair task cannot be scripted however, as the rater focus is on the way candidates manage the interaction. Just as discourse varies when candidates participate in an unscripted interaction with their examiner (Clark, 1978; Bachman, 1988; Stansfield, 1992; McNamara, 1996), variability is also likely to occur when candidates elicit language and interact with each other while performing an unstructured task such as the paired format.

In sum, we have seen that speakers affect each other when performing a paired oral for a test, leading to variability in performance, and hence scores. This is of concern to testers. However, this 'unpredictability' (Morrow, 1979; Fulcher, 2000) in the way each pair deals with a topic and how changes take place between them should be considered in light of the fact that a candidate's next utterance is bound by what has been said previously. From Conversation Analysis we know that all interaction is highly structured from this point of view. Although it appears that candidates are able to say what they want, how and when they want, in fact they are limited by the task, by their language level and the time factor. Candidates are necessarily implicated in the construction of each other's output: "the language used by the participants at any given point in the communicative exchange affects subsequent language use" (Bachman and Palmer, 1996:55).

It is not surprising that despite the affective support for paired format, Foot (1999) was less than favourable in his appraisal of paired orals, particularly in the paired context of the UCLES exams (see also Norton, 2005). Foot's greatest concern was the issue of mismatch in proficiency levels of the candidates and its effect on their scores. However, empirical research has shown that mismatch in proficiency levels of candidates is not a cause for concern regarding scores (as detailed below in §2.2.1).

Foot (1999) also takes issue with regard to low-level candidates being paired, arguing that two struggling inexpert users cannot reach their true potential. However, this depends on how 'true potential' is defined. Is it the result of a very accommodating interviewer who makes the candidate appear better or worse than he or she is? Or is

‘true potential’ managing communicative interaction at whatever one's level regardless of the proficiency and or training of the other interlocutor? Regardless of whether the paired format is used alone or as part of a test battery, research has shown that while candidates speak more when paired with a higher-level candidate, the score was not higher (Iwashita, 1999). So candidates speaking to an evidently higher-level interlocutor are not necessarily going to achieve a higher mark.

The position Foot (1999) takes is that for struggling low-level candidates, even in a test, the co-construction of the performance requires 'scaffolding' provided by a trained interviewer ('scaffolding' is a term from SLA, implying supporting a second language speaker through a conversation (Gibbons, 2002). The amount and the quality of the ‘co-construction’ between the candidate and the rater interviewer was the focus of investigation in Brown (2003), based on IELTS interviews. Brown (2003) examined variation in interlocutor discourse by comparing a supportive teacher-like examiner with a casual and less supportive examiner. Evidence of differing amounts of support given by examiners is found in the discourse. The result is a variation in the mark awarded due to a different style of interaction (caused by differing co-construction or scaffolding) with the candidate. The implication of Brown (2003) is that the scaffolding that Foot (1999) suggests is required by lower level candidates is not always helpful to the candidate because interviewers differ in the way they accommodate the candidates. A lower mark because they are not scaffolded well would support Foot’s (1999) position, but in a peer task there is no 'expert' Native Speaker (NS) to scaffold - so pairs are in the same position. Nonetheless, the amount of accommodation between pairs in peer test tasks and the effect on scores is an unresolved empirical question.

Continuing on the point of accommodation, Foot (1999) was also concerned that a pair of test-takers as interlocutors are not as successful as trained interviewers. As shown in Brown (2003), one of the identified causes of unwanted variability with trained raters is that the amount of scaffolding differs between interlocutors in unscripted interviews. The issue then would be not whether a rater or a peer is trained or not and better for performing with in an oral, but whether the rater or peer is more or less supportive in the co-construction of the display talk in the test. As yet it is

impossible to know whether variation in the amount of scaffolding from accommodation is greater in pairs than in an interview format.

An additional point to consider in relation to proficiency mismatching is Lazaraton's (2004) comment about the Cambridge suite. In this suite, the test batteries are divided into language levels, and it is thus unlikely that very disparate pairs would arise. Lazaraton (2004) points out that there is not a paired option because students of a wide range of levels sit it on demand, which would be of concern for matching levels in a paired task in the oral.

The findings for pair dynamics reported by Storch (2001), Galazci (2004) and May (2006) demonstrate that different kinds of relationships exist in pairs in the SLA context and the testing context. Should raters have the choice of giving separate marks for interaction to reflect the communicative reality? This is known as the separability issue and is addressed in Chapter 6.

We are left with the undeniable inherent unpredictability of communicative interaction. While a task in a pair is adaptive to a degree, could it become potentially a source of measurement error? Could the undeniable inherent unpredictability of communicative interaction "easily jeopardize the fairness and the generalizability of conclusions" (McNamara 1996:3) that are reached about the candidates in the pair? How great is the concern about the validity of the assessment when we use another candidate as interviewer?

With these issues in mind, we move on to the literature on interlocutor effects on scores in speaking tests.

2.2 INTERLOCUTOR EFFECTS ON SCORES IN SPEAKING TESTS

The main issues dealt with in this next section are: the proficiency effect (§2.2.1), the familiarity effect (§2.2.2) and the personality effect (§2.2.3). Some remaining issues concerning interlocutor effects on scores are described in §2.2.4.

2.2.1 The proficiency effect

In addressing the question of fairness, research (e.g. Csepes 2002) examined the effect on scores of pairing candidates of different proficiencies. This is known as the proficiency effect: would the type of candidate you spoke to affect your score?

When the pair format was introduced as an option in the FCE, differences of ability in paired candidates was a concern mainly because there was not much published research validating it. Foot (1999:38) observed that teachers were "concerned about how differences in the ability of paired candidates can affect performance". This same point was raised by Iwashita (1999) for paired orals for Japanese at university level, by Csepes (2002) for pairs in English orals at high school level and most recently by Nakatsuhara (2004) for the Cambridge paired format

In a study of 24 university students (12 male and 12 female) learning Japanese as a foreign language, Iwashita (1999) found that between mixed level (high-low) dyads, more negotiation occurred when a learner was paired with a learner of a higher proficiency than with a learner of their same level. The learners were also found to increase speech output because the conversations lasted longer when paired with a higher proficiency partner. In those pairings the quantity of speech was affected but notably, the score was not.

Csepes (2002) paired 120 candidates from 10 different high schools in Hungary with partners of the same level, as well as higher and lower levels. The results similar to those reported by Iwashita (1999): no statistically significant difference was found between the scores of the candidates in examination conditions with higher or lower or matched level proficiency partners.

Nakatsuhara (2004:55) used 24 students in a simulated CFE pair format paired at different levels and found that students "were likely to obtain rather identical opportunities to display their communicative abilities with the use of similar conversational styles" regardless of their being paired with someone of higher or lower proficiency. The proficiency level was decided based on results of commercial

tests, which students had taken beforehand. The study had 12 same-level pairs and 12 different-level pairs. Nakatsuhara (2004) found that there was only a slight impact on discourse and outcomes for pairs of the same, or of different, proficiency levels. The explanation offered was that some candidates support their partners in dyadic interaction, and this type of accommodating behaviour could contribute to the balance found in the conversation data.

In conclusion, the three studies reported here demonstrate promising support for the argument that there is no significant difference in candidate performance whether proficiency level is matched or unmatched.

2.2.2 The familiarity effect

Tests are stressful. The study of the effects on performance of someone you know is known as the familiarity effect. With reference to whether peers knew each other or not before performing the test, Foot (1999:37) was concerned whether "there are actually two different tests; one if the candidate is a friend and a different one if the other candidate is a stranger".

An early study on the familiarity effect is Iwashita's (1999) study, which reported that English medium university students felt less anxious when paired with their classmates instead of with an interviewer in a test of Japanese language. In a questionnaire "performing tasks with a non-native rather than a native speaking interlocutor created a non-threatening environment and made test taker feel more relaxed" (Iwashita 1999:62). In addition to the perception questionnaire, in the vein of much of the early affective reaction research into paired orals, this early study looked at the two areas that were to be most researched with regards to the paired format: scores and discourse. Through scores, discourse transcription and questionnaires the study examined whether the proficiency of a non-native speaker interlocutor had any impact on the score assigned or on the nature of the discourse produced during task performance. Twenty candidates (10 high- and 10 low-level Japanese learners) took the test twice with a Non Native Speaker (NNS) interlocutor of a different level to the participants. The findings showed that "subjects' anxiety rate and confidence level in relation to the proficiency of interlocutors affect the assessment scores and the

amount of talk differently" (Iwashita 1999:62). The learners talked more with a higher-level pair, but talking more did not necessarily attract a higher score.

Katona (1998) also looked at the familiarity effect, but on the negotiation of meaning between interviewers and 12 adult candidates of English as a Second Language (ESL) in Hungary. The test had an interview, a picture task, and a role-play. The two situations that were compared were a practice with the known teacher and the live exam with two unknown examiners. While there is very little at stake during the practice, the anxiety felt during the live test would have been a contributing factor to changes in the exchanges between the candidate and live test interviewer. The findings "indicate noticeable differences in the ways meaning get negotiated in two testing conditions" (Katona 1998:262). This is problematic because the first condition is practising, not testing. It is argued that familiarity affects the frequency and type of negotiation in exchanges between interviewer and candidate but the sample is not taken under comparable conditions: two tests or two practices as well as having a very small number of participants. Under these conditions it is difficult to sustain claims about the familiarity effect.

This issue of familiarity was taken a step further by O'Sullivan (2002), who examined the effect of familiarity on scores in pairs of (female) Japanese learners of English. Thirty-two students took part performing a task in pairs consisting of a personal information exchange. In order to measure whether scores were impacted by pairing with a friend or a stranger, the measure chosen was the accuracy and complexity of the language. The findings suggested that the degree of familiarity of a test taker with his or her interlocutor, in the Japanese context, affected performance significantly. It was concluded that candidates should be allowed to choose their own partner for the orals because "when the participants paired with a friend the resulting measured performance is significantly superior to when their partner is a stranger" (O'Sullivan, 2002:286). The implication was that candidates speak more accurately with people known to them and consequently, familiarity affects performance in terms of accuracy and complexity.

O'Sullivan (2002) was the first to empirically demonstrate that familiarity changed scores. Other contributing reasons beyond familiarity resulting in greater accuracy could come to light from a microanalysis of the test discourse between the pairs. The study concluded that while the results may be predictable and significant in the Japanese context, the variable of familiarity may be culture specific. O'Sullivan made reference to Porter (1991), whose findings showed no evidence in support of the effect of familiarity on performance in an Arab context. With respect to affective features on large-scale tests, Porter (1991: 101) recommends that it would be dangerous not to heed possible effects and that in a practical sense test construction and administration could consider

"some selection of the features to include - including what is felt to be important and excluding what is felt to be irrelevant. Research into what the significant affective factors are, the scale of their effects and their field of operation (what topic areas, what cultural backgrounds) will be necessary to inform the selection process".

These recommendations were echoed by Iwashita (1999) and O'Sullivan (2002), who agreed that taking a closer look at the ways in which different variables interact in the paired task context will lead to a clearer understanding of the task and the constructs upon which it is based.

2.2.3 The personality effect

In addition to the familiarity effect, Foot (1999) also raised the question of matching a slightly reserved candidate with an over-assertive candidate. Randomly paired or grouped candidates can hypothetically be paired with different personality types. This is known as the personality effect on performance, and scores.

In a study on the effect of different personality types on task type, Berry (1996) investigated a group of 32 male and 22 female students. They were studying in the School of Economics in the University of Hong Kong. Their first language was Cantonese. Using the results of a psychometric test, Berry divided the students into two groups: introverts and extraverts. Students were subsequently tested in individual interviews and collaborative-paired tasks, in pairs that were homogeneous and heterogeneous. The principal finding was that introverts perform better on individual tasks and extraverts perform better in paired tasks, whether the pairing is homogenous or not. On the other hand, introverts perform slightly better on paired tasks, but

significantly better when paired with an extravert. Berry (1996) suggested that because of their personality, introverts on the same task have taken different tests and those different tasks, such as paired or individual, elicit different types of performances from different personalities.

This is an important issue in terms of the use the paired format, because of the possible effect on scores. The issue is whether the elicited performances from different personality types can be scored accurately. The findings show that extraverts speak as well as they write, and the introverts do not. However, personality type is an inherent characteristic. Therefore, whether extrovert or introvert, test results will need to be representative of other speaking performances if results are to be transferable beyond the testing situation. But what if the candidate had been paired with someone else in the test? Their results would have been different, and this is unfair. One way to further investigate this issue in Berry's would be to look at the discourse produced by homogeneous and heterogeneous pairs.

In addition to proficiency, familiarity and personality, all raised by Foot (1999) as issues of concern for speaking tests, some other interlocutor effects are outlined below.

2.2.4 Other interlocutor effects

Other interlocutor effects on discourse and/or scores have also been studied, such as the effect of first language (L1) on pronunciation, the effect of background culture on amount of talk, and the possible effect of gender.

Jenkins (1997) found that mutual intelligibility between candidates of the same L1 was increased when candidates used their accent to be better understood by their partner with the same L1. As a result, students would do better with partners belonging to the same L1 because they could make themselves understood to their pair, though not necessarily to a rater. From a validity point of view, the rater's inferences would not be transferable to a situation in which candidates speak to others not from the same L1 background.

Moving from pronunciation to quantity of talk in a test, this can be affected by the background culture. Young and Halleck (1998) compared American Korean, American Mexican and American Japanese pairs and reported on the issue of 'talkativeness'. The findings were that Japanese candidates spoke more slowly and changed topic less frequently than the Mexican candidates, who spoke more quickly and change topics more, contributing more to a test conversation.

There are as yet no studies on how gender affects scores on the paired format. Based on the interview, not the paired format, O'Loughlin (2002) studied interlocutor effects in the IELTS interview, including the impact of gender. That study found that the gender of the interlocutor does not have a significant impact on performance. Brown & McNamara (2004) summarize research on the impact of gender in language assessment which is at "the intersection of two sites of social power and control" (Brown and McNamara, 2004:524).

In sum, research on the interlocutor effects on score covers mismatches in proficiency, familiarity, personality, L1 culture and gender. In §2.3 we examine the effect of interlocutors not on test scores, but on test discourse.

2.3 INTERLOCUTOR EFFECTS ON DISCOURSE

This section is concerned with interlocutor effects. In interviews (§2.3.1) this means how the interviewer, as interlocutor, affects the candidate's performance. In a paired task (§2.3.2), the interlocutor effect refers to the candidates' affect on each other.

2.3.1 Interlocutor effects on discourse in interviews

Interviews are claimed to allow learners to take part in a communicative event and demonstrate different components of their communicative competence (Ross, 1996). The interviews can be unscripted or scripted. In the unscripted type there are the Australian Second Language Proficiency Rating (ASLPR) (Ingram 1986) or the International Language Roundtable (ILR) type, for example where guidelines are provided with topics to direct the questioning focus. The questions are reformulated for each candidate. In the scripted oral proficiency interviews, there is, for example,

the revised IELTS. In this test, raters have a predetermined set of questions and scripted prompts resulting in each candidate hearing very close to the same words from each interviewer.

An increasing body of research is concerned with the discourse of both the interviewer and the candidate. This stems from a line of inquiry into the use of interviews as measures of non-test communicative ability (van Lier, 1989).

While interviews are not the major focus of the present study, the research into interviews pre-dates paired format research and laid the foundations for what is continuing today. Four main areas of interviewer research can be identified:

1. Studies that identify how similar or different interviews are to non-test interaction (e.g. Lazaraton, 1992; Ross & Berwick, 1992; Young & Milanovic, 1992; Johnson & Tyler, 1998; Katona, 1998).

2. Studies that are concerned with the effects of accommodation on candidate speech (Young, 1995; Young & Halleck, 1998), or on the manner in which interviewer speech is adjusted (Ross & Berwick, 1992; Cafarella, 1994).

3. Studies that investigate the impact of the interviewer *type* on ratings (e.g. Ross & Berwick, 1992; Ross, 1996; Berwick & Ross, 1996; Lazaraton, 1996; Brown & Lumley, 1997; McNamara & Lumley, 1997; Reed & Halleck, 1997; O'Loughlin, 2001; Morton, Wigglesworth, & Williams, 1997).

4. Studies on the effect of individual interviewers on ratings (e.g. Reed & Halleck, 1997, Brown & Hill, 1998; Brown 2003).

"Communicative competence or effectiveness is an abstraction that is rarely defined with any precision in terms of test performance" (Brown, 2003:20) and this is a large part of the problems faced by individual raters and interviewers. How raters/interviewers or candidates affect the communicative competence of the other can be shown in terms of power and asymmetry. Status or a position of power, invested in both the interviewer and the Native Speaker is found to be one of the main factors influencing candidates' output in the oral interview.

Power can be a question of asymmetry, as in the position of candidate versus interviewer, or the inequality of the NS versus the NNS position. There is a significant body of research that critiques interaction in the oral interview in relation to these positions of power.

Asymmetry of power was addressed by van Lier (1989) in a paper on the nature of the "activity and work done by participants in the Oral Proficiency Interview". He was not "primarily interested in rating procedures and reliability" (van Lier 1989:492). He was more interested in identifying and describing "performance features that determine the quality of conversational interaction" (van Lier 1989:497). Van Lier pointed out features of the Oral Proficiency Interview that were to be considered in the light of their effect on interaction.

One of these features was symmetry of power in the two modes of social interaction in question: the conversation compared to the interview. Van Lier demonstrated the differences between the two, and showed that the interaction was controlled by the interviewer, who has a position of power, and is also a NS. To balance the asymmetry, van Lier put forward a possible solution: to transform interviews into conversations at some point in the interview process by requiring candidates to perform task based assessment in peer groups. Notably, this was taken up by UCLES.

Following van Lier (1989), other studies (Perret, 1990; Johnson, 2001; Skehan, 2001) have considered the effect that the structure of the interview has on interaction. Asymmetry of power between the interviewer and the candidate can affect the conversational skills displayed by the candidate. Studies have shown the paired format to be more interactive and less restrictive when compared with the Oral Proficiency Interview (OPI), because either candidate can direct the discourse as they co-construct the dialogue. Perret (1990) analysed six unstructured ASLPR interviews and found restrictions on the candidates' output because of the nature of the interview. Candidates did not elicit information, and the interaction was predominantly a one-way information exchange. In this way, the candidates were not

able to demonstrate a variety of speech functions (for elaboration on speech functions see Hatch, 1992).

Power relations were also examined by Young and Milanovic (1992). Their focus was on dominance, defined as "the tendency for one participant to control the discourse by various means" (Young and Milanovic 1992:406). Dominance was also examined by observing goal orientation, which reflects the internal goals of the speaker. This is manifested for example by maintaining topics over a number of turns. Another way that Young and Milanovic (1992) look at dominance is by observing interactional contingency, which means how participants react to each other. Their analysis of 30 interviews of the FCE showed less variety of interaction with interviewers. Interviewers control the "other participants access to the floor by means of interruptions and questions" (Young and Milanovic 1992:406) and have the undisputed right to initiate or terminate a topic.

The effect that 'status' has on the ability of the interview test format to elicit different functions has played an important role in the introduction of the paired format, which attempts to correct the power asymmetry that caused dissatisfaction with oral interviews (Swain 2001). With the introduction of paired tests (or their variant the group test), candidates are interlocutors, they co-construct the performance (Brown 2003) and are jointly responsible for more symmetrical output (Iwashita, 1998; Egyud & Glover, 2001; Taylor, 2001; Lazaraton, 2002).

Johnson and Tyler (1998) looked at turn taking and topic in the Oral Proficiency Interview, and found that the OPI was not an example of 'everyday life' conversation. The interaction pattern has been shown to be more varied in a paired oral (Saville & Hargreaves 1999; see also David 2000 who reported on turn taking and selection of the next speaker in transitions in a group oral). More opportunities to interact and to direct the discourse also meant increased opportunities to demonstrate conversation management skills.

The research on interviews raises all sorts of questions on issues of status and power and tries to demonstrate how limiting oral interviews are as a sole task for measuring

oral proficiency. This sets up a discussion of research into the way paired tasks affect discourse in §2.3.2. The question to consider is whether peer interaction is any better or whether it also raises the same difficulties.

2.3.2 Interlocutor effects on discourse in a paired task

In Young and He's (1998) seminal volume on interview discourse, interaction was brought to the forefront of oral language testing issues, but discourse studies on peer talk have not yet been represented in this way. Interaction has been shown to have a similar range of features within a conversation and an OPI interview (Lazaraton 1992). Although much research has examined the discourse of the OPI, less research has investigated the type of conversation in a test between peers in a paired/group task such as the one on the Cambridge First Certificate of English (FCE).

The discourse study conducted by Brown (2005) brought the 'interlocutor effect' to the fore. Using Conversation Analysis (CA), the study analysed rater variation across International English Language Testing System (IELTS) interviews. By applying the micro analytic tools offered by CA, the differences in the manner in which two different interviewers elicited the same candidate's speech sample were highlighted through the interviewer's use of questions, topic management and feedback.

The findings demonstrated that because different interviewers had different interactional styles, raters and candidates co-constructed the interaction differently. The issue is not that it was different, so much as that the difference was shown to negatively impact on test outcomes. As a result, two sets of raters in Brown's (2003) study (one set that interviewed and one set that rated from a tape) had varying impressions of the same candidate's speaking ability.

Despite the fact that paired orals have been optional since the 1980s, only a few studies have been carried out on the effect of paired tasks on the interaction that is produced in the discourse *between* candidates. In such a case the interlocutors, two candidates, affect each other. The studies reported on in the next section show

language sampling, increased functions and conversation management affected discourse.

2.3.3 Increased functions in paired or group tasks

Language functions that are expected in a paired format speaking test are: informational, for eliciting and passing on of information; interactional, for taking turns and managing the discourse when speaking to another interlocutor; and also conversation management, to enable the interlocutors to perform the first two functions with ease (Weir, 1993). Inclusion of a group task in an oral test elicits a broader range of speaking styles, which enables examiners to better discriminate between different levels (Shohamy, Reves and Bejarano, 1986).

Studies based on the Cambridge ESOL suite paired format, report that the paired orals were less restrictive compared to interviews, with a larger number of functions. They also report that paired orals were overall more interactive: when speaking to their peers, and not being interviewed, candidates could ask questions and manage the direction of the discourse in which they were participants. A closer look at transcriptions of the discourse produced in paired orals reveals that interviews allow only a restrictive output for candidates. Recall from §2.3.1 that in a study on 30 pairs for the Cambridge FCE, candidates and interviewers were shown to be in an asymmetrical relationship (Young and Milanovic 1992), which led to restrictive output.

Research on the Cambridge FCE has shown that the paired format is more interactive (ffrench, 2003). Based on Wier's (1993) speech function taxonomy, three categories were included: informational functions, interactional functions, and managing interactional functions. Ffrench (2003) made a comparison between paired tasks and the interview in which the paired tasks had a greater percentage of interactional functions in a pair. The paired conversational management tasks were 15% compared to 5% found in an interview which supported the claim made by Taylor (2000) reporting that the paired format was more interactive. Conversation management in paired orals is examined in the section that follows §2.3.3.

Studies have looked specifically at the number of functions used in different tasks in the paired format in the FCE. Lazaraton (2002), in her work on qualitatively validating oral tests, looked at the number of speech functions elicited in 20 transcripts of the FCE. The sample included a candidate-to-candidate paired task and an interlocutor-led discussion with the two candidates. Out of the total of 15 functions found, five were more common in the paired task, but nevertheless, the functions were not restricted by task but were distributed through the other sections of the test.

There have been recent developments in the revised Cambridge Proficiency of English (CPE) exam, which has included the paired format since 2002. These came about as a result of research into operational findings and statistical data gathered in a study validating revised assessment criteria (Weir & Milanovic 2003). Based on 11 raters and 24 candidates, Weir and Milanovic (2003) compared the use of functions in the interview and in the paired format. They found that informational functions accounted for between 72 % and 93% in the interview. In one extreme case, interactional management was not manifested by the candidate at all. In the paired format, however, the functions were spread over informational function (55%), interactional functions (30%) and interactional management (15%). As a result, the assessment criteria for the CPE have broadened from interactive communication to include discourse management along with grammar, vocabulary and pronunciation.

In an early study on paired discourse, with a small number of participants, Egyud & Glover (2001) analysed the language in peer-peer talk in a test for English, in Hungary. The paper was written in response to Foot's (1999) doubts about the validity of paired tasks, which had been incorporated as part of the Cambridge ESOL suite without any published validation research at the time. The authors illustrated the claim that candidates produce 'better English' (Egyud and Glover 2001:70) in a pair without operationalizing 'better English'. That is to say the authors did not define what 'better English' meant, in order to enable it to be measured. Egyud and Glover (2001) looked at the type of interaction produced and claimed that pairs help produce 'better English' compared to an interview. A paired task also was also claimed to

provide candidates with an opportunity to produce 'their best' (Egyud & Glover, 2001) language for assessment.

Excerpts of transcriptions of conversation made from videos of paired performances selected from 14 students (Egyud & Glover 2001) at a vocational secondary school were used as examples. The selection from the transcript the authors provided as an illustration for 'better English', showed regular even length of turn between the two candidates. The regular even turns were compared, on the same task, with a candidate and an interviewer. The transcription showed the dominance of the interviewer by the length of their turn. The non candidate interlocutor held the floor which did not allow much candidate participation. This was considered in that study, to be an example of students not producing 'their best' (not operationalised either). One could argue that it is equally an example of an interviewer not doing their 'best'. Only one example of each type of discourse was shown and the sample was small (14). Perhaps 'better English' and 'their best' meant a greater range of functions or improved conversation management but the small sample was not analysed to such a degree.

More examples of what happens in the interaction between pairs are needed for claims of improved language sample to be sufficiently supported. But, despite the small study and the lack of operationalized concepts, it was a classroom teacher's voice calling for research to support experience (Egyud and Glover 2001:76). Specifically:

"Innovation has always been a difficult task, but without it progress is impossible. Problems and difficulties should be addressed, and possible solutions must be found: this is the only way of making progress. We are convinced that the paired format offers students and teachers opportunities for development and an escape route from the dire one-to-one situations."

Other discourse studies, which follow below, *have* documented the range of functions and conversation management skills exhibited in samples from paired tasks.

2.3.4 Conversation management in peer tasks

The skills relevant to conversation management are turn taking, topic organization, sequence and overall structure of the discourse.

A recent study (Galazci2004) reported that pairs awarded high marks on the FCE paired task had higher scores on the criterion 'interactive communication' (a Cambridge test criterion). The success of the pairs on the paired task could not be demonstrated within the constraint of an interview because the candidates achieve these higher scores when they make topic-building moves. The topic-building moves were found through CA analysis of 40 tape-recorded speaking tests provided by Cambridge during 6 administrations of the FCE.

Galaczi's (2004) study is important because it shows which features of communicative interaction are correlated with successful performance or high marks. By adapting Storch's (2001) patterns of dyadic interaction, Galazci was able to identify three major patterns of interaction in the way that test-taker dyads produced "topicality": Collaborative, Parallel, and Asymmetric interaction. A fourth category, a combination of any two of these, was labelled a 'Blended' interactional pattern. The main findings were that higher scores on interactive communication were usually accompanied by topic-building moves, extending topics with follow up questions, less topic initiation, more speaker nominations through questions, and more supportive conversational features such as latching, overlap and interruptions. These results imply that there is a need for the elaboration of criteria used for marking interactive communication in the FCE. The paired format is more interactive precisely because of the choices available for turn taking, which are part of conversational management.

The focus of the review of the literature in the area of speaking assessment has so far covered testing speaking in pairs (§2.1.3), interlocutor effects on scores in speaking tests (§2.2) and interlocutor effects on discourse in the interview and on the paired task (§2.3). Much of the discussion was concerned with the effect on the discourse and scores of a test task, whether an interview or a peer paired task. The main finding is the increased function and conversation management skills associated with the paired format.

Between the task and the score lies a very significant aspect of the performance test: the rating scale. "[I]f the task types reflect the developer's view of the communication needs of the candidates concerned, the rating scale (or scoring rubric) reflects the developer's view of language ability" (North 2003:1). With this in mind, the literature review moves from the candidates' performance to the tools used to assess it. Section 2.4 deals with issues related to rating scale development and section 2.5 focuses on the assessment of communicative interaction.

2.4 DEVELOPING RATING SCALES

Rating scales usually mark out a series of levels, each of which has descriptors that include characteristics of the performance expected at that level. The sample of candidate discourse used to assign a score is understood to derive from underlying language abilities, or the construct being tested. Scales with number points along a continuum - with band descriptors specifying what learners can do in a language - have been used since the inception of modern day language testing.

In a comprehensive report for TOEFL, North (2003) surveyed scales for rating language performance and recommended scoring rubrics for the revised speaking paper. North described four issues of particular relevance to Chapter 6 of this thesis, each outlined in turn below.

The first issue concerns the basic problems with rating scales. North (2003:3) put forward three points: "[t]he questions of the match of the scale purpose to the descriptor, the extent to which the categories in the scale have a base in linguistic theory and the extent to which the scale has a base in measurement theory". Of these, the first is most relevant to this thesis, which is concerned with rating interaction between peers in the paired format. Part of the research process would be to match the descriptor to the scale purpose. On this point, North (2003: 10) criticised the ACTFL for "burying the assessment criteria in holistic descriptors". North concluded that the assessor should be born in mind and that it is "the qualitative aspects of communicative language competence which can be deduced from performance which

should be assessed". Both of these points relate to the aim of this thesis, which is to rate interaction based on samples of candidate performance.

The second issue raised by North (2003) involved categories used for description in rating scales. In reviewing current practice in 2003, North divided these into three: firstly, scales which are based on the four skills, without taking into account language use; secondly, scales which describe performance which are reported as salient and thirdly, those based on a model of communicative language ability. Of those three categories the second is most relevant to this thesis. As will be reported in Chapter 6, the EBB rating procedure is based on discussion of the performances and by using videos of the performances. In other words the resulting procedure is very much embedded in the performance. According to North (2003:2), the main advantage of this kind of performance based scale is that "the development of the categories tends to involve detailed investigation and discussion of the performances involved and … the categories selected as a result are embedded in the context concerned". The focus on the performance is a strength as long as inferences that translate to an ability in relation to performance in the real world can be made about the candidate's performance in the test.

The third of the four issues concerns the distinction between levels, which can involve abstract, concrete and quantitative divisions. The concrete distinction between levels, relevant to this thesis, is a major focus of the EBB methodology. The main argument for using concrete divisions between bands is that the scales are formulated in concrete terms, or in criterion statements to which 'yes' or 'no' can be answered. According to North, the concrete formulations can be *qualitative,* such as the EBB procedure, or *quantitative,* such as the discriminant analysis used by Fulcher (1996) to develop accuracy/fluency scales.

The final issue presented by North in the process of rating with scales is the presentation of the final product. Three types of scale types were presented: scale, grid and checklist. None of the types mentioned represents the method used in this thesis but it is important to understand how the EBB differs from the usual types of rating tools.

There remains another issue to consider which is highly relevant to the chapter on rating procedure development in this thesis. This last and most relevant issue is that there are two ways of developing scales that are used to guide the rating process in oral assessment: the intuitive method and the empirical method. The intuitive method was described by North & Schneider (1998: 220) as "appeal[ing] to intuition; the local pedagogic culture and those scales the author has access to". This has been criticised because scales used to rate second language performance devised in this intuitive way are not derived from *actual* performance output. The result is what Clark (1985: 348) described as "descriptions of expected outcomes, or impressionistic etchings of what proficiency might look like". The intuitive method was broken down by Fulcher (1993) into three subcategories:

- Expert judgement: An experienced teacher or tester writes a scale.
- Committee: A group of experts agree on the levels and the wording of descriptors.
- Experiential: A scale that is developed with either of the two categories above, and develops and evolves over time as it is used.

These three categories overlap. For example, expert 'intuition' is also an inherent part of developing empirically based rating scales, but the important difference from other scale development methodologies is that the empirically-based scale development utilizes samples of candidate performances in the test setting. The empirical methods, set out in Fulcher (2003), are:

- Data-based or data-driven scales: requiring the description of key features on a task
- Empirically derived, Binary choice, Boundary definition scales (EBB): speech or writing samples are divided by criterial yes/no questions that lead raters to the final score.
- Scaling descriptors: band descriptors are collected then ranked by difficulty and a scale is devised from the sequence of ranked descriptors.

The process of scale development has traditionally been based on expert intuition, because the development of rating scales has most often been left to expert judgment. Scales based on expert judgement have been criticised for lacking empirical underpinnings (Fulcher, 2003) in that they have not evolved from, and are not connected to, the language sample elicited in the test. In fact, it is the exception rather than the rule for information to be made publicly available on how individuals or committees come to agreement on descriptors for rating scales or for detailed information of the development of a scale to be published (Fulcher, 2003).

Aside from use of intuitively devised scale development without transparent processes, the area of scale development has been criticized on other counts. According to North and Schneider (1998), there appears to have been a lack of attention to theoretical issues, because scales have not always been based on models of measurement or models of language use. Criticisms also extend to practical and administrative issues, such as whether descriptors are relevant to users, or whether a scale is being used in the context for which it was intended.

Turner and Upshur (2002) report that rating scales have also been criticised for producing scores with low validity and reliability. Offering a critique of the practice of intuitive scale development, problems they cited involve:

- The ordering of scale criteria may be inconsistent with the findings of Second Language Acquisition (SLA)
- Criteria may be irrelevant to tasks and content
- Criteria may be incorrectly grouped at different levels
- Scales may lead to raters making false judgements, because of relative wording

These critiques are central to this thesis, because there is a great contrast in procedures between what is normally performed in rating scale development and what is suggested as an alternative in chapter 6 for example, which is the use of the EBB procedure.

The focus of this review now moves away from language rating scales generally to communicative interaction and speaking scales. In doing that it is necessary to reflect on the fact that although communicative interaction is included in scales, such as those for the Cambridge suite for example, there is no evidence that scales for interaction in pairs or groups have been *developed* by observing direct examples of live or taped performance. Are they intuitive or data based development procedures? It appears that the scale makers have used transcriptions of performance to retrospectively validate scales, as for example Lazaraton (2002). It would seem that the criteria on the scales have been *validated* post hoc (Weir 2003) with data drawn from tasks designed to elicit communicative interaction. It would be fair to claim that they are not data based procedures: up till now developers have not directly observed performances during the scale development process.

With respect to criticism on speaking scales, Fulcher (1987) argued that assessment scales were based on a theory with little justification. The theory that scales are based on, according to North (2003), has three areas that almost follow chronological developments in theoretical, then operational models of language:

a) Pre-communicative scales that are generic. Those scales are based on the 1960's elements: grammar, vocabulary and phonology.
b) Performance based scales that are contextualised. They describe features observable during a particular task in a context.
c) Communicative ability based scales that are generic. Scales are based explicitly on a model of communicative language ability.

Bearing in mind that there are different theories, Fulcher recommended that a construct in communicative oral tests could "be empirically tested… in discourse analysis" (Fulcher, 1987: 291). This means that a rating scale for 'communicative ability' in an oral test could have an underlying construct based on what is actually said by candidates in a communicative task on a test. The discourse would be subject to discourse analysis, thus validating the inferences made from that performance. A test developer would be in a position to claim that such a scale described candidate

performance and the discourse showed that it was a true reflection of performance. This position on empirically testing scales was supported by Matthews (1990: 120), who in a discussion on the difficulty of applying rating criteria on the ELTS (now the IELTS) added that "descriptions of sufficient accuracy to allow objective decisions to be made are simply not available" because the scales at the time were intuitive and not empirically validated.

Further criticism of 'intuitive' scales has most recently been expressed by Cummings, Kantor, Powers (2002, p. 68) who, specific to speaking scales, say that "most holistic and other analytic scales lack firm empirical substantiation in respect to evidence about L2". Holistic scales describe a number of levels of overall ability. These "holistic and analytic scales are based upon a general theory of the development of language abilities", to quote Turner & Upshur (1996:56). Analytic skills separately rate a number of different attributes, which is a useful tool for focusing raters on important areas to hold in balance while rating. Nevertheless, the resulting scales could be criticised for not being closely connected to what candidates produce in the test.

Upshur and Turner (1996) draw a contrast between the holistic and analytic scales, referred to above, and multi trait and primary trait scales, which are more closely linked to task and educational objectives, and more closely reflect what candidates produce in a test. Multi trait scales aim to empirically develop scales from work samples in a context but they are not tied to a particular task or context. Primary trait scales are task specific and are developed from work samples to suit a particular context. The choice of presentation format analytic, holistic, multi trait or primary trait is ultimately influenced by the context and purpose of the scale.

North (2003) lists four types of primary trait scale: task-focused, objective checklist, ability, and task-focused and binary scales. The latter is most relevant to the study reported in Chapter 6. For the purposes of this thesis, binary scales will be referred to as 'rating procedures', 'methods for arriving at judgements' and occasionally 'scales' when quoting directly from Upshur and Turner (1996) who also referred to them as 'scales'. The chart of binary decisions that is the visual representation of this technique

has a series of numbers for scoring, but it should not be associated with band scale descriptors that are commonly used. This point is discussed further in Chapter 6.

Returning to the most common way of arriving at rating scales, using the intuitive methods criticised above, the call for experts to develop scales using empirical data based methods has now been taken up by experts developing procedures such as the EBB method. This method works from samples of student oral or written work, connecting these samples to the score through yes/no questions. The method has so far been employed for the productive skills of speaking and writing.

There are two methodologies for data based scale development, as identified in a survey of the literature (North & Schneider1998):

- Data based rating scale design for interviews and group oral fluency using discriminant analysis (Fulcher 1993, 1996)
- Data based level scale design for story retelling and information gap tasks for speaking using boundaries in EBB scales with the identification of decision points made by raters (Upshur and Turner 1995, 1999, 2002)

In both of these approaches, key features are identified from a language sample. Fulcher (1987) called for data based criteria for tests of oral performance, after he investigated raters using the ELTS scale and found four native speaker participants would not meet the criteria. More specifically, Fulcher analysed a very short transcribed dialogue from native speakers recorded by a hidden tape recorder. He suggested that the ELTS scale was "based on the functional notional categories, [and is] attempting to describe not what *actually* happens in communicative situations, but what communicative theorists *think* happens in communicative situations" (Fulcher, 1987: 290). More than ten years later he developed and validated a rating scale design for interviews and group oral fluency using discriminant analysis (Fulcher 1993, 1996).

The second example of evidence-based rating scale design uses the identification of decision points made by raters at level boundaries (Upshur & Turner 1995, 1999)

during the development process. Raters divide levels by looking at a representative range of language samples from a single task and a single population such as a story retell or a paired information gap task. This scaling procedure results in no levels being included that did not have performances to match. According to Turner and Upshur (1996: 60), the procedure has "a hierarchical sequence of attribute checks and the scales require raters to make a series of binary choices about features of student performance that define boundaries between score levels". Because the procedure was ***E****mpirical*, involved ***B****inary* choices and defined ***B****oundaries* it was known as an EBB rating procedure.

We have seen with the two data based methods in this section that in empirically based rating scale development there are several procedures that can be adopted. One of them is the identification of key features by experts (Upshur & Turner 1995). The other involves close scrutiny of the data (Fulcher 1993).

The fact that little had been written about data based scales provoked a call from researchers (Young, 1995; Upshur & Turner 1995, Fulcher, 1996) for methods to "derives scales empirically" (Chalhoub-Deville, 1995: 28) as opposed to the more usual intuitively based rating scales mentioned in the section above.

2.4.1 Evidence-based scale development

The Upshur and Turner (1995, 1999) EBB methodology combines several methodological steps in the development process. This style of evidence-based scale combines two steps in the process, which results in two methodologies each validating the other. The scale development process uses key features to separate levels by closely scrutinising the data, and rater consensus is used to achieve a hierarchy of rating levels. North (2003) suggests that EBB offers the possibility of scoring rubrics for modest achievements. The language level of the learners in this thesis is Beginner Spanish which would be considered to be at the modest achievement end of the proficiency spectrum. I put forward that the ideal for the context for this study fully described in Chapter 3, would be to use the EBB methodology.

2.4.2 Data based scale methodology: issues with EBB rating scales

The use of expert input on data based empirical scale development could possibly limit the use of the resulting scale to one specific context, meaning the scale would be criterion-referenced and task-based. This is not the case according to Upshur and Turner (1995). The EBB uses performances to guide the scale development and then uses consensus to complete the scaling process. As such, the scale can be used to describe aspects of proficiency by demonstrating how candidate output data corresponds to a particular level agreed on by experts, and not just a particular task in a particular context.

Turner and Upshur (1996) looked at the difference between levels on a scale, focusing on the key feature that separates two levels of performance. They applied it to written compositions, and to narrative monologues (Upshur and Turner 1995). It is important to note that in these two studies the authors marked out the difference between levels, instead of attempting to describe all possible features of a level. The key feature of difference found between levels was subsequently used as an empirical basis for raters to find cut off points between levels. These cut-off points were based on the overall quality of performances, in the two groups (according to the answer to the yes/no criterial question).

The EBB scales are devised from work samples in a particular context. The concrete rather than abstract formulation means that people who use the scale (whether they are raters, professionals, or outsiders) are more likely to interpret the scale in a similar way (North 2003). The EBB scales are developed through 'hands on' practical methods that make the scale development process transparent.

In the EBB scale development method there are several issues to consider: is there a broad enough range of data on which to base the scale? Is there a range in number and experience amongst the raters? There is also the issue of the number of raters providing expert information, and the quantity and quality of candidate data input. In addition, there is the question of how raters arrive at a consensus while marking out

the different levels (North 2003). If there is no agreement, the scale development process cannot continue to the next stage.

One might ask whether test method affects rating scale construction. Test method is usually seen as raising issues of validity, with data driven scales causing tension between the construct and the inferences being made. Upshur and Turner (1999) analysed discourse produced by candidates to address test method effects on discourse. It was concluded that rating scale construction required analysis of the discourse produced on the task. By doing this the rating scale would then be task specific. If the scale is too task specific then the only inference that can be drawn is that candidates can perform the task, with no reference to a construct beyond it.

The scales are based on real performances but the issue remains that they need to be related to other performances. Rating peer interaction requires two peers to speak together. When candidates are rated for ‘communicative interaction’, their speaking mark will allow one to infer how well they can ‘speak to others’, not just to each other. Therefore it is not task specific.

In EBB, raters do not need to interpret and apply band scales. The issue of interpretation by raters in this context is that of understanding the meaning of the criterial questions and their relevance to the new performance, not the one the scale was based on. The scoring method that is used once the procedure is developed is fixed for a particular trait. Raters using EBB do not need to ‘judge’. They answer 'yes' or 'no' to a series of questions in a preset order. Their answer reflects their interpretation of a performance; whether it is the performance the scale was made on, the performance being judged or the future performance about which inferences can be made.

Having dealt with issues related to data driven scales we now move to a discussion on validating score inferences.

2.4.3 Validating scales through verbal protocol analysis

Verbal Protocol Analysis (VPA) has been used to investigate the rating process raters in ESL second language writing (Cumming, Kantor, & Powers, 2002; Lumley, 2002). Although VPA is widely used, the increasing number of studies of rater cognition of first and second language in *writing* contexts is not matched by studies in the *oral* context - because it is not possible to interview and rate while simultaneously recording a verbal report. As Wigglesworth (2005: 103) states, Verbal Protocols are "not appropriate for use with listening or speaking data because they necessarily conflict with the communicative nature of such activities". The situation is different for research into cognitive processes for assessing writing in both first and second language contexts where it is possible to research concurrently while rating written tasks (e.g. Huot, 1993; Milanovic, Saville, & Shen, 1996).

VPA has nevertheless been used for rating on ESL and English for Academic Purposes (EAP) speaking tasks (e.g. (Brown, Iwashita, & McNamara, 2005). Research into rating oral proficiency is limited to rater orientations (Pollitt & Murray, 1996; Brown, Iwashita, & McNamara, 2005; May, 2006), and scale validation (Meiron, 1998; Brown, 2000). Both rater orientation and scale validation are relevant to this study, which includes reports on rater behaviour such as rater focus and orientation.

Meiron (1998) reports on the rating of a monologic task from the Educational Testing Service (ETS) test: Speaking Proficiency English Assessment Kit (SPEAK). The SPEAK test is used for assessing the English proficiency of non-native speakers, particularly graduate teaching assistants in the American college and university system. The task Meiron reports on involves a narrative where learners use picture prompts to retell a story.

The Meiron (1998) study provided insight into two aspects of rater behaviour: the approach raters adopt when carrying out ratings, and their rating focus. With respect to the first aspect, two approaches emerge: (1) an analytical style where discrete features in the criteria are weighed up; and (2) a more holistic approach where raters do not home in on any particular feature when deciding the final score. With respect

to the rating focus, Meiron (1998) found that it went beyond the specific criteria, to include the use of other self-generated features not on the scoring rubric. This is particularly relevant as a precedent to the studies carried out in this thesis in Chapter 4, where raters will be asked to orient to rating 'interaction' without any criteria at all.

Brown (2000) finds that IELTS rater focus - when rating an interview task - goes beyond the specific criteria in the scales used and includes additional criteria of pronunciation, fluency and communicative skill. Communicative skill as defined by the raters in that study included a range of learner behaviour such as use of communicative strategies, comprehension of the interviewer, ability or willingness to speak at length, and ability or willingness to take the initiative and the organization of structure of discourse. In addition, raters' view of functional fulfilment of the task was found to vary greatly, and they also drew inferences on whether test taker could cope with the real (academic) world based on the candidates' content and interactional styles. A further study of rater orientation (Brown, Iwashita and McNamara, 2005) shows that domain experts are able to describe and distinguish qualitatively different performances by using their own generated set of criteria - which are similar to the ETS draft scales. This is a relevant point because in the study reported in this thesis the raters will similarly need to generate their own set of criteria through the EBB process.

In an examination of rater orientation to a paired candidate discussion task study, May (2006) reported that trained and experienced raters attended to many non-criterion features of the pair discussion performances. Each of the different categories were fleshed out from Verbal Protocols verbalising how raters interpreted in-house intuitively derived rating scales and scored pairs of candidates. Two raters observed 6 pairs taking the discussion task. Raters commented on the four different criteria: 'fluency', 'accuracy', 'range effectiveness', and 'overall'. The pairs took the test twice - once with a similar level partner and once with a partner of a different level. May (2006:47) observed that of particular interest was the manner in which:

"a rater acknowledges co-construction of the paired performance and the impact this has on the final rating. The rating scale requires raters to view the paired speaking test

as if it were the product of two solo quite distinct performances, which ignores the inherent co-construction of the performances"

The implications of May's study are that either the rating scales need to be modified or those raters needed improved training. This issue is further discussed in Chapter 6, in order to shed some light on May's (2006) findings.

Meiron (1998), Brown et al (2005) and May (2006) provide a strong theoretical support for the use of Verbal Protocols in order to create and validate scales using raters, the focus of the present research. None of these studies found differences between rater focus according to level, while Pollitt and Murray (1996) did, in their study on raters assessing performance on the Cambridge Certificate of Proficiency in English (CPE). Raters in the latter study were found to concentrate on the lower level candidate of a mixed level pair. Depending on the proficiency level, different performance characteristics were more or less salient, which prompted Pollitt and Murray to remark that scale development should start with rater perceptions of proficiency. In line with this finding, rater perceptions of 'peer interaction' are investigated first in Study 1 Chapter 4, before the development of the rating procedure is described in Chapter 6 of this thesis.

An insight into precisely what candidates notice during these test taking processes (as opposed to what raters notice about them) can be drawn post hoc from a type of verbal report called Retrospective Verbal Protocol (RVP) (Ericsson & Simon, 1993).

For example, in an introspective investigation of listening test strategies, Buck (1991) uses questions to prompt students to recall their test taking procedure, with immediate retrospective recall following their hearing an audio text. Gruba (1999) also used retrospective verbal reports, to investigate listening strategies while students watched videos.

Verbal reports as used for validating tests have been thoroughly described by Green, (1998). It is assumed in this thesis that Verbal Protocols for speaking in language

testing may also be appropriate for evaluating speaking skills in a pair or for validating the paired task as part of the test process.

2.5 THE ASSESSMENT OF COMMUNICATIVE INTERACTION FOR PAIRS AND GROUPS

The rating of interaction changed once there was recognition of two-way interaction in the oral proficiency testing literature (Brown & Lumley, 1997; Brown & Hill, 1998). Teaching practice had moved away from teachers interrogating students in front of the class to include more pair and group work in class, naturally leading to testing in pairs or groups. While the disenchantment with interviews as the only task progressively led to the introduction of pair or group tasks within tests, the inclusion of pairs was more importantly a reflection of contemporary best practice in teaching.

Speaking in pairs implies interaction, but to express this interaction in rating scales is complex, and not always satisfactory. It is difficult to access many rating scales, but North (2003) surveyed a range of those publicly available. Of particular interest to our discussion is the term 'interactional fluency', meaning "[t]he ability to judge when and how to take the turn, (turn taking), the capacity to work with other people, inviting their views, commenting on the contributions of others (cooperating). This focuses on the interactional nature of discourse as participants weave contributions into joint products" (North 2003:91).

From the studies carried out so far, a number have investigated the difficulty scale makers face in devising scales that adapt to the paired and group context. Nunn (2000) tackled the problem of designing rating scales for small group interaction during classroom activities, as distinct from paired tests. The study reported on the role of the rater and the rating scales for pairs and groups. He proposed a rating scale that incorporated language proficiency and communicative performance for scoring within a group. What separates this study from others on pairs or group assessment is that the focus was on aligning teaching goals (including interaction) with assessment. By implication, rating scales included group interaction. Nunn (2000) acknowledged that for groups and rating scales "the considerable difficulties of reliability and validation need to be fully understood and the facile extrapolations about how

students can perform in real life should be avoided" (Nunn, 2000:178). Despite the recognition of a difficult problem, Nunn suggested that teachers recognise that "the question is not whether to do it but how to do it as fairly and efficiently as possible" (Nunn, 2000, p. 178). Nunn proposed that the scales can guide teaching, define principles of assessment and provide teachers and students with achievable goals. . How one empirically develops Nunn's scales still remains unresolved, regardless of the scope of their intended application.

While the scales suggested by Nunn are analytic, elements identified by Van Moere (2006) are more holistic and intangible. Van Moere (2006:436) focused on "the more intangible interpersonal factors in the way group members react to each other". The 'intangible' is an issue that remains so far unexplained by raters or candidates in the peer testing context. The interpersonal factors arising out of communicative interaction need to be described by those involved, both the candidates and the raters. If this is achieved then we may be a step closer to capturing issues related to the intangible in a scale. This may reduce variability caused by the person-to-person factor. Although the person-to-person communicative skill is desirable in proficiency testing, the construct is not well described and as a consequence is not reflected in the scales.

Orr (2002) analysed verbal reports given by raters on the decision making process during the rating of the UCLES First Certificate of English (FCE). Thirty-two raters completed verbal reports (Green, 1998) on two separate pairs of candidates performing the paired task from the FCE under test conditions. In that study Orr reports most compromising results, in which raters were not consistent. Orr's findings about raters confirmed those of Brown (2000) and Meiron (1998). These two latter studies are not for rating pairs, and are detailed in §2.4 on scale development. Orr's raters were found to: (1) apply different standards because they varied in severity; (2) focus on rating criteria in different ways; and (3) vary in the amount of non-criterion information they noticed for each candidate. These findings are true of all forms of rating of oral interaction (see McNamara 1996). However, a certain amount of non-verbal communication, for example eye contact and body language, was included in the non-criterion information while rating the paired interaction. The raters' had

varying perceptions of the performance, but this was not reflected in the scores. This makes it difficult to understand what FCE speaking test scores represent, and ultimately undermines the validity of the paired oral, because the scores mean different things to the raters that arrived at them.

Perhaps the most important finding in support of scales for group orals emerged with Bonk and Ockey (2003). In their thorough many facet Rasch analysis of a second language group discussion task, Bonk and Ockey found that, despite differences in severity, "rater and scale reliability were achievable under real testing conditions even when the discourse went largely uncontrolled". Their statistical examination showed, however, enormous differences in the severity of raters, which would have had a major impact on scores and except that the study controlled for severity.

In their research into rating scales, Bonk and Ockey (2003) found varying rates of rater severity to be of concern. Bonk and Ockey (2003) pointed to the difficulty in interpreting FCE speaking scores, because raters vary greatly in their perception of performance. They researched the paired task included as a task within oral proficiency interviews, where the interaction between candidates is symmetrical (as opposed to the asymmetry in an oral interview). As a result, the variable 'interlocutor' can either be an interviewer/assessor or another candidate.

2.6 CONCLUDING DISCUSSION

From a background to testing speaking through paired discourse, the discussion of the literature moved to interlocutor effects on scores and on discourse and then moved on to the inclusion of communicative in speaking scales. Finally there was discussion on the development of rating scales, focusing on data based scales in particular. The practical focus of this thesis relies on theories drawn from overlapping areas of language testing research presented in the literature review.

The paired format has been found to "generate language performances that allow us to test much more complex constructs than in a traditional OPI" (Fulcher 2003:189) but we still need to find what that complex construct represents. The construct may involve non-verbal communication, as raised by Orr (2002), who showed eye contact to be a feature of interaction that raters focus on, but not one included in rating scales.

The literature also raised the question of interactive listening, e.g. Pollitt and Murray, (1998) who considered whether comprehension forms part of testing speaking. This was echoed by Galazci's (2004) speculative questions concerning the ability to change position from listener to speaker as levels of interaction proficiency increase. The correlation of topic management with higher marks in interactive communication in Galazci's (2004) study also raised questions about rater focus on turn taking and topic management.

This thesis does however take into account the Meiron (1998) study which reports that raters self generate criteria not in the scoring rubrics. This thesis also bears in mind the Brown (2000) study which reports that the focus of raters beyond the scales, particularly on 'communicative skill', is broad ranging. Taking both of these into account and adding to the balance the study by Pollitt and Murray's (1998) and their consideration that scale development should start from raters' perception of saliency in what they are rating, all three studies are shown to be, in combination, the foundation upon which this thesis will extend into a new area for empirical scale development: interaction.

The literature on oral interviews, group and pair orals presented in this chapter focuses on the manner in which different variables affect test discourse. It raises questions about: the unpredictability of interaction, the construct being tested, and how to rate paired interaction. Issues such as interactional patterns between peers, and differences in the discourse of peer interaction that have already been identified in the literature demonstrate the interest in paired orals and the importance of extending the current body of knowledge.

The difficulty in rating pairs has started to be examined and, in light of varying rater perceptions and severity (seen in section 2.4 above), two very important questions touched upon in chapter 1 need to be remembered. Firstly, whether the process of 'communicative interaction' (as it is called by UCLES) is a construct that can be adequately operationalized for raters to perceive the intent of the construct through the rating criteria they are trained to use and secondly, whether the communicative

interaction construct is scalable in the traditional sense with band levels and accompanying descriptors.

With the narrow range of scales that include interaction available for inspection for paired orals, it is difficult to speculate what needs changing, if anything or what should be included. From the Orr (2002) and the Bonk and Ockey (2003) experience of varying rater severity, it would be helpful to build scales using the features of interaction that raters themselves identify and focus on. Conversely, it may be better to build scales based on features empirically found in the interaction, as in the recommendations made by Lazaraton (1998) and Galazci (2004) for FCE criteria. The question remains whether either of these proposals, if taken up empirically, would have any effect on the relative severity issue which is a permanent feature of rating behaviour. The studies in this thesis ask raters to observe live test interaction, in order to build scales from features they identify. In other words, it combines both suggestions.

Brown (2004:15) argues that the conceptualization of what speaking tests measure has shifted, which "means that the interviewer is likely to be even more closely implicated in the construction of the candidate as communicatively effective" (compared to when assessment was focused on linguistic form rather than on communicative interaction). This new type of performance calls for examination of co-construction in pair or group oral interaction. In the words of Luoma (2004:190):

> "[w]hat we need to understand better is how one person's performance affects the others. There needs to be additional investigation into what it is about an examinee's talk and his or her accommodation to the conversation partner that should be appreciated in order to make evaluations in a fair way".

The issue is that discourse between the pairs is unpredictable, so improving the fairness of rating for paired tasks requires a more thorough investigation of the examinee's talk. A more thorough investigation of this type would involve including the examinees themselves as 'experts', and incorporating their experience of interaction in a paired performance. One of the three studies in this thesis, Chapter 5

on candidate orientation, asks candidates to comment on their experience of the interaction while performing the test with a peer.

Empirical features of the co-constructed candidate output that are relevant to the construct have been brought to light by research (Lazaraton 1998, Galazci2004). So too have the features that are attended to by the raters (Orr 2002, Bonk and Ockey 2003). In addition there have also been candidate studies (Brown 1993, Luoma 2004). The combination of all these perspective would mean a possible solution to the problems identified by the few studies on paired candidate output and on paired rating scales. To start to understand co-construction as the 'cause' of the problem, interaction needs to be described from as many perspectives as possible: that of the raters, of the candidates as well as studying the discourse. Only then will be there be enough known about the development of an empirical rating scale for score inferences about candidates performing 'communicative interaction' to be validated.

Assessment of speaking in pairs based on a scale with detailed descriptors aims to assist raters to be explicit when awarding scores. Speaking is a complex skill that is difficult to narrow down to discrete points such as lists of competencies, so the scale descriptors that raters use to award scores should reflect the levels of complexity involved in speaking. Speaking in pairs adds a further dimension to the interaction, and hence another level of complexity to challenge scale development procedures.

The findings from research into different types of speaking test formats such as interviews, pairs or groups combined with the findings from studies on scale development processes, such as intuitive and empirical, together inform the present study.

2.7 CHAPTER SUMMARY

The literature review first surveyed how testing in pairs emerged as a format. Secondly, we reviewed the interlocutor effects on scores in speaking. Thirdly, we examined research findings on interlocutor effects on scores. Finally, the findings on the effect that pairing has on discourse were presented before concluding with the assessment of communicative interaction

Chapter 3: ASSESSING THE PAIRED TEST: MOTIVATION AND METHODOLOGY

3.1 INTRODUCTION AND CHAPTER OVERVIEW

This chapter covers several areas that provide the site context and background to the motivation behind and the methodology employed in the main studies that make up this thesis. Section 3.2 provides a rationale for the introduction of a paired test. Section 3.3 deals with the choice and trial of a paired task. Section 3.4 outlines the key outcomes of the task trial, and the implications for the main study. The research agenda (§3.5) lists the motivations for each of the studies that make up the thesis. An overview of the methodology for these main studies is held in §3.6.

3.2 RATIONALE FOR THE INTRODUCTION OF THE PAIRED TEST IN THE SPANISH PROGRAM

We saw in Chapter 2 that paired task discourse in a paired oral task in a beginner foreign language achievement test has not been researched, and neither has there been development of empirical data based rating scales for such a context.

The Paired Test (PT) used for this study will be referred to as such in order to distinguish it from any other paired test. Before the introduction of oral assessment in the Spanish program in the site for this study, assessment varied from level to level and class to class. There were marks awarded for class participation as in the beginner language courses where students were given a mark from 1 to 10 for each class. (The fact that students were being given a mark for participation by the teacher without explicit criteria was of course problematic as a way of assessing spoken language skills). This mark was given for each class then averaged out for a final end of semester speaking participation mark in the class. There were other courses of an intermediate level where students were individually interviewed in their teacher's office at the end of semester. In other courses where students were more advanced, there were oral presentations in front of the class, also at the end of semester, which were frequently read.

If students were not spending time being interviewed in class then assessing interview response skills was not a natural progression from class practice. It is the task and its relationship to learning that matters more. These three methods being used to 'mark speaking' raised questions about fairness. Threats to fairness included the interlocutor effect on scores, discourse patterns in interviews, and lack of transparency - where no criteria were made explicit for participation (see discussion in Chapter 2, §2.2 and §2.3).

In addition to the problem with the task being used, there were issues with the assessment criteria - or rather the lack thereof. Teachers were not given a guide for the class participation mark, although both the interviews and the class presentation had a list of criteria that students' performances were marked against. The criteria were intuitively developed, as was usually the case in modern language departments. No rater training sessions were offered for new tutors or moderations across the levels to ensure fairness.

In order to address the issues of fairness and transparency, a new type of assessment was called for, and the paired test task was developed and trialled in 2001. The reasons for the introduction of the paired format for the speaking test were:

- To be representative of class activity
- To cut time and cost
- To remove the rater from being implicated as an interlocutor in the task
- To make rating more transparent to students
- To standardise marking across large first year groups of beginners

We address each of these issues in turn.

To be representative of class activity: The introduction of a speaking test in pairs in this setting was a natural progression from Communicative Language Teaching (CLT) in that the test reflected the task and the type of interaction students were accustomed to in the L2 classroom (Taylor 2001, Egyud and Glover 2001). Students were being taught in the target language within the framework of CLT. Tasks that required pair and group work made up a high proportion of the available practical

class time. So the overriding aim for the PT was to reflect the amount of class time spent working in pairs.

To cut time and cost: Assessing students in pairs would simplify test administration, by cutting time and cost. If students paired up instead of entering the tutor's office individually there would be a major saving (Hildson 1990, Swain 2001).

To remove the rater from being implicated in the task: There was, and is, a changing population of casual teachers involved with the course. Removing the interviewer from the interaction, by pairing students, was intended to avoid an interviewer effect (e.g. Brown 2003). Removing the interviewer to a rater/observer position would counteract the different training and experience the tutors brought to eliciting talk (though not to assessing orals). It would also lessen the impact on the reliability of ratings when many different tutors handled an interview and rated in idiosyncratic ways (Shohamy 1994, Brown 2003 et al, 2004), as the pairs of candidates could be taped for double marking and rater training. Without the participation of the interviewers on tape, they would be off the record and uncompromised.

To make rating more transparent to students: Pairing the students in a new task was intended to be a springboard for the development of new criteria. A new task and new criteria were aimed at improving the oral testing for the absolute beginner program by making the entire assessment procedure more transparent from task to assessment to feedback. Transparency would be gained by developing an evidence-based rating scale, by training raters and by providing students with scale specific feedback on their performance.

To standardise marking across large first year groups of beginners: With a new task and new criteria there would need to be rater training and moderation. This would be a good opportunity to involve the tutors in some professional development.

These are the five practical reasons why the PT was introduced for beginner Spanish students. The uniqueness of the setting for the study is described below.

3.3 CONTEXT FOR THE TASK TRIAL BEFORE MAIN STUDY

The setting for the trial of the task and for the test is described in detail here because so little is known about paired orals, at least beyond the context of high stakes English exams such as the ones run by UCLES (Galazci 2005). The detailed description that follows is an important part of the thesis, as it completes a picture of the context in which the test task trial took place and from which the main studies reported in this thesis emerged.

The site was chosen because it has a long established Spanish program with a large enrolment of beginners. The researcher had easy access to the site and the site participants would have enough in common with those in other similar university Ab initio or Beginner language programs for the research to be seen as a potentially useful case study.

The Spanish program has an average of 250 students studying at the beginner level. The students are taught by lecturers (the Australian equivalent of Assistant Professors in other educational systems), teaching assistants from Spain and casual staff. The test, as it is currently run following the introduction of the PT, is taken by students at the end of each semester of teaching.

The task was trialled and adapted several times in class as a paired activity in the lead up to the first full test task trial. In class it involved speaking in pairs continuously for ten minutes on three topics introduced during the course. Students were asked to write three topics from the course on a paper. Students all paired up and swapped the sets of three topics with another pair so neither of them had written the topics. All pairs started and spoke for ten minutes, introducing their topics and answering questions on the other partner's topics.

Then the paired orals were conducted for the first semester task trial and the end of first semester studies. The orals formed part of the end of semester examinations and they were audio-taped for second marking of grades. Audio recordings of the students' orals are required for double marking where the grades are too high or constituted a borderline pass/fail at the university. It was from these audiotapes that, with the students' permission, data was initially collected from one class consisting of twelve pairs in the first semester. This was used to gauge the success of the task trial.

In order to understand the place of the task in the test and how different it is to other test tasks, momentarily we move out of the test development narrative to more general considerations. The types of tasks generally used in speaking tests add to the variety of language output. The most common of these tasks are role-plays, picture-description tasks, information gap exercises or discussions.

- Role-play is a technique where the learner assumes the role given on a task card and acts out a situation, imagining how to converse appropriately within the role assigned.
- A picture task is based around a visual stimulus that the candidate describes, followed by further questions on attitudes or opinions regarding the picture.
- An information gap requires each student to complete a task by asking questions that will uncover information that the other candidate has, but they do not, and vice versa.
- A discussion, which involves more than two candidates, can be based on readings or on a mini-lecture.

Unlike any of the four task types listed above, the test task in the PT was put together as an oral achievement test based on material taught on the course. Candidates performed a familiar task, which involved introducing up to three topics in a ten-minute interaction and cooperatively negotiating this conversation under test conditions with their partner. Even though students practised talking with a familiar and self-selected partner, the test remained as spontaneous as possible, for a test, by leaving the exact topics of the test or their order unknown until the speaking test day. This is important because the candidates were *not* to rote learn a dialogue

The paired oral task set out in the test procedure below in §3.3.1 was developed to match student achievement at a particular level of the course. The differences between the FCE paired task and the PT discussed below are that the students perform a single task, there is no interview component, it is an achievement test and the candidates choose their own partners.

The paired discourse from the task used both for the trial in §3.3 and the main studies reported in Chapters 4 and 5 could be compared to the paired format in the high stakes test run by Cambridge ESOL (see §2.1.6). However, the task is different in most ways to the one on the Cambridge Oral Proficiency Interview - principally because it does not take place in the context of an interview since candidates only perform one task. Because so much research has already been carried out on the Cambridge paired task, the differences between it and the PT must be made clear before proceeding.

One major difference between the two tasks - one that forms part of the Cambridge interview, and the other stands on its own as an achievement test after a semester of work in a university course - lies in the fact that the talk arising from the Spanish task does not include the rater at all. In the PT, the rater is a participant only as an active listener - and in such a capacity is not permitted to guide the group or pair at any time during the task.

The task is also different to the role-play task used in the Cambridge exam, because in the latter, candidates take on assigned roles. However, in the PT candidates are being themselves. They are displaying to the rater what they can talk about together within the time constraints of the test and on the topics for the task.

The section that reviewed the literature on familiarity §2.2.2 in Chapter Two showed that students in pairs were more comfortable and performed better when they knew their partner. As far as matching candidates is concerned, although students take Cambridge tests when they reach a certain threshold at each new proficiency level, it

is still a 'proficiency' test, not an 'achievement' test. The fact that students are within a range at the same proficiency level because they are all beginners allows the PT to offer paired tasks. This 'range within a level' as it applies to the achievement test in this study, is different to the Cambridge tests only in that students select their partner in advance of the test date.

These students are within a similar proficiency range because they have been studying as a beginner for the same amount of time. Students have all taken the same number of contact hours and are deemed beginners, allowing pairing between them to be viable just as in the Cambridge tests.

We have seen above the reason why a Paired Test is used for beginner Spanish students in a university setting for practical reasons. The PT has been compared to a task on a large volume, high stakes testing system in the global testing context and the uniqueness of the setting for the study has been described.

3.3.1 Test procedure

Consistent test procedure is important in order for the candidates to have an equal chance of doing well. The procedure for the task trial (and in the main study reported in Chapters 4 and 5) was outlined on a sheet with instructions for the raters. It stated that in the testing room the students were to be seated facing each other with a tape recorder between them. The rater was to have a sideways view of them but was not to interact. The positioning is important to prevent the rater being selected as a speaker and brought into the conversation.

Each student was asked if they understood the task and what was expected of them before they commenced. If the reply was no, then it was explained that they were to speak with the other student, including the topics in whichever order they chose. After the tape recorder was switched on, the examiners were to introduce themselves and the students identified themselves. Then they were handed the task cards. The cards were written in Spanish and the topics were taken directly from the topics taught

during the course. After they were given the task card, as shown on the sample, below, the candidate started on the task.

Figure 3: Task cards

Tiene 10 minutos en total La familia Los días festivos Los amigos

Tiene 10 minutos en total Los fines de semana Los pasatiempos Las vacaciones de verano

As they talked for ten minutes the rater listened and rated each of them. Two minutes before the end they were flashed a warning sign by the rater. The timer sounded, the students finished and they left the room for the rater to give each of them separate marks for communication, comprehension, grammar and vocabulary. The rating criteria, described below, are in Appendix 1, in Spanish).

3.3.2 The rating criteria for the trial

The only existing rating criteria in the program at the time of the trial of the paired format were used by lecturers to assess the native speaker oral interviews and class presentations in the advanced stream. These existing criteria were not expected to cover beginner level, so new criteria were intuitively put together (Appendix 1) and trialled by the teachers involved in marking the orals. As a result, as well as trialling a new task we were also trialling new criteria.

At beginner level there were five band levels of proficiency under four criteria: two for linguistic competence (grammar and vocabulary); and two for interactional competence (comprehension and communication). The change in the rating criteria was intended to operationalize, for the raters, the change in procedure from interview

(which was used for advanced level students) to conversational interaction. The vocabulary and grammar remained unchanged.

3.4 Outcomes of the trial of the paired task

In the end of semester program meetings, the raters discussed the trial PT. They reported that it was easier to rate the pairs that 'sounded' like a conversation as compared with those pairs who did not sound like a conversation but rather interviewed each other in a stilted fashion. Whether, as a result, better marks were being awarded to those students whose interaction was closest to 'a conversation' - in the raters' opinion - needed to be investigated.

The thesis was borne out of the curiosity to investigate this impression about the differing qualities of the discourse, and the need to improve the criteria and to provide validity evidence in support of the score inferences from the test. By attempting to describe in detail, through context and discourse, the different processes in the paired interaction; processes such as rating the test and taking the test, an attempt would be made to uncover what made paired interaction deemed to be 'successful' or otherwise, for raters and candidates alike.

The three key outcomes of trialling the task of the PT were firstly that the task was successful in eliciting 10 minutes of assessable talk; secondly, that it was adopted by the Spanish program at beginner level; and finally, the intuitively devised consensual scale needed revising.

3.4.1 Implications of the trial of the test task on the main study

The issue of concern that emerged from the trial was that there appeared to be evidence of communicative interaction skill in the discourse. This was gauged from the rough transcriptions of the test discourse, and from anecdotal evidence of differences in rating expressed in teachers' meetings. It was also evident from listening to the tapes from the task trial: the inferences that could be drawn from the candidate performance had changed because it was not an interview but a peer-to-peer

paired task. These inferences were linked to the findings pertaining to topic changes and maintaining coherence, through relevant question forming within each topic, before engaging in a smooth transition to the following topic. Candidates were responsible for initiating new topics and responding to introduced topics, so the roles played were different to those in an interview.

These skills had not been detailed in the old rating criteria for the non-paired tasks used in the program previously. They were also not adequately detailed in the new criteria for the task trial. Developed intuitively, they incorporated *questioning* under the communication criteria and *responsiveness* under the comprehension criteria. Despite these intuitive inclusions to rate the new 10 minute paired speaking task, the result was that the raters felt there were insufficient criteria relating to the responsibility of two parties in the peer-to-peer interaction.

The rating scales used in the task trial focused on four criteria. With a view to incorporate questioning and responsiveness by both candidates, the rating criteria were arrived at intuitively and collectively. Two criteria were anchored in a pre-communicative four skills model (Lado 1961) for rating using grammar and vocabulary, and two criteria were anchored in a communicative ability-based model. Both of the ‘comprehension’ and 'communication’ communicative skill criteria (see Appendix 1) needed to be revised after the trial. The linguistically based accuracy and range descriptors were not of concern where raters' focus remained on the traditional criteria used for oral proficiency interviews, with which they were already experienced. The main concern was how to better rate the interaction, intuitively and collectively operationalized as ‘communication’ and ‘comprehension.’

The evidence found in the tapes of the trials called for further research into the communicative skill audible, but not visible in the task trial tape. This is detailed in Chapters 4, 5 and 6. If successful communicative skill could be perceived by raters, in Study 1 Part A, then further research would need to be carried out to develop rating procedures, in Study 2. The first step was to demonstrate that raters perceived differences in communicative skill across the different pairs.

Another important finding was that oral text cohesion was found to be part of the pair's communicative skill and was commented on by the raters of the task trial data. Text cohesion was missing from the intuitively developed scale used in the task trial. To rectify the situation a new scale, reported on in the main study, was developed and validated.

3.4.2 Task success

The task designed for the pilot was found to be suitable to test both the linguistic and interactional features at beginner level. It would be adopted, because it met the criteria outlined in §3.3.2 above. Specifically, it was found to be representative of class activity, and, being administered in pairs, it cut time and cost. The rater was no longer implicated in the task because the role of the rater was to listen and rate without speaking to the candidates.

3.4.3 Site adoption of paired oral interaction

The Paired Test continued to be used throughout the beginner program from the trial onwards during the development of the new scales. While the new criteria were being developed, those outlined in §3.3.2 were used on a temporary basis.

3.4.4 The consensual scale

Two of the reasons for trialling the PT were to make rating more transparent to students and to standardise marking across large first year groups of beginners.

A problem with the scales devised for the trial was that they were difficult to use. According to the raters, and anecdotally documented in meetings, the scales failed to sufficiently distinguish between levels of peer interaction for the raters to apply them confidently. To rectify the situation a new scale was developed and validated (cf. Chapter 6).

The conclusion from the task trial was that further research was needed. There were implications for the rating criteria from the new task. The changes in the test

discourse produced by the candidates in the PT, made it different from the interview or the class presentation discourse for which raters had used and devised criteria previously. If pairs were to be used then it would be necessary for new criteria to be devised and written into the level descriptors which accommodated the extended communicative skills identified by the raters discussing the task trial tapes. The intuitive rating criteria devised for the trial which included communication and comprehension would require further examination.

The trial task demonstrated, however minimally, that new rating criteria were needed because different discernible features were identified in peer-to-peer paired discourse data which raters in the post trial discussion commented on as being more 'conversation like' and less like an interview, as noted above. The issue that emerged was how the new criteria were to be identified and described, developed into scales and validated in order to best to rate the peer task. This is addressed in the next section.

3.5 RESEARCH MOTIVATION

3.5.1 Motivation for Study 1 Part A

Study 1 sets out to explore the peer construct and the meaning of the word 'interaction' to teacher/raters. Because in the trial the researcher only taped one class of pairs on audiotape, the non-verbal communication could not be observed by the researcher. A way of clarifying what was not evident in the taped data was to first video the test in the main study, then to ask raters and candidates how they oriented to the paired interaction.

Part A of Study 1, in which raters were asked to comment on interaction, was motivated by the initial findings of the task trial. These discourse samples appeared to include examples of turn-taking, back-channel, overlap and topic management that were noticeable to the raters. (A glossed transcription of 10 minutes of beginner level talk from the trial is included in appendix 2). The samples from the task trial were deemed to be 'conversation like' by the raters.

The fact that raters focus on criteria beyond the ones set down for them to use as rating criteria has been shown in the Meiron (1998) study and in Brown (2003). In their capacity as domain experts, raters have also been shown to devise a relevant set of criteria from candidate oral discourse provided to them for that purpose (Brown, McNamara and Iwashita 2005). Using these studies as evidence that raters produce the relevant and necessary criteria to mark when asked, in Study 1 for this thesis the raters were asked to identify the features of interaction that they were aware of in the same peer task as was used in the trial. The aim was to identify the features first then develop a scale to incorporate then in Study 2.

3.5.2 Motivation for Study 1 Part B

Before developing the scale, another complementary perspective needed to be gathered. We saw in §2.3.3 that Galazci (2004), who studied the Cambridge paired format, had called for studies to ask candidates to help interpret the interaction with their peers.

Part B of Study 1 was motivated by two reasons. The first reason was to see if the interaction that was visible to the raters was also an aspect of interaction that the candidates were aware of. The second reason was that a goal of the study was to develop and validate crucial aspects of the newly adopted task.

3.5.3 Motivation for Study 2

Building on Study 1, if the teacher/raters oriented to certain features of interaction without being asked specifically to identify them, it would suggest that the teacher/raters themselves could be used to develop an empirical data based rating scale using student discourse (Turner and Upshur 1996). Devising such a rating procedure would demonstrate a close relationship between the task, the linguistic output and the rating scales. In the words of Fulcher (1993: 99), scales produced through empirical data based procedures "are typically assessor oriented, require holistic or multiple trait scoring and have a construct focus". In the task trial, assessors had issues with two of the traits on the test: the rating of communication and

the comprehension (a copy of the criteria, in Spanish, is in Appendix 1). Assuming that successful communication relies on *both* candidates during peer task performance, then the rating of communication and of comprehension would combine as part of rating interaction. This motivated the new scale, which would answer Fulcher's call for "correspondence between the speech sample generated and the descriptors in the rating scale" (Fulcher 1993: 95).

The motivations for the two parts of the study have been put forward above. §3.6 provides an overview of the methodology employed for Studies 1 and 2.

3.6 OVERVIEW OF METHODOLOGY

This section outlines the methodologies used for Study 1 Part A (Think aloud Verbal Protocols), Study 1 Part A (Stimulated Retrospective Verbal Recalls) and Study 2 (the Evidence-based rating scale method).

3.6.1 Methodology Study 1 Part A

Study 1 Part 1 attempts to address **research question 1**: What features of peer interaction do raters attend to in paired task test performance? As we saw in Chapter 2, Verbal Protocols used for validating tests have been thoroughly described by Gass and Mackey (2000). Verbal Protocol methodology is discourse based and qualitative. The protocols can be segmented, coded and then analysed statistically if required. In VPA the data is analysed either by content and sequence or just content. In this study the focus is on the content alone.

Verbal protocol analysis is used here as a methodology because what the raters verbalise in the study is specific to what they are attending to in the video of the candidate's performances. The individual verbalisations are a way of accessing information about perceptions of interaction in a paired task. In the word of Gass and Mackey (2000) "protocol analysis may help elucidate the very "constructs" that tests seek to measure". It is argued here that Verbal Protocols for speaking in language testing may also be an appropriate methodology for evaluating the paired form of assessing speaking skills or for validating the paired task as part of the test process.

There are various categories of verbal report procedures, which vary according to the manner and the circumstances under which they are collected. There are three basic variations on the protocols: (1) talk aloud or think aloud, (2) concurrent or retrospective and (3) mediated or non- mediated (which together can result in eight combinations). To help decide the category of verbal report that will be collected, these three variations on the protocols can be broken into three questions. These questions are listed below with the answer relevant to the research undertaken in the thesis.

(1) Talk aloud or think aloud?

Verbal reports that are think-alouds occur in Study 1 when the raters say what they are thinking about the performance - making it a 'think aloud'. They express thoughts about non-verbal behaviour, about how candidates listen or gesture or take turns in speaking.

(2) Concurrent or retrospective?

The reports are concurrent in real time, because raters watch videos and almost simultaneously verbalise what they observe in the pairs' interaction. Raters could not be rating a 'live' oral and be speaking aloud for the research purposes to gather a verbal protocol. They had to *simulate* **concurrent** rating at the time the data was gathered.

(3) Mediated or non- mediated?

The reports were **non-mediated** because the raters worked at home and recorded the VP on their own without the presence of the researcher.

In sum, in Study 1 Part 1 the data gathered is a think aloud, concurrent and non-mediated verbal protocol.

3.6.2 Methodology Study 1 Part B

Study 1 Part A addresses **research question 2**: How do candidates view interaction in a paired oral? Following the choice of appropriate type of Verbal Protocol, in §3.6.1,

Study 1 Part A uses a retrospective non-mediated think aloud protocol (following Green 1998). It is impossible to gather data during a speaking test performance, so the candidate is asked to verbalize retrospectively after the test has been completed. The stimulus is the video of their performance. The protocol is unmediated: the candidates have the remote control for the video in hand; they speak into a continuously playing tape recorder when they find something to say.

3.6.3 Methodology Study 2

Study 2 involved devising a data based scale in order to answer **research question 3:** Can candidate peer performance samples from a paired test form the basis for developing a rating procedure for interaction? As we saw in §2.4.2, in the Upshur and Turner (1995) study, a video of ESL learners performing a 1 minute story telling task was used for raters to develop Empirically-based, Binary-choice, Boundary-definition scales. The method is described here step-by-step:

- In Step 1, the performances to be rated are selected.
- In Step 2, the experts divide the performances into two groups, impressionistically separating and defining one group as being better than the other.
- In Step 3, a question that divides the two groups is formulated by the scale developer/experts. The question could be, for example, "do they speak without hesitation?" The answer 'yes' would mean the sample belongs to group one and the answer 'no' would mean the sample belongs to group 2. The 'binary choice' is made based on a question, which is the boundary definition between levels.
- In step 4, the experts divide group 1 following step 2, then group 2 following step 2. The divisions and the number of levels are not known in advance. The number is only determined when then there are no further divisions to be made to the level. These steps are repeated, until the required number of levels is reached.

Each group of samples, which represents a point on a scale, is distinguished from the others by a series of binary choices based on the features of performance that describe that level and, critically, have been focused on by the raters. The difference between

what is stated in the 'criterial questions' and other more common types of rating criteria, is that this system focuses on the differences between each level rather than the similarities within each level (Upshur and Turner 1995:10).

3.7 CHAPTER SUMMARY

This chapter has provided an entry to the main study by providing background on the rationale for the introduction of the Paired Test. It has described the task and the trial. The key outcomes of the task trial provided the motivation for the two studies that comprise the thesis, to which we now turn.

Chapter 4: RATER ORIENTATION TO PEER INTERACTION IN SPEAKING TESTS (STUDY 1: PART A)

4.1 CHAPTER OVERVIEW & INTRODUCTION

4.1.1 Overview

This chapter contains Part A of Study 1, which is concerned with rater orientation to peer interaction in the speaking test. Chapter 5 contains Part B, which reports on individual candidate orientation to their personal performance with a partner in a paired speaking test.

This chapter is organized as follows. The introduction is followed by a description of the selection of the participant candidates and raters (in §4.2). Next, §4.3 describes the collection of the candidate video clips during the actual live test and collection of the rater orientation Verbal Protocols. Then the section describes the conventions used for transcribing, segmenting and encoding the data. §4.4 outlines the data analysis and §4.5 summarises the chapter.

4.1.2 Introduction

This study is the first of three studies (Study 1 Part A, Study 1 Part B, and Study 2) that constitute this thesis. The studies deal with three separate but connected issues that emerged from the task trial. This first study deals with rater orientation towards the interaction between two candidates performing a paired task in an oral test.

Rater orientation, to the interaction between peers, is examined from the point of view of Spanish speaking second language specialists: tutors and lecturers. As these specialists are trained to mark the speaking test used in this context, they are considered trained raters not 'naïve raters'. The way in which success in paired interaction is conceptualised by these specialists is the basis for building the evidence-based scale described in Chapter 6. The conceptualisation of 'successful interaction' is explored by having the language specialists watch video clips of students taking the test in pairs. The specialists watch the interaction between the pairs of candidates

speaking to each other in the test and are asked to simultaneously comment on what they *notice* about successful interaction - they are not required to 'rate'. This method of data collection is used here with the intention of revealing the meaning that language specialists attribute to the concept 'successful paired interaction'.

A set of 17 pairs of candidates was videoed while they were performing a peer oral paired task under live test conditions. These videos were used for this study and subsequently used for Study 1 Part B and Study 2. In this chapter the 17 paired test performances are used to elicit the unguided perspectives of raters on paired interaction.

The larger project of empirical scale development is divided into two studies over three chapters that include the collection, description and analysis of two connected sets of data and their use as the basis for building a scale. Study 1 involves collecting and using paired oral data to make explicit raters' and candidates' orientation towards interaction. This is written up in two separate chapters. Study 2 involves incorporating the language specialists' perceptions of interaction in the development process for an evidence-based rating scale.

4.2 THE STUDY: PARTICIPANT SELECTION PROCEDURE

The study involved two kinds of participants: candidates (students) who took part in the orals (§4.2.1), and their teachers who rated their performance (§4.2.2). Details will now be provided of the recruitment of each group.

4.2.1 Candidates: beginner Spanish language students

The videoed tests used in the three studies that comprise this thesis were collected from a large pool of 128 individual candidates across first year level in a beginner language program.

In terms of the candidates, the 64 pairs of students who took the undergraduate Spanish beginners oral in October 2002 were invited to return shortly afterwards to watch their videoed test in order to receive feedback on the fluency and accuracy of their performances. Of these, 25 candidates accepted the invitation. The 17 pairs used in the three studies that make up the thesis all involve at least one of the 25 candidates

who returned: there were 8 pairs and 9 candidates without their partners. The missing nine partners had given consent to take part in the study but their participation in this part did not eventuate. The 25 candidates participated in the second part of Study 1 by completing a questionnaire and giving a verbal report on their own performance of the oral.

The candidates were first year undergraduate beginner Non Native Speakers (NNS). The 17 pairs of candidates were all acquainted, and self-selected as pairs from their cohort. Personalised individual feedback on their performance was received in exchange for volunteering for the project (details are given in Chapter 5).

As the live videoed test data was going to be used as input for scale development, it was important that it represent the normal range of output of students at this level. In order to check this, the scores gained by the pairs were considered.

Students had been given marks with four-criteria, five level rating scales being used at the time of the oral. Candidates were marked on grammar, vocabulary, comprehension and communication, with each criterion having 5 possible levels of achievement. On an aggregate score out of twenty, which combined all four traits for speaking in pairs, the candidates fell into five groups, as illustrated in Table 4.

Table 4: Candidates and scores

Number of candidates n=34	Oral score out of 20.
2	**9**
3	**10/11**
10	**12/13**
6	**14/15**
10	**16/17**
3	**18**

At the lower end, two candidates achieved 9 out of 20; three achieved 10 or 11 out of 20; ten achieved 12 or 13 out of 20, six achieved 14 or 15 out of 20; ten achieved 16

or 17 and at the higher end, three achieved 18 out of 20. Almost as many students whose mark was below 13/20 as 14/20 or above took part. For a convenience sample it was reassuring to have an apparently representative group take part in the study.

4.2.2 Raters: Spanish language specialists

The other participants in the study were twelve Spanish language specialists who were approached and agreed to take part in the first part of the data collection. All had teaching experience on university level Spanish Language courses and had taught on and assessed the beginners' program. They had a range of qualifications from undergraduate degrees to doctorates and a few had postgraduate teacher training. Three were, or had been, language-teaching assistants funded by the Spanish government. They all were all very experienced in teaching oral communication skills in the Communicative Language Teaching style. They had all taken part in rater training and paired oral examinations using the rating criteria employed by the Spanish Program.

The raters included eleven native speakers and one non-native speaker (from Portugal). Together, they represented varieties from Spain (8) and Latin America (4), and were male (3) and female (9).

4.3 DATA COLLECTION

Two separate data collections took place. Firstly the candidates' performance was videoed during their actual semester test in pairs. (See Appendix 4 and 5 for ethics clearance.) Secondly the raters provided Verbal Protocol on the performances.

The video collection took place during the end of semester speaking tests. In the following week, the candidates returned to observe their own performance with their partner, as described in §3.3.1.

4.3.1 Verbal report data collection

As described in §3.6.1, this chapter involves a think aloud, concurrent and non-mediated verbal protocol.

Each of the 12 raters received a CD-ROM (convenient for moving easily from one pair to the next on a computer), a set of instructions and an audiotape for recording their comments independently. They were asked to individually record a verbal report after watching the video clips of three pairs of candidates assigned to them. Each pair of candidates was commented on by at least two different raters. The pairs were distributed among the raters to prevent the same pair repeatedly appearing first or last for a rater to comment on, and thus to avoid an order effect. The instructions provided to the raters were:

"Watch the 3 different pairs and comment in English into a tape recorder about the interaction. While observing this pair of candidates' performance comment on what makes interaction successful or not. Please comment on all aspects of the performance including, but not restricted to, what or how something is said.¨

The comments were requested in English despite that fact that they were all, bar one, native speakers of Spanish, and the researcher is also a Spanish speaker. It was of concern that it would be difficult to find a suitable qualitative researcher to carry out the intercoder reliability in Spanish. Intercoder reliability involves two raters coding a transcript and then comparing to identify whether they have assigned different or identical codes to segments of the transcript which have been coded with a predetermined coding scheme.

For each performance, the verbal reporting consisted of two steps. First, the raters watched the entire video clip performance. They then recorded a summary of their impression of the interaction. They were to give any of the reasons they thought could account for the success or otherwise of the interaction between the candidates. This first step was intended as a practice at verbal report production. Raters made general comments about the pair and sometimes about particular candidate behaviour.

In the second step, they were asked to watch the performance for a second time. This time raters stopped the tape at significant randomly selected intervals. Raters recorded what they had noticed about the pairs' interaction.

These first steps were performed individually by each participating rater, using the verbal report method. The two hours of the data collection were carried out in private, with no discussion between raters as to what 'interaction' meant. This was an important factor in obtaining unguided orientations. These two steps were repeated three times till the allotted three ten minute clips had been first summarized and then commented on in detail.

The tapes were collected and were listened to before being transcribed. One tape was blank and another only had the first summary step recorded because the rater could not complete the task due to illness but wanted to offer a contribution by the deadline. This rater then re-listened and completed the detailed verbal report. The rater with the blank tape was happy to repeat the task but said that it would be more concise as she had already said it once. One other rater commented on four different pairs to those assigned. It nevertheless contributed to the data set, though it meant that some pairs did not have descriptions from as many raters and others had more than were required.

4.3.2 Transcribing verbal reports

The verbal report data from the raters were transcribed orthographically without capturing any non-verbal features such as pauses, intonation or emphasis. The tape was replayed and the transcription was checked against the original. Each report was segmented into ideas units (Appendix 7 has a sample of one rater speaking on three candidate performances).

After reading through the transcripts the raters' comments were found to roughly divide into comments on pairs or comments on individuals. The trend appeared to be that comments on a pairs increased the better the pair worked together. Conversely, the more the candidates worked as individuals, the more the comments were made about individuals.

It is clear from the transcription of the raters' comments that some pairs inspired more commentary than others. It was also found that some raters were not as eloquent as others; they appeared not as capable of expressing and verbalizing what was occurring in the interaction before them.

4.3.2.1 Segmenting protocols

Raters varied in the style of protocol that they produced. Three of the twelve gave short dot point like oral descriptions of what they considered to be important regarding interaction. The other nine spoke more discursively.

The first step in analysing all the rater orientation discourse was to scan for comments on the same theme or same idea. This concept is an 'ideas unit', defined by Green (1998) as what is said in relation to a single aspect of whatever event is being focused on. However, Brown, Iwashita and McNamara (2005) redefined ideas units as "a single or several utterances, either continuous or separated by other talk falling within the same turn, with a single aspect of the performance as a focus." This latter definition was chosen over Green's, because it was more practical to apply to the protocols in this study.

Having decided where to draw the boundaries between the ideas, the next step was for each new aspect of 'interaction' commented on by the language specialist to be counted as 'new' for each rater. That is, when the raters repeated themselves, it was not recorded as a new entry against that rater or that candidate. Within one segment, sometimes a lengthy one, a point was only counted once until the full turn finished. For example, a comment made on 'questions' in the paired interaction may have had a 'rater reflection' on an aspect not relevant to the ideas unit 'question'. In these cases the ellipsis […] was used to indicate that there was some other speech that had been left out from that category because it was not relevant to that ideas unit. To illustrate this, an extract from rater 9 appears below.

> *"Um in terms of the physical interaction between the two of them well 1 think it is very telling about them she has her legs crossed and one of her arms or*

> *hands is between her legs as if she is trying not to move that hand that much maybe I put a lot of emphasis on the movement of the hands being Spanish but 1 think it is very telling.[…] In the first pair number 7 he had a very successful interaction because he was using his hands and he was really into it but um these two pair number 8 are really not very interested in communicating that way um"* (Rater 4 on pair 9)

This rater's protocol was easy to split into ideas units because each new idea started with 'um…'. However, in the middle of the idea about non-verbal communication, there is an aside where the rater makes a comment comparing pair 8 to pair 7 (in italics). This would have the three-dot ellipsis to take out the irrelevant bit, now in italics for the example. The rest of the ideas unit, without the ellipsis, would count as one in the coding, while the ellipsis could be moved and counted as another ideas unit.

This redefining of the categories was most useful where interruptions of a long turn in a rater's reflection excluded the beginning of a new ideas unit, because the rater returned to clarify or expand the original idea. This occasionally resulted in coding differing between coder and checker. However, if one ideas unit section was taken out and replaced by ellipsis, then it could be moved, and the point of disagreement between coder and checker was resolved.

4.3.2.2 Developing an encoding scheme: thematic analysis

The categories were originally divided into multiple subcategories at the ´discovery stage´. The discovery stage is a part of qualitative analysis that involves allowing categories to emerge form the data being analysed before grouping them in related areas during the process. It is a little like a brainstorm in reverse: you have all the ideas but you need to tie them together into connected patterns. This is represented by the last column on the right labelled 'discovery stage labels for the coder', in Table 5, below.

Defining categories involves repeated data reduction, rearranging and recoding as the researcher cycles through the data, grouping into small categories then regrouping into larger ones as categories become clearer. Green (1998) recommended finding a balance between too many and too few categories to achieve rater reliability. The categories were grouped by theme after more coding, seen in the middle column,

labelled secondary group refined categories in Table 5. This grouping procedure was repeated until five final categories emerged, shown in the left-hand column. These final 5 categories were defined before coding and checking.

Final coding categories	**Secondary grouping Refined categories**	**Primary discovery stage (details for second coder)**
Non-verbal communication	VISUAL SIGNALS	*Any eye movement* *Where they look* *What & who they look at* *How long they look at each other*
	BODY LANGUAGE	• *Hand gesture* • *Movement or position of feet, legs, body posture*
Interactive Listening	WORDS SIGNALLING COMPREHENSION	*Cooperation* *Making relevant comments* *Offering help* *Filling a silence/provide missing word* *Clarifying before continuing* *Giving an example to help comprehension* *Demonstrating comprehension*
	SOUNDS SIGNALLING ENGAGEMENT	• *Back-channel* • *Encourage the other speaker to go on* • *Show a keenness to be involved and* • *engaged with other speaker* • *Sound interested* • *No demonstrable comprehension*
Topic management	TOPIC COHESION	*Develop conversation* *Grow topic extend* *Connect a topic*
	TOPIC QUESTIONS	• *Reference to formulating questions to ask* • *The word question appears*
Turn taking	TURN SPEED	*How fast /slowly they respond* *Flow / natural /automatic*
	TURN LENGTH	• *How long / short they speak for*
	TURN DOMINATION	*Direct reference to how one person dominates and interferes with the others' capacity to demonstrate fluency*
Comments not related to interaction	REFLECTION	• *Compare two candidate pairs* • *Anything about exam , task, or examiner* • *How pairs connect, familiarity* • *Talk not on interaction* • *Observations /general comments* • *She looks…/ he appears…/she seems*

Table 5: Rater content analysis expanded coding grid

To illustrate how the reduction of categories worked, I will use the first row in the table, working back from the right to the left. The data was first coded under the idea ‘any eye movement’, ‘where they look’ ‘what and who they look at’ and ‘how long they look at each other’. Each of these small clusters was then joined under one umbrella category called visual signals. Meanwhile on a related but different thread, any comments made on hand gestures, body movement or posture were coded to hand or body until they were subsumed under ‘body language’. By the third cycle through the data, visual signals and body language became subsumed under the same feature commented on by the raters: ‘non-verbal communication'.

This process of cycling and recycling through the data resulted in the establishment of five categories: non-verbal communication, interactive listening, topic management, turn taking, and comments unrelated to interaction. Each of these will now be discussed in turn.

The first category in the left-hand column, non-verbal communication, amounted to observable body language between the candidate pair. These observations could still be made if the sound was turned off and the clip was watched in silence. As shown in Table 5, they fell naturally into two further categories: one for eye contact and another for general body language.

The second category in Table 5 refers to the support provided on the part of the ‘active’ listener. This is another new conceptual category for analytic rating scales. ‘Active’ here contrasts with ‘passive’ as in the role of the rater who listens but is not expected to contribute. Unlike the support signalled by body language, this second category is audible and can be of different types. For example, the interlocutor can provide feedback such as back-channelling while the other speaker maintains the floor. Alternatively, the interlocutor can provide support by filling a silence or providing the word the other partner is searching for but cannot produce fast enough. Both of these types of support enable the interaction to continue. The ‘word versus sounds’ distinction adds a layer to separate comprehension from non-comprehension while allowing a candidate to still remain an interactive listener.

The third category in Table 5, topic management, emerged from the cohesion of questions asked by the speaker. This involved making sense within a turn and within a single topic. Any rater comments that were connected in any way with questions fell under this category. This topic management category also was broadened to include a second type of cohesion extending over the complete oral text. This second type refers to cases where candidates introduced new topics successfully, when the timing was appropriate, so that the other speaker took up the topics. Rater comments that mentioned developing conversation, extending the topic or connecting the topic all fell into this category.

The fourth of the broad categories, in Table 5, was managing the interactional fluency of the conversation. This included comments made by the raters on turn speed, on turn length and on domination within the paired discourse. Comments on turn speed included topics by the raters on the flow, the automaticity and the naturalness of the turn speed. This was broken down into comments that related to how fast or how slowly the candidates in the pair responded to their partner. Finally, turn domination subsumed comments that made direct reference to how one candidate dominated and interfered with the other's capacity to demonstrate fluency.

The first four categories accounted for four fifths of the data. The remainder were observations such as the initial summaries, which raters had made in their protocols. The researcher found that these were not always directly relevant to the 'interaction' between the candidates - the focus of this study. They also included appraisals of the performance, how the pairs connected or the candidates' level of familiarity. Apart from the appraisals, all comments made about the exam, the examiner or the task, were also grouped in this section. Finally, comments made on linguistic resources such as vocabulary or grammar rather than successful communication vocabulary were also placed here.

From this initial analysis of five categories, the researcher developed a protocol from the raters' comments describing the features that each category included, with summaries of the types of comments the raters made to clarify what was meant by

each feature. This information was placed on a grid to be used as the basis for the final coding of the rater's orientation to paired test interaction. The distribution of the topics and percentages is displayed in Table 6.

Table 6: Percentage of coding: rater orientation

Rater Coding	**Percentage of coding**
Interaction irrelevant observations	16.3
Non verbal interpersonal communication	17.4
Interactive listener	17.6
Topic management	28.9
Turn taking	19.8

These five initial categories drawn from the qualitative analysis of the Verbal Protocols were distributed into comments on non-verbal communication (17.4%), the support of the listener (17.6%), topic management within and across topics - including questions (28.9%), turn taking (19.8%) and rater reflection not on interaction (16.3%).

While nearly a third of the comments by the raters were on topic management, this was not further divided into use of questions and topic management because topic and question were thought to be mutually dependent as a category. Nonetheless, we note that 57% of the topic management comments referred to topic, and 43% referred to questions.

4.3.2.3 Calculating encoder reliability

Independently, the researcher and the coder coded a third of the data set, one protocol out of three for each of the 12 raters. Using the Hatch and Lazaraton (1991) formula, the total number of agreements was divided by the total number of coding and the intercoder agreement was estimated to be 84.5% as in Table 9 below. Gass and

Mackey (2001) indicated that the level of agreement is often in the vicinity of 80 per cent, as does Storch (2001) for discourse studies, meaning these results fall within an acceptable range.

Table7: Percentage of coding: rater orientation

Rater number	1:	2:	3:	4:	5:	6:	7	8:	9:	10:	11:	12	Total
Candidate number	2	2	13	11	1	6	1	8	14	5	5	7	
Coding agreement	38	11	20	17	23	10	22	31	20	20	34	15	261
Total codings	42	13	25	22	32	12	25	34	25	21	42	16	309
Intercoder agreement													84.46

4.3.2.4 Categories reduced as a result of intercoder reliability

Going over the coding with the code checker highlighted some overlap between 'taking a turn' and 'managing a topic'. This was found to be the cause of some intercoder disagreement. In consultation with the coder, the two categories were deemed unclear and were joined. This became a new large category: 'interactional management', including turn taking and topic management as subcategories. Therefore, questions were coded under turn taking instead of topic management. This gave the 84% level of agreement.

Note that the difficulty of separating topic management and turn taking was also acknowledged by Galazci (2004). She noted that "the multi-functionality of questions as both topic management and conversation management devices caused some discrepancy in the coding" (Galazci 2004:97).

4.4 ANALYSING DATA

In the final analysis four main categories emerged from the content analysis:

1. Non-verbal interpersonal communication: A speaker in a pair demonstrates visual and corporal non-verbal communication

2. Interactive listening: A supportive listener in a pair offers verbal signs of comprehension or provides audible support to the speaker
3. Interactional management: A speaker in a pair cedes to the listening interlocutor at a comfortable speed after a reasonable turn length and demonstrates the ability to change or develop connecting topics
4. Interaction irrelevant observations: A rater comment about pairs other than about interaction

Each of these categories that emerged from the raters' comments is defined and expanded below.

4.4.1 Non-verbal interpersonal communication

This first category refers to non-verbal interpersonal communication between the candidate pair. The raters would be capable of commenting on this category even if the sound was turned off and the clip was watched in silence. A flow of eye movement and of gesture and body positioning physically supports what takes place verbally in the interaction.

As an example of the gaze subcategory, a rater referred to whether or not the candidates looked at each other during the performance:

"*what works really well with them is that they look at each other and never lose the thread of the conversation* (rater 1 pair 1)*"*.

Candidates that did not look at each other attracted comments about the lack of 'gaze':

"they look at the paper not at each other (rater 6 pair 13)*"*.

Body language, the other subcategory for non-verbal communication, also involved both negative and positive comments. The use of hands was positively appraised by rater 2:

"I find the use of the hands…. I find the girl with the glasses uses her hands when she talks. It gives a nice colour and is more in tune with the Latin American speech and culture; and it's her way of expressing her feelings. It helps her interaction to be more positive and fluent (rater 2 pair 3)*"*.

However the use of hands imparts the opposite reaction when the hands are used to support difficulties in conveying meaning. Another rater commented that

> *"There are too many gestures and that gives me the impression that they lack verbal resources* (rater 5 pair 12).

With the rater commenting negatively or positively on the same features, a plus or minus coding system could have been developed. However, there were insufficient comments in the data to warrant the development of such a system.

4.4.2 Interactive listening

The second category refers to raters' comments on the candidates' manner of displaying attention or engagement while listening during the interaction. Listening as part of successful interaction was divided into two subcategories: comprehension and a type of listening termed 'supportive listening'.

Comprehension, via verbal support, is a means for the listener of showing engagement, of giving encouragement for the speaker to continue, or of demonstrating comprehension as a listener. It can mean raters notice that candidates are filling a silence, asking for clarification or comprehension. In the case of filling a silence, the raters noticed that the candidate provided the word the other partner is searching for, as in the example: "*She sometimes filled in with a missing word to help (*rater 1 pair 3)". This shows the partner has been attending and comprehends sufficiently to predict a missing word, which enables the interaction to continue. Where one of the speakers did not fill a silence, this was also remarked on by the raters. The following examples involve two different raters on the same pair:

> *"I think she could have helped out finding the words when the other girl was thinking for a long while. You have to be patient but she could have helped with a word or a question"* (rater 10 pair 17)

> *"The girl is struggling with the next question. There is no attempt by the partner to help her out or to put words into her mouth. She just sat there and waited for her to get out of the predicament"* (rater 8 on pair 17).

Here the raters express how the person who is listening is not engaged, or not supporting the speaker, by either signalling for the speaker to go on or by taking the

floor and offering to break the silence. Candidates manifesting such behaviour represent un-interactive or 'unsuccessful listening' to the raters.

Comprehension, as a category, seems to indicate an individual and internal cognitive process. But what the raters were identifying in the peer discourse was that candidates were making attempts at 'constructing comprehension' or were 'negotiating comprehension':

> "*what he says relates to what has been said before. They make comments about what each other is saying and it makes the conversation look they are really interested* (rater 2 on pair 5).

What the raters are picking up on is the cohesion when there is a relationship between what is said and what follows. The comment 'makes it look like a conversation' refers to an engagement on behalf of the candidates, despite the fact that it is a case of test talk or institutional talk. This is what makes the fact that they sound interested in each other more surprising to the raters watching the video.

Comprehension also included cases where candidates offered clarification as they negotiated meaning. Clarification, as part of comprehension, also includes asking as in the example "*another way of facilitating is to clarify what the other person has said"* (Rater 10 pair 2). This enables the dialogue to continue.

Apart from the support offered during the negotiating comprehension, which requires clarification and filling silences, another type of support offered through listening attentively is not unlike non-verbal support. The difference between negotiating and maintaining comprehension and the latter is that the back-channel such as 'si, si' or 'aja' in the Spanish beginner dialogues are audible as well as non-verbal. This kind of support provides feedback while the other speaker maintains the floor. These are sounds that can be heard being made by the candidate who is interactively listening. These may be accompanied by non-verbal communication, such as gesturing or back-channel as in 'mm' 'si' 'ah' as in the example below.

> *"the girl with the glasses used a lot of back-channel a lot of confirmation mm, si, ah si she was very responsive (…) she used a lot of physical prompts to*

continue the interaction with more inflection and intonation in her voice" (rater 2 pair 3).

The distinction being made is that back-channel, which is audible, and gesturing, which is visible both serve as signs that the candidate is interactively listening and negotiating comprehension.

Despite encouraging the other speaker to continue, the sounds made by the person listening may not mean that person has been able to make meaning out of what is being said. The point is that speakers sometimes pretend to listen until they understand, or they choose not to listen but feign understanding because they are thinking of what to say next. The sounds included in 'supportive listening' do not indicate anything beyond interest or engagement from the listener for the speaker to maintain the floor.

Therefore, of the two types of interactive listening, only the first requires evidence of being able to negotiate comprehension by the listener. The second, back-channel, does not necessarily require an attempt at constructing comprehension, just audible support with sounds. Both types of interactive listening support the interaction and were attended to by the raters.

4.4.3 Interactional management

The third feature of successful interaction identified by the raters emerged from comments on the management of the topics and turns. This can be theorized from different perspectives. Between adjacent turns on the same topic it could be viewed as horizontal management, making the conversation flow. However, across topics it could be termed vertical management, exhibiting flexibility that allows switching between topics.

The raters' comments on turn taking show how interaction is managed horizontally when relating to speaker change. Thus elements connected to speaker change, such as speed of response, turn length or domination come under this category. For example:

"They are incapable of comprehending and replying quickly. He takes a long while to answer" (rater 10, pair 2)

> *"It is important to listen and to allow time to respond, to be sensitive to taking turns and not dominating" (*rater 10, pair 17)

These comments refer to two aspects of turn taking. Firstly there is the need to reply within a comfortable time, but it is also necessary to leave time for the other participant to respond. In the first example above (i.e. rater 10, pair 2), the candidate has the added problem of negotiating comprehension (cf. §4.4.2, above).

As long as candidates made sense across turns, by asking relevant questions for example, they were connecting horizontally. As in

> "*If one does not know what to say the other helps by changing the topic" (*rater 5, pair 12).

This example shows that the raters orient to how the topic is changed in order to keep it going. On the relevance of questions that keep the turns changing, the same rater continues with:

> *"and another thing that makes it natural is that for example they both interrupt each other with more questions" (*rater 5, pair 12).

The second type of interactional management is vertical management, which connects topics vertically down the complete oral text. Raters noted candidates' ability to connect topics when constructing comprehension. For example a rater commented:

> *"She is following the conversation and as a consequence it is all interrelated* (rater 1 pair 2)"

The 'it' that the rater refers to is the flow of the dialogue, which as a result of the two candidates negotiating comprehension is interrelated and judged by the rater to be successful. Similar comments were made on the candidates' ability to develop the conversation by extending the topic, such as:

> *"He finds something related to what has gone before"* (rater 1 pair 2)

Raters also commented on whether candidates facilitated interaction, such as in the following two comments:

"*topics are connected and he is following the conversation he tries to make sense with the person on the right"* (rater 6, pair 15)

"*In this conversation they are asking the right kinds of questions and it gives coherence to what they are talking about; quite good for beginners"* (rater 9, pair 2)

Both the turn change and the topic cohesion (i.e. horizontal and vertical management) were found by the raters to be indicators of successful interaction.

4.4.4 Interaction irrelevant observation

The final category was interaction irrelevant observations in which appraisals were made of the performance, of how pairs connected, or of the candidates' level of familiarity. These comments are irrelevant because raters were asked only to comment on successful interaction - appraisals of this type did not appear in the rating scales for comprehension and communication (Appendix 1). In addition to these appraisals, all comments made about the exam, the examiner, the task, as well as grammar and vocabulary were included in this category. This section is not expanded because it is not relevant to developing the construct definition of peer interaction.

4.5 SYNTHESIS OF FINDINGS & CHAPTER SUMMARY

4.5.1 Synthesis of findings: raters' views on successful interaction

The findings presented above can be divided into three categories:

1. Non-verbal interpersonal communication: *A speaker in a pair demonstrates visual and corporal non-verbal communication.*

2. Interactive listening: *A supportive listener in a pair offers verbal signs of comprehension or provides audible support to the speaker.*

3. Interactional management: *A speaker in a pair cedes to the listening interlocutor at a comfortable speed after a reasonable turn length and demonstrates the ability to change or develop connecting topics.*

What do these three categories mean to raters? What are raters observing that is different to what they have been asked to observe before? In response, it could be argued that all three categories show speaking to be social because it is not carried out in isolation. In particular the categories 'listening' and 'gesturing' drawn from the Verbal Protocols can be inferred to represent a measure of success in the social aspect of test performance.

This is an important development because if there is a problem currently faced by researchers, it is the need for language test-tasks requiring candidates to be 'communicative' with each other. Part of the problem is not having a precisely defined idea of how 'the partner' impacts practically or theoretically on the outcome of the candidates in the pair. Analysis of the introductory remarks in the Verbal Protocols presented in this chapter showed ways in which the presence of 'the partner', in this communicative situation, is perceived by the raters. In this study one position taken by the raters is that each partner in the pair is seen as 'an individual' taking part in a *paired* task. The other position found is one in which the rater sees both of the candidates as part of a whole 'a pair' performing a task

On this point, recall from §4.3.1 that in the Verbal Protocols, raters initially remarked on their first impressions, after having observed the entire performance. These opening comments state whether the rater thinks the paired interaction has been successful or not. These comments were used by the researcher to label the pair as 'successful' or 'unsuccessful'.

In the 'successful' pairs, the raters referred to the candidates as the plural subject pronoun 'they' in the comments during the remainder of the protocol. These instances were counted up and reported in Table 8 below.

For pairs labelled 'unsuccessful', based on the opening comments, the rater protocols referred to the individual candidates with the singular subject pronoun 'he/she' - with a much smaller number of 'they' comments.

Table 8: Instances of 'they' in successful versus unsuccessful interaction

	Candidate number	pair	Instances of 'they'	Instances of 'he'	Instances of 'she'
Unsuccessful	13	11	10		7
	14	11			7
Unsuccessful	3	4	5		7
	4	4			11
Successful	1	1	48		7
	2	1		9	
Successful	21	16	42		4
	23	16			7

Notwithstanding the small scale of the analysis, it shows how raters' judgements about interaction can be affected by how candidates position themselves within the pair. The inclusive 'they' references in the comments on the 'successful' pairs, show that raters view the pair to function as one social entity in the test context. Using 'they' as opposed to 'he/she' as indexes reflects a deeper perception of the 'oneness' of the pair. However, is this view is based on how the candidates handle interactional management? Or do raters perceive this 'oneness' from the non-verbal communication and the manner in which candidates listen?

It appears that raters' choice of pronoun for describing the participants reflects their view on the relative success of the paired interaction. Either the raters perceive the candidates as temporarily co-constructing one entity for the task, or as two individuals with a perceived lack of co-construction and perhaps less success in 'paired interaction'.

4.5.2 Chapter summary

This chapter explored the features of interaction in a paired oral that raters focus on as they observe a paired task. The aim was to identify features observable by raters in the set of peer performance, to be used as a basis for scale criteria.

We examined what L2 Spanish tutors claim to notice as salient features of interaction on paired tasks. We focused on the nature of language teacher raters' orientation to 'interactional features' and their perceptions of what interaction consists of.

The research question for Study 1 was **research question 1:** What features of peer interaction do raters attend to in paired task test performance? The analysis of Verbal Protocols suggested that raters attend to three main categories of interaction: non-verbal interpersonal communication, interactive listening, interactional management and a fourth that contained interaction irrelevant observations.

The results presented in this chapter provide a first perspective on how language specialists construe interaction on a (peer) paired task. This has implications beyond this study, for tests that employ the paired test as part of the oral battery for oral proficiency beyond this context. Bachman (1990:50) argued that we must provide

"clear and unambiguous theoretical definitions of the abilities we want to measure and specify precisely the conditions and or operation that will follow in eliciting and observing performance."

The results presented take us one step closer to this goal. Possible incorporation of some of these features into rating scales for paired tasks is a feasible option. Rather than attempting an impossible, inaccurate and impractical description of 'levels of peer interaction', the method used here, following Turner and Upshur's (1996) EBB rating scale, identifies the *boundaries* between levels rather than *describing* each level is proposed, as we will see in chapter 6.

But before moving onto scale development, we recall that as well as raters, candidates were also asked for input. The candidate view of peer-interaction was collected in order to ensure that the test conformed to their expectations of the peer testing

process. The candidate data aimed to validate what the raters thought appeared to be happening when they observed the clips. The researcher aimed to uncover whether raters observe the very same processes that candidates claim to have undergone. This made up Part B of Study 1 and is described in the following chapter.

Chapter 5: CANDIDATE ORIENTATION TO PEER INTERACTION (STUDY 1 PART B)

5.1 CHAPTER OVERVIEW & INTRODUCTION

5.1.1 Chapter overview

This chapter reports on the second part of Study 1, which considers candidate orientation to peer interaction. The chapter explores individual candidate orientation to performance in a paired task during a speaking test. Following the introduction, §5.2 describes the methodology. The data analysis in described in §5.3, and the findings are synthesized in §5.4. §5.5 maps the findings to those described in Chapter 4, which enables a clearer picture of both candidate and rater orientation to peer interaction.

5.1.2 Introduction

The purpose of the second study, that is. Part B of Study 1, was to explore the test candidates' orientation to the same paired task that was the basis for Study 1 Part A (i.e. Chapter 4). This was intended to determine whether candidates could make appropriate and meaningful comments on their paired performance in order to contribute to scale development and interpretation.

Candidates were asked to verbalize the process of interactive communication through stimulated verbal recall after taking their speaking test. What was said in the verbal protocol was analysed to see the extent to which candidates were aware of 'interaction' between them while they spoke to each other during their test. The intention was to answer the question: "How do candidates view their interaction in a paired oral?"

In Chapter 4, test-takers' success in interaction was judged by raters as a function of interactive listening, conversation management and nonverbal communication.

Overall language processing involves two levels: lower-level processing (e.g., automatic, unreflective and unconscious processing, especially in familiar and easy tasks) and higher-level processing (e.g., conscious, intentional strategic processing such as monitoring, planning and evaluation). In an attempt to understand the reasons why successful spoken interaction between the peers affects successful test performance in the way that it does, the notion of orientation is explored. Orientation is defined as an individual's ability to perceive and verbalise aspects of performance. Candidate orientations are used here to identify indicators of successful interaction.

The study investigates how the candidates say they work in the pair. Candidates' cooperation is in a crucial position between the task they are given and the discourse that emerges. Their views on interactional management provide insight into candidates' own perception of the construct 'peer interaction'.

To date, few studies in language testing research have attempted to understand the nature of L2 test-takers' paired interaction and factors that in turn, may affect successful interaction in L2 test performance as studied by May (2006). At present little is known of the extent to which *individual* candidates are aware of the factors that effect success during interaction. And if they are aware, it is not known the types of features that they notice. This chapter investigates these issues, focusing on **research question 2:** How do candidates view interaction in a paired oral?

5.2 METHODOLOGY

As is the case with an increasing number of studies, the research reported on here overlaps with several areas of applied linguistics. It is a qualitative language testing study looking at candidate cognition. This is facilitated by protocol analysis, which identifies themes in verbal reports elicited from participants as they verbally recall the demands made on them while speaking in a paired test. As detailed in §3.6.2, the Verbal Protocols here are Retrospective, non-mediated, think-alouds. Through

analysis of these Verbal Protocols, this chapter attempts to understand and explain the process of taking a language test with a partner.

In a verbal report, researchers typically ask participants to act and speak contemporaneously. As Wigglesworth (2005:103) states, this is "not appropriate for use with listening or speaking data because they necessarily conflict with the communicative nature of such activities". Here, to gain insight into what candidates notice, the protocol is drawn post hoc, making it retrospective stimulated recall.

5.2.1 Participants

Originally a large pool of students volunteered to participate in the study, but not in all stages. The 17 videos of pairs of candidates performing the test that were used in Study 1 part A all included one or both of the candidates in each pair who had consented to take part in the follow up candidate orientation study reported on in this chapter.

Out of the 17 pairs that took part in Study 1 Part A, 25 single candidates took part in the verbal protocol study. Both individual candidates of the eight pairs numbered pair 1, 2, 3, 4, 8, 11, 12 and 14 made up 16 of the candidates who returned to take part in the candidate orientation. This resulted in a complete data set for eight pairs. There were 9 candidates remaining who also returned for feedback but without their partners. These nine, together with the sixteen paired candidates, totalled 25 candidate participants in the study.

As mentioned in §4.2.1, candidates received personalised individual feedback on their performance in exchange for volunteering for the project. The feedback focused on errors in pronunciation, grammar and vocabulary. As the candidates performed the verbal protocol the researcher listed errors, and this list was given to the candidates as they left the protocol session. Apart from the error correction, the opportunity to watch their interaction during the paired performance was felt to be helpful. This was expressed in comments in some of the protocols.

5.2.2 The verbal protocol collection

The candidates watched the 10-minute video clip of their performance in the week following the oral. The stimulated verbal recall sessions lasted not much longer than 15 minutes. The candidates watched themselves on video, with a remote in hand. When they wished to comment, candidates set the video on pause and spoke into the tape recorder. They were asked the following:

> *"By observing your paired performance with another candidate comment on what you recall was 'happening' in the interaction. Please comment on all aspects of the performance including, but not restricted to, what or how something is said. Comment as you see examples of successful or unsuccessful interaction".*

A limitation of the methodology was that it was not possible to train the candidates because training time would have exceeded the feedback time. This could have compromised the data. During the recall a candidate said: "I remember all of this. I remember everything I said" (pair 4 candidate 2), which suggests that at least this candidate held the performance very vividly in her short-term memory. A summary of the data collected for the candidate study is set out in the table below:

Table 9: Candidate Data Table

Data	**Data set**	**Analysis**
17 x 10 min. videos of operational paired candidate orals Verbal reports by 25 candidates	Individuals watch their performance: 25 taped protocols 25 candidate verbal reports transcriptions	Content analysis of Verbal Reports into coding grid

5.2.3 Transcribing verbal reports

It has been suggested for verbal reports that the spoken discourse features such as intonation, pauses and speaking rate need to be accounted for (Afflerbach & Johnson, 1984). This study is interested in the content of the reports that is,. what is being said and not how it is said. Therefore it was unnecessary to transcribe spoken discourse

features and the reports were transcribed orthographically. Once complete, the transcriptions were checked against the tape for accuracy.

5.2.4 Segmenting

The sections of the transcription were delineated when candidates stopped and started the tape to comment. The protocols were segmented by these naturally occurring pauses. Each time the tape was stopped, a particular topic was discussed by the candidate, as prompted by the performance being observed on video. The segmented units were large enough to be coded in one idea or category. Neither their length nor their complexity made them difficult to code as a single category (see appendix 8 for examples of three candidate stimulated Verbal Protocols).

Themes and categories relevant to interaction were identified through a qualitative analysis of the transcripts. The researcher looked for parallels with the key features identified by the raters in part A. This resulted in grouping the protocols by types of comment similar to those made by the raters. The same categories that had been generated by the raters were used in order to see if candidates noticed the *same* features in their own interaction that raters had also found salient. Cycling through the data helped to clearly define the thematic categories.

It is acknowledged that the decision to use the same categories of analysis as those identified by the raters is a limitation of this study. If open coding had been used there would have been consequences for section 5.5 below where the two sets of data are discussed in relation to each other. Nevertheless this pragmatic procedure was pursued with the intent of uncovering whether raters and candidates perceived similar traits in paired interaction.

A coding grid with three columns was developed to code all the data, seen in Table 10. The last column on the right shows the categories that were identified in the first coding cycle. The middle column resulted from the second cycle of coding and shows a reduced set of data. It is important not to have too many categories because they can lose meaning or become unwieldy for subsequent intercoder checking. The first column shows the result of the final cycle through the data.

Table 10: Candidate content analysis expanded coding grid

Code name	**Feature**	**expansion of category**
Candidate interaction irrelevant reflection	Appraisal by candidates of own or other's performance	Good performance bad performance
Candidate interaction irrelevant reflection	Comment:	vocabulary or grammar other language interference on teacher/ partner video camera/2 minute buzz preparation
Interpersonal non-verbal communication	Visual signals Body or gesture facial	Comments: -On eyes up or down or gaze -On hands, legs, posture -On appearance, nerves
Interactional management	Topic change Speaker change organization Turn length Flow/ speed Silence	Getting new topic in Predicting/planning Taking floor or dominance Saying enough or too much Continuous speech Filling the gap
Interactive listening	Comprehension	Negotiating comprehension Listening to predict Not listening

Each of the categories listed under the column 'code name', that is, the left-hand column, correspond to the categories resulting from Study 1 Part A (Chapter 4). In other words, the candidates *had* commented on each of the same categories while watching their performance. These were:

- *'Interaction irrelevant' comments* where candidates spoke for example about how they had prepared for the oral
- *'Interpersonal non-verbal communication'* where candidates commented on where they were looking, how they appeared and made references to their body or their partners'
- *'Interactional management'* where the candidates commented on predicting and planning within the interaction and taking a turn or introducing a new topic
- *'Interactive listening'* which encompassed negotiating comprehension, listening to predict, and filling gaps

These categories can be seen in the left-hand column of Table 10. Using these categories as a reference point in Study B, the coders were able to code for them by redefining the categories from the candidates' perspective and then making them

consistent with the rater content analysis data. So the coding categories remained, but the features they encompassed were modified. A distinction was drawn between interactional management from inside the mutually dependent context between the candidates and for an outside observer of that interactional context.

5.2.5 Calculating encoder reliability

From the coding grid in Table 10, above, another rater could recode the data to check intercoder reliability. All the data was coded twice.

The second coder was an applied linguist with experience in qualitative research, in another country. The coder was sent the material to code in Excel files. The segments had been turned from text in word to tables in Excel. Of the 465 segments coded, 372 were coded the same by both the researcher and the code checker, and 93 were coded differently. The second coder and the researcher arrived at an 80% agreement, which, as noted in Chapter 4, falls within the acceptable range (Gass & Mackey 2000). Following the completion of the independent coding process, areas of disagreement were clarified and consensus was reached for the remaining 20%. Table 11 shows that of the total 424 segments, the smallest category (candidate reflections on gesture and body language) had 14.4% of the comments.

Table 11: Percentage of coding: candidate orientation

Candidate content analysis	**% of 465 segments**
Other reflection not on interaction	28
Non-verbal interpersonal communication	14.4
Interactional management	28.3
Interactive Listening	29.3

The other comments were divided fairly closely to make up the remaining 85.6%: interactive listening (29.3%), interactional management (28.3%) and reflections not on interaction and appraisals (28%). Each of the four categories listed in Table 11 is explained in turn in §5.3.

5.3 DATA ANALYSIS

5.3.1 Non-verbal interpersonal communication

Although non-verbal interpersonal communication was the smallest stand alone category numerically, it is nevertheless still important. It can be inferred from the comments made by the candidates that they use non-verbal communication for providing cues during the interaction, as well as for reading interactive cues from their partner. The comments that candidates made about non-verbal interpersonal communication were grouped into gaze and gesture, laughter, body position and facial expression. Each of the four groupings is dealt with separately below, with examples taken from the transcriptions of the Verbal Protocols.

5.3.1.1 Gesture

In the comments below, the candidates refer to their use of gestures to explain themselves. Gesture is an extension of their vocabulary.

> *"I tried to explain handball with my hands but I got lost so I didn't know how to say it so I just laughed"* (Pair 1, candidate 3)
>
> *"I didn't know how to say that with my hands"* (Pair 4, candidate 21)
>
> *"I was so nervous and it was so good that she did that. That helps that hand movement"* (Pair 4, candidate 2) (from watching the video of the performance that the comment was made about one can see that candidate 21 is miming eating so that candidate 2 can understand.)

In contrast to the comments above where the gestures are used for vocabulary, in the comments below, the candidates use a wave or a nod consciously to provide their partners with a signal to proceed with the next question during the turn taking in the interaction.

> *"I did that hand wave for her to ask the next question"* (Pair 8, candidate 27)
>
> *"I am waiting for him to pause and nod like… next one… so I know from the nod…"* (Pair 10, candidate 26)

These gestures organize the interaction, rather than conveying meaning as in the other comments on gesture listed first in this section. Using the candidate stimulated recall enables clearer comprehension of the discourse and interpretation of the interaction

between them. Candidates' awareness of the fact that gesture can contain meaning or can organise interaction shows that gesture contributes to success in this mutually dependent context.

5.3.1.2 Gaze

The comment below is on gaze. It explains how an intense gaze is actually giving the candidates thinking time in which to prepare themselves for the next move in the interaction.

> *“he was losing me you could see the blankness in the eyes he had lost me. Instead of freaking out and looking away we locked eyes and I worked out he was thinking I don’t know what you’re saying”* (Pair 13, candidate 6)

This gaze allows one candidate to decipher that the other is unable to continue the interaction. The second comment, below, is from the same candidate, referring to the co-constructed nature of the interaction:

> *"It is a just a comfort sort of thing to look down, when you speaking you are in nowhere land where you have to speak, and we are both looking down for just something”* (Pair 13, candidate 6)

When the candidate says "when you speak you are in nowhere land" he means that as a beginner monolingual he is out of his comfort zone. He must build the test outcome with his partner. Candidate 6, who made the comment above, was very prepared. The other candidate, his partner, thought he did not need to prepare, as a bilingual Italian speaker. In the oral they had a lot of trouble communicating. They were both mature age students and candidate 6 was very keen to do well, because he wanted to apply to go to Mexico on exchange to study a post grad diploma in film. By looking down at the paper "for just something" he takes time out of the intense situation they find themselves in.

5.3.1.3 Laughter

There were different references to the function of laughter by the candidates. The first example shows how affiliation is expressed and acknowledged through laughter:

> *"I think a lot of my laughter is because I am trying to affirm her. She is always saying "you know more than me. I don't know much" Girls are taught to apologise and make the person comfortable all the time"* (Pair 16, candidate 19)

In the second extract the laughter is an expression of nerves.

> *"I laughed and made a noise out of embarrassment"* (Pair 1, candidate 7)

Lack of comprehension also results in laughter from candidate 7 who is embarrassed:

> *"I am laughing because I don't understand anything"* (Pair 4, candidate 2)

It can be inferred from these comments that laughter has different functions in interaction and the candidates appreciate the differences between them.

5.6.1.4 Body position

Candidates were not likely to have been aware of their body position. They are unconscious of how they look and how they impact on the other person.

> "*shaky legs don't look good I must have been nervous"* (Pair 13, candidate 6)
>
> *"I look so weird I look so upright" (*Pair 4, candidate 2)

These two comments express surprise at the body position they are observing as they watch the video. They are most likely unaware of the visual image they are projecting and are commenting how that kind of position is not what they expect for communication. A 'stiffly upright' demeanour is not encouraging for a partner to warm towards another candidates if they look like they will not be flexible or move to

cooperate with the non-verbal fluency which is 'the dance' of non-verbal communication happening as people engage in interaction with each other.

5.3.1.4 Facial expression

The comment below on facial expression refers to the lack of authenticity of the activity.

> *"Look how well she is pretending to look interested. . She has heard it all before"* (Pair 14, candidate 17)

The candidate making the comment observes that the listener should be interested in the speaker, even if they have heard it before.

In sum, laughter, body position, gaze, gesture and facial expression were all an expansion of the feature 'non-verbal interpersonal communication' in the coding grid for the content analysis. It is important to note that comments from the candidates are evidence that non-verbal communication, in all the instances above, except for body positions, is integral to candidate performance. We saw in Chapter 4 that raters also commented on candidates' use of non-verbal communication.

The changes in gaze, gesture and body position described by the candidates could be described as a visual non-verbal fluency. This visual non-verbal fluency is a spatial layer of meaning added to the verbal and audible fluency of 'words' and 'other sounds' that convey meaning.

5.3.2 Interactive Listening

5.3.2.1 Comprehension

There were many comments on 'signalling the state of comprehension to the interlocutor [or other viewer]' in the candidate protocols. Listening interactively is the mirror image of speaking in a conversation. While not speaking, listeners need to be in a position of openness to negotiating comprehension with their interlocutor. Most of the comments were about signalling, and how accurately it reflects the actual comprehension:

"I am listening because I didn't know where she went to school and that" (Pair 11, candidate 1)

"I am laughing because I can't understand anything" (Pair 4, candidate 2)

"I don't know what she is talking about I just say 'si'" (Pair 4, candidate 2)

"I had no idea what she was saying I just kept saying si all the time" (Pair 7, candidate 5)

"I tried to concentrate to see we understood each other" (Pair 16, candidate 19)

"she said something I couldn't understand so I said no to stop her saying more" (Pair 9, candidate 15)

"most of the time I could understand what she was saying but at that point I just couldn't under stand so I just kept saying si si" (Pair 12, candidate 25)

Many of the examples above support part of the definition that evolved, in section 5.3.2, which says 'to provide audible support to the speaker'. In other words, they had listened but had not understood at all. Listening and comprehending are different parts of successful interaction. The fact that candidates are so aware of their ability to signal comprehension, or lack of negotiating it, underlines that they are very dependent on this skill. The impact of not listening, or, on the other hand, listening to predict an interactive move, is discussed separately below.

- **Not listening**

Not 'listening' and avoiding negotiating comprehension can effect the interaction and some candidates admitted that they were not paying attention during particular moments of the interaction. For example:

"She is kind of listing things so I don't have to pay much attention" (Pair 1, candidate 8)

In this comment the candidate feels that there is no need to negotiate comprehension because their partner is wasting valuable opportunities by listing details. This encourages the partner to shut down the negotiating of comprehension, knowing that he/she is not part of the listing, which is just for display in the test.

"the problem is that I wasn't really listening to what she was saying. I should have helped her out" (Pair 8, candidate 27)

This is another example of the absence of interaction and an acknowledgement by the candidate of the joint responsibility for success.

- **Listening to predict an interactive move**

The amount of candidate attention to predicting and connecting within the dialogue is evidence of how acutely aware candidates are of 'making sense' of what has gone before and predicting what may lie ahead.

"I was trying to figure out what he as going to ask about my parents like how old they were" (Pair 1, candidate 7)

"you look at your thing and you think ok I can ask that later on but you keep on listening you have to think of something that is related" (Pair 11, candidate 1)

"I am listening and trying to think what she would say" (Pair 11, candidate 3)

Here it is not a case of listening for 'understanding only' in the psycholinguistic sense, that is, an individual cognitive process without the need for 'an interactive other'. As candidates shift from the position of speaker to the position of hearer they still have a responsibility to move the dialogue forward, as active listeners. In the test task the candidates needed to introduce three topics, in addition to responding to their partner's questions on other topics. The success of the interactive communication for them is a combination of listening and taking a turn or introducing a topic, which falls under 'interactional management'.

5.3.3 Interactional management

Analysis of candidate protocols shows that candidates are able to focus on their own performance and verbalize their interactional management. An example of a candidate recalling a moment during an exchange, and trying to predict what to do next, is the following:

"I am thinking. Where is she taking this? Where is she going to go next?" (Pair 1, candidate 8)

A content analysis of the candidate protocols in the interactional management category shows that candidates focus on three areas: topic change, turn organization and turn length. These areas are defined and expanded below with examples taken from the data.

5.3.3.1 Topic change

Candidates face an internal struggle to read the interactional cues in order to facilitate a topic change.

> *"I think I am waiting for a change of topic"* (Pair 16, candidate 19)

This is an example of the candidate waiting instead of being proactive and directing the discourse to the next topic.

> *"I was trying to think how to introduce this into the conversation. How do I ask without sounding as if I was interrogating her"* (Pair 16, candidate 23)

The concern shown by the candidate above exemplifies the interactional task of not interrogating, but managing to steer to the new topic.

In the following quote the candidates refer to a previous test experience, in which a candidate felt a lack of control of topic:

> *"We were quite concerned about getting through all the topics as we missed some last semester. I think that is why we had lots of questions and answers and not long bits"* (Pair 14, candidate 17)

Similarly:

> *"I am not helping her I felt I didn't have control at all over the situation over the way we were going"* (Pair 14, candidate 18)

Each of the quotes above illustrates the manner in which candidates recall the internal struggle to read the interactional cues.

Candidates are aware that in order to facilitate a topic change they must choose the right moment, as in the following examples:

"just had to get my topic out and he sort of looked at me and went ah now we are talking about this" (Pair 1, candidate 7)

"we had gone through the neighbourhood shops I am trying to figure out how to get to the next topic" (Pair 1, candidate 8)

"I was trying to get back to our topics" (Pair 3, candidate 12)

"I didn't really know where we were going but it was good 'cause I generally tried to say something" (Pair 7, candidate 5*)*

"I didn't know how to ask about it so I just avoided it" (Pair 8, candidate 27)

"I didn't like the questions so I just changed it to what I wanted to talk about I was thinking at the same time and I knew which one I was going to ask next" (Pair 8, candidate 27)

"the topics were cool but what kind of a question the other one was going to ask was yeah a bit of an unknown" (Pair 13, candidate 6)

All examples above illustrate how aware the candidates are of entering the interaction appropriately.

5.3.3.2 Turn organization

This category refers to the organization and planning involved in changing speaker, in response to the partner in the interaction. The candidates had not planned or prepared their sequences, so that the turns occurred as 'naturally' as can be expected during a test:

"We started talking and then after that it was whoever got in next" (Pair 6, candidate 10)

Candidates innately know that long silences are not right in an oral. The candidate in the quote below explains that the silence was due to the time needed to process what to say.

"I know it's my turn. There's a silence. I feel it is my turn…There were a lot of silences. We were trying to think of what to say and the best way to say it and it took a while" (Pair 12, candidate 25)

The difficulty of organizing the turns is also candidly acknowledged in the instances below:

"it was a bit hard to pick up when to talk to each other" (Pair 3, candidate 12)

"I thought we worked well together we helped each other out when both knew when to ask a question. We asked the questions randomly and answered and took it from there from wherever it went" (Pair 10, candidate 26)

"you think about the next question and what you expect her to say so you plan your answer" (Pair 10, candidate 26)

"We both knew we both had to have even turns so" (Pair 12, candidate 25)

*"we started each topic then we would to and fro when you sort of get tired of a topic you think OK you go now. For them to ask a question and then you are a bit more relaxed" (*Pair13, candidate 6)

The examples illustrate the expectations candidates have of each other during the turn taking.

5.3.3.3 Turn length

The relative supportiveness of the candidates signals a cooperative engagement between speakers on turn length. There is a 'helping' element as candidates support each other through the interaction. This is illustrated by candidates' reflections on the decision-making that occurred during the performance. Decisions whether to wait or to take a turn affect both parties.

"I am trying to think of what to say next. I am thinking like, should I help him or should I let him keep on going." (Pair 8, candidate 2)

"I was sort of thinking at the same time and I knew which one I was going to ask next I should have helped her out there." (Pair 8, candidate 27)

The candidate protocols show two main decision making issues. One is managing the content; that is, thinking what to say. The other is managing the interaction; that is, choosing the moment when to take the floor from the partner and for how long.

Speaker change as connected to turns and length was also a constant thread through the simulated recalls:

"we both went for as long as we could go and if they asked a question we got the other to elaborate" (Pair 13, candidate 6)

"if there was a break or a pause we just asked another question" (Pair 10, candidate 21)

"so then I just filled in the silence" (Pair 1, candidate 8)

"she went ahead and asked all the questions which made me nervous because she talked a lot" (Pair 3, candidate 12)

"we had planned we would ask a question each but she dominated the conversation" (Pair 3, candidate 12)

"I am waiting and thinking how am I going to answer" (Pair 7, candidate 5)

"now it is your turn I am thinking what more can I say finding words I remember" (Pair 8, candidate 28)

"I found it a bit difficult some times because I waited for her to ask a question back, but she didn't so I had to fill in" (Pair 9, candidate 15)

"I was going to say something more but I forgot so she saved me" (Pair 11, candidate 1)

"there is a silence I know it is my turn" (Pair 12, candidate 25)

These examples taken from the content analysis on interactional management, shed light on 'speaker change' during paired interaction which has been shown to be difficult to describe and rate (Nunn, 2000; Orr, 2002; Van Moere, 2006).

5.3.4 Other candidate reflections: a comment about pairs

The comments in this section are not relevant to communicative interaction. They refer to exam preparation (§5.3.4.1) and linguistic strategies for vocabulary and grammar (§5.3.4.2).

5.3.4.1 Preparation

Preparation is part of what is expected of a student: they are meant to prepare for exams, which could explain why they comment on it as they observe their performance.

"when we practiced we went off the topic a lot more. It was more relaxed in the practice" (Pair 7, candidate 5)

"we never practiced we never prepared a thing" (Pair 4, candidate 21)

"we practiced so much before we came but for one reason we got really nervous" (Pair 8, candidate 28)

5.3.4.2 Linguistic comments

Second language speaking strategies such as circumlocution or a visual recall were used as a memory aid:

"I tried to say rent videos but then I couldn't so then I had to think of a way around it" (pair 9, candidate 15)

"I am doing the verb thing in my head you know you see the verb page in the book to find the endings" (pair 14, candidate 18)

The comments are from within the candidate, before the utterance is made. They are included to illustrate the category, but no further inferences were drawn from the candidate comments grouped here.

5.4 SYNTHESIS OF FINDINGS: DEPENDENCY IN INTERACTIONAL MANAGEMENT

This section synthesizes the findings from the content analysis of the Verbal Protocols and considers how the candidates position themselves and view their partner through the paired discourse. It continues to answer the research question informing Study 1 part B of the complete study "How do candidates view their interaction in a paired oral?"

From the categories drawn from the candidate comments in their stimulated protocols, it could be inferred that candidates had different levels of dependency. Students made comments about themselves and their partner in the protocols regarding dependency as they negotiated the co-constructed dialogue. These levels of dependency are set out in three categories below: co-dependent (§5.4.1), inter-reliant (§5.4.2), and inter-dependent (§5.4.3). These levels are defined by how different candidates position themselves vis-à-vis their partner during the test.

The degree to which candidates were successful in their communicative interaction appears to depend on how the candidates positioned themselves. That is, as individuals trying to be part of a pair, or as an individual subsumed into a pair.

5.4.1 Co-dependent

The comments of this type are about the candidate being 'an individual' that has to survive the paired interaction. They consider, but compete against their partner. In the extracts it can be seen how much candidates affected each other in this type of co-dependency. This is a negative situation. It is the opposite from 'co constructing' a dialogue. As they interact it is for themselves, to benefit their own performance.

> *"I was wondering what to say. I was thinking that he was talking too much and I needed to be more active. It's funny 'cause he also thought that….I tried to move into another topic but I wasn't quick enough so he jumped in"*
> (candidate 15, pair 24)

Here even by quoting just one, it seems that both candidates know they need to display their language proficiency and both think the other is talking too much. The co-dependent partner thinks of *themselves* primarily during the interaction. They think their success lies in the other allowing them the space to talk. They do not want to construct a dialogue that might benefit the other - they want to work to achieve the best outcome for themselves.

5.4.2 Inter-reliant

The comments of this type place *the problem* in the interaction with the partner. It is a better situation because candidates attempt interaction - including and not competing against the partner. It differs from co-dependence, in §5.4.1, in that they are not two individuals competing against each other. Here the focus is on the partner's skill as an interlocutor. The inter-reliant individuals come to the paired interaction with *the intention to attempt to interact*. During the test, the inter-reliant candidates are aware of the impact that their partner's shortcomings have on them:

> *"We had planned we would ask a question each but she dominated the conversation. She went ahead and asked all her questions which made me nervous because she talked a lot and put me off a bit and made me nervous."* (pair 3, candidate 12)

They blame the other for a problem or for the lack of success. But on the other hand if you provide your partner with the opportunity to ask questions, but they don't, then as a candidate you have a reduced chance to demonstrate your ability. For example:

> *"I was waiting for her to introduce one of her topics so I jumped back in with one of mine… A lot of the time when she spoke I couldn't understand what she said but I just picked out the key words and then put them together"* (pair 9, candidate 15)

In this case the other partner is being blamed for not asking a question, but also for not being comprehensible.

There is also a problem of not taking the floor despite the fact that it is offered, for example:

> *"She had run out of questions. Every time she asked a question I didn't know the answer, so she had to move on to another one"* (pair 4, candidate 2)

The candidate is aware of relying on the other to keep the conversation moving, and that the fact that her partner had run out of questions or strategies for keeping the dialogue going was unfair. The fault or the problem for lack of success lies in the partner. Though willing to interact they find fault with the other even as they are performing.

The difference between this category and co-dependence is that the candidates in the latter think only of themselves. In this second category the candidates blame the other candidate for whatever issues may have arisen. This could be the partner dominating, insufficient questions, no new topics, incomprehensibility or not answering, all of which emerged from the discourse. Candidates are aware of these matters and are concerned about the impact of their partners on them with regard to conversational management.

5.4.3 Inter-dependent

Comments of this type demonstrate how one party builds on the other for success in interactional management. In this case the individuals are totally subsumed into the pair. It is a positive situation. For example, candidates try to change topic and wait for the right moment. They work at the interaction together on the assumption that it will be a success for them. The comments from both candidates in pair 1 show how their thinking is parallel regarding the interactional management:

> *"I was trying to figure out what he was going to, what he was going to ask."* (Pair 1, candidate 7)
>
> *"I am thinking where is she taking this where is she going to go next."* (Pair 1, candidate 8)

The effort put into predicting means that the interaction makes sense because they pick up the interactional cues, though presumably all of the pairs are faced with the same task and the same dilemma. Comments in this category show the pairs were confident and glad to work with each other. They helped each other and managed the interaction in a way that that they both felt comfortable. This is borne out by their comments on their paired performance:

> *"I was glad to be with him because I thought that we worked well together we helped each other out."* (Pair 10, candidate 21)
>
> *"I asked a question. I thought it was her turn. That's what I was feeling confident with."* (Pair 10, candidate 26)

They knew how each of them would respond to questions or to trouble in the interaction. The pairs that supported each other demonstrated how aware they were of what their partners were going through.

> *"I was going to say something more but I forgot so she saved me"* (Pair 11, candidate 1)
>
> *"There she probably didn't know what to say so I thought I would ask her a question to get out of it"* (Pair 11, candidate 21)

This fine-tuning of the interactional management requires a high level of engagement during the interaction.

In sum, candidates were not only conscious of the cues from the context that they need to change turns or change topic but they were acutely aware of each other. Candidates were at different stages, for example the isolating 'what will I do next' co-dependent situation; and the inter-reliant stage of noticing what the partner was doing wrong. The best interactional management requires inter-dependent candidates.

5.5 MAPPING THE TWO DATA SETS PART A AND B

We have reached the end of Study 1 that aimed to define peer –peer interaction from two perspectives. The questions asked were:

> **Research question 1** What features of peer interaction do raters attend to in paired task test performance?
>
> **Research question 2** How do candidates view interaction in a paired oral?

We now have two sets of analysed data showing the orientation of candidates and raters towards paired interaction. The mapping of the data set is presented in two sections.

5.5.1 Differences between raters' and test takers' orientation to peer interaction

The methodological similarities are that the data collections both come from think aloud Verbal Protocols and they are both carried out using the same set of paired performances of paired speaking tests. The differences are that in Part B the Verbal Protocols were retrospective whereas in Part A the observer rater used cues to judge and interpret what occurred between the candidates. In Part B the candidates described what they thought was happening or what they were aware of during the interaction with their candidate partner. In mapping the findings from the two Verbal Protocols and comparing them, we can uncover the focus of raters *observing* successful interaction, and the focus of candidates *creating* successful interaction to 'pass'.

The raters and test-takers oriented towards the same three categories of features: interactive listening, conversation management and non-verbal communication. This reflects both the raters and the individual's sensitivity to what comprises effective interaction in a paired peer task. The test-takers were acutely aware of the demands made on themselves and their peer pair candidates during interaction in a paired performance. These demands were heightened particularly when the level of difficulty in interacting increased by having an 'uncooperative' partner.

The inferences on dependency made from the candidate comments on each other imply that 'easy to interact with' or 'difficult to interact with' partners affect the nature of successful speaking performance which in turn affects the way that test-takers are perceived by raters.

5.5.2 Similarities in awareness between raters and test takers

In Chapter 4 we saw a set of key features that raters oriented to while observing paired interaction. These same categories were used again when cycling through the transcriptions of the candidate retrospective verbal recall. While it was intended that findings would be compared according to the *same set* of key features, the close overlap between them was surprising. Table 12 shows the percentage of comments made by candidates and raters on the same key features that were commented on in interaction.

Table 12: Quantity of key features identified by candidates and raters

Key features identified through content analysis	**Percentage of total candidate coded comments**	**Percentage of total raters coded comments**
Reflection not on interaction	28%	16.3%
Interpersonal non -verbal communication	14.4%	17.4%
Interactional management	28.3%	48.7%
Interactive Listening	29.3%	17.6%

Raters of oral proficiency are trained to focus on a specific area in an oral test. If they are asked to comment on interaction, it would be expected for them to leave out extraneous comments. Candidates on the other hand are not trained to focus on language performances from an observers' perspective. They were asked to comment on their interaction. It was not surprising that nearly a third of their comments, 28%, were on topics other than the interaction.

Of the total number of segments coded as non-verbal communication there was a proportion of 14.4% for the candidates and 17.4% for the raters. This shows that raters and candidates are equally concerned with the impact of non-verbal interpersonal communication in spoken performance assessment.

It is interesting that the candidates gave equal weight to comments on listening (29.3%) and on interactional management (28.3%). This shows that candidates are aware of the importance of moving from speaker to listener during dialogue. As they speak, their focus and full attention was on reaching the listener. On the other hand the raters, for whom focusing on the listener - not the speaker - is unusual, made only 17.6% of their comments on how candidates listened. The fact that trained raters focused on interactional management, at 48.7% of the comments, is not surprising because although such a direct report from raters using VP has not been carried out by Cambridge, existing studies based on CA analysis of candidate discourse (Lazaraton 2002, Weir 2003) showed that working in pairs allowed the functions to be displayed which only appear when candidates can initiate dialogue. Candidates also focused on interactional management (28.3% of their comments), which was almost the same proportion as for their comments on listening.

In answer to the second research question for this thesis, which asks how candidates view interaction in a paired oral, it can be said that candidates view interaction as a combination of three features: interpersonal non-verbal communication, interactive listening and interactional management.

The results of this candidate orientation study have provided empirical evidence validating raters' perception of interaction and these results have also provided a

framework for describing candidate and rater ways of conceptualizing interaction. The differences between candidate and rater lie in the percentage of focus on the categories in Table 12, but the key features still confirm each other. In other words, they are not at odds with the raters, who notice the same features that candidates focus on. The results of this chapter have also shown that candidates make appropriate and meaningful comments on their paired performance, which could contribute to scale development and interpretation.

5.6 CHAPTER SUMMARY

This part of Study 1, Part B, has explored the issue of rating pairs by asking candidates how they view paired interaction. Language testers can aim for more exact measurement, by untangling how candidates work separately or interdependently. Following this 'untangling', the type of mark (joint or individual) could be more justly decided for 'interaction' in a pair. This foreshadows an implication of this study involving single or joint scores, which will be taken up in the final discussion in Chapter 7.

The two parts of Study 1, that is Chapters 4 and 5, investigated orientation to peer interaction in order to help define it from a rater and from a candidate perspective. By mapping the findings of Part A and Part B it was found that the test-takers and raters oriented to similar features that demonstrated successful or unsuccessful interaction with a pair in a speaking test performance.

Not all candidates or pairs were successful in terms of interactional factors (recall Chapter 4). It was found that the test-takers that lacked in the three components that the raters had previously oriented to were correspondingly less successful (§4.5). The fact that those pairs or candidates that were more successful manifested increased amounts of the qualities oriented to by raters might explain why successful interaction was easily identifiable (in Chapter 4) in their opening comments (§ 4.5.1) after a first viewing. Empirical scales will be devised in Study 2, which will reflect raters' view of paired interactional management.

Chapter 6: DEVELOPING A DATA BASED RATING PROCEDURE FROM OBSERVED PEER INTERACTION (STUDY 2)

6.1 CHAPTER OVERVIEW & INTRODUCTION

6.1.1 Chapter overview

This chapter introduces the second study in this thesis**,** which is concerned with developing a data based scale from observed peer interaction. §6.2 reviews the research context and site, and the methodology is described in §6.3. The scale development procedure is described in §6.4, including the adaptation of Upshur & Turner's (1996) method. In the final section there is a discussion.

6.1.2 Introduction

Study 1 Part A and B utilized Verbal Protocol analysis to explore orientation towards interaction between candidates in a paired oral. Think aloud Verbal Protocols were elicited from university tutors and lecturers, who were Spanish L2 specialists and raters, and from candidates who were L2 Spanish beginner students. The results from both raters and candidates uncovered the key features that reflected the view of the majority of the participants.

The purpose of Study 2 is to take the key features from Study 1 and further define them through a data based process of scale development. This practical procedure involves an empirical scale development process, using teachers as scale-maker and experts.

6.2 REVIEW OF RESEARCH CONTEXT AND SITE

Two strands of research provide the background to this study. One strand, in §6.2.1, is the development of rating scales, and in particular data based scales. (Recall that the concept of data based scale was introduced in Chapter 2 on issues relevant to assessing speakers in pairs.) The other strand, in §6.2.2, is concerned with rating spoken interaction, in particular between peers.

6.2.1 Developing empirical rating scales

As reported in Turner and Upshur (2002), rating scales have been criticized for producing scores with low validity and reliability. Improving the rating criteria could improve reliability and validity (Hamp-Lyons, 1991; North, 1995; North & Schneider, 1998b, 2003).

Scale development methods are basically divided in two types: intuitive and evidence-based. Although the intuitive method (using prior knowledge and consensus among experts) is by far the most common way of arriving at rating scales, the evidence-based empirical method, which works *from* language output samples *towards* the descriptors, is the method chosen for this study. A rating scale based on what raters observe and notice during peer interaction aims to address the validity problems noted above. It answers calls from the literature such as those in Chalhoub-Deville (1997), who cautions that theory alone is insufficient to produce task specific scales and Fulcher (2003) who directly calls for empirically developed rating scales.

The development of evidence-based scales for rating paired orals is further motivated by the fact that paired orals have been included comparatively recently into test batteries. With less research into the peer-peer construct, it is difficult to gauge the features that theoretically may be salient to raters in peer interaction. It has been said that assessment that takes the salient features of a task into account can improve measurement (Pollitt & Hutchinson, 1987) but taking salient features into account is difficult if they have not been shown empirically to be salient from a rater perspective.

6.2.2 Rating paired orals

In Chapter 4, on rater orientation to peer interaction, raters were found to heed eye contact and non-verbal communication. A content analysis of the protocols suggested that the raters oriented to three main features as salient for interaction: interactive listening, non-verbal interpersonal communication and interactional management. The next step is to build scales with raters in the role of scale makers. The focus of this chapter fits with a longstanding call from the field for including "in a scale what raters attend to" (Pollitt and Murray, 1996).

In light of the varying perceptions by scale makers, and the varying severity that results, the difficulty of rating pairs has been of interest. Two very important questions arise that must be addressed. Firstly, we should consider whether the process of 'communicative interaction' (as it is called by Cambridge ESOL) is a construct that can be adequately operationalized in such a way that raters "understand the model of communicative ability on which rating scales are based" (Orr 2002:153). Secondly, we should consider whether communicative interaction is scalable in the same manner that linguistic abilities have traditionally been scaled, by band level with accompanying descriptor.

Recall the preliminary discussion of the EBB scaling procedure in Chapter 2, including some of the concerns raised by Turner and Upshur (1995). The EBB procedure is described in more detail to enable a replica of the study to be carried out.

6.2.3 Review of the site

The aim is to develop a measure of speaking ability, in particular for 'communication and comprehension'. The scale will be developed from raters' observations of live test performance. The type of rating procedure to be developed is an Empirically-derived Binary-choice Boundary-definition (Upshur and Turner 1995). This method is based on samples of candidate performance, which mark the difference between levels with a criterial yes/no question. The research question that focuses Study 2 is **research question 3:** Can candidate peer performance samples from a paired test form the basis for developing a rating procedure for interaction?

The results from Study 2 will generate an empirically developed rating procedure, relevant to the performance of tasks in pairs. The rating procedure will reflect the interactional features attended to by trained raters and participating candidates in their observation of peer test performance.

6.3 METHODOLOGY

The original methodology of the Empirically-based, Boundary Bound, Binary-choice (EBB) method (Turner and Upshur 1996) is outlined in §6.3.1. In the section that follows, the adaptations made to this scale development process for the present study are detailed.

6.3.1 The EBB scaling procedure

Turner and Upshur (1995), and follow up studies Upshur and Turner (1996, 1999) describe a scaling procedure that "is *empirically* derived, requires *binary* choices by scale makers and defines the *boundaries* between score levels" (1999:82). It leads to "a hierarchical sequence of attribute checks" (Turner and Upshur, 1996) requiring raters to make binary choices about the salient features of student performance. Upshur and Turner's scale development project was conducted in a French medium school in Montreal, Quebec and aimed to provide reliable assessments of ESL speaking ability. The scale development was based on a sample set of 12 performances on each of two tasks: a story retell and an audio-pal which involved a taped 'oral' letter. The participants were twelve teachers, as test developers and scale makers and 36 grade 6 ESL students.

It is important to bear in mind the points from Upshur and Turner (1996:61) for this type of scale development, in which the scale developers:

1. Let actual performances tell what elements of the property space actually occur
2. Do not assume what variables are important at different levels
3. Let scale include only as many discriminable levels as raters can use reliably
4. Assure that all levels are used
5. Procedures for constructing scales are explicit
6. Scoring is efficient both in training time and rating time
7. Incorporate knowledge and procedures followed by experts

Following these recommendations, and using the EBB method of dividing the discourse sample into groups, Scales are empirically derived by using a question that marks the point of difference. Turner and Upshur (2002:55) summarise the procedure:

"a group of scale constructors, generally L2 teachers, is given a sample of writings or recorded oral performances. Working without a rating scale, the raters first arrive at a consensus on assignment of the sample performances into an identified number of levels and then identify and describe salient features that distinguish performance at adjacent levels". The main feature of the procedure is the focus on the point of difference between the levels. This focus differs from what usually is the norm for creating descriptors, which is to continue describing what can be done at each level till there is sufficient difference between the levels being described. (For greater detail on rating scale development see §2.4).

There are five tasks to develop the scale. Here they are first presented as Turner and Upshur intended for them to be followed. The adaptations are then described in §6.3.2. The order of the EBB as set out by Turner and Upshur (1995) is as follows:

Task 1:

- Rank the candidate performances.

Task 2:

- Divide the sample into two groups: an upper level and a lower level
- Identify the most salient attribute of interest that divides the sample of collected data. Form a yes/no question about the attribute that divides the sample into those with or without that attribute. The question should refer to an observed difference that is relatively easy for teachers to recognise.

Task 3:

- Identify how many score levels the sample can be divided into.
- Rank the upper level sample, *with* the salient feature, from task 2. Identify the most salient attribute of interest that divides the level. Divide the sample into two groups with or without that attribute. Form a yes/no question for that attribute.
- Rank the lower level sample, *without* the salient feature, from task 2. Identify the most salient attribute of interest that divides the level. Divide the sample into those with or without that attribute. Form a yes/no question for that attribute.
- Repeat until there are no more viable divisions.

Task 4:

- Set out the questions needed to sort the samples into score levels.

Task 5:

- Provide a score level description based on the salient features used to divide the sample into all the clusters, as identified in task 3 and set out in task 4.

In their subsequent research (Upshur & Turner, 1999) identified three major concerns with these tasks. The first concern was that an analysis of salient features was useful for scale development, but that features that did not distinguish between different learner levels do not necessarily emerge. The second concern was that "when using empirical methods of scale construction the composition of construction teams and the make up of the samples of performances may have effects that deserve study" (Turner and Upshur 1999:107). Turner and Upshur addressed this issue themselves in their 2002 study for rating student writing, *not speaking*. Three teams of raters were provided with two samples of writing each from a group of learners from which to build empirically derived scales. The researchers observed that the "scale development team had a minor effect on ratings" (2002:65). As we shall see, the study design in this thesis allows careful consideration of this point. Turner & Upshur's final concern was whether these types of scale were task specific: they refer to the "tension between the need for accuracy in assessing a particular performance and the generalization to broader domains of language use" (Turner and Upshur 1999:107). The peer interaction construct, defined on the basis of an analysis of the discourse, was newly re-defined by the scale makers to include listening and nonverbal features (Chapter 4). It would not be useful if the rating scale operationalizing paired interaction only applied to the particular test performance. Performance on the task and the demonstrable skills that are rated based on the output need to be separable. The manner in which candidates interact with a peer is deemed transferable to other peer non-test situations because interaction is a demonstrable skill: interaction is not *the task* in itself it is a *result* of the task. In order to make the scale more robust the paired candidate video speaking samples were very carefully chosen to represent a range of performance types and candidate characteristics.

6.3.2 Adaptations of the EBB procedure

The EBB procedure, as presented above, had been used by the researchers that developed the procedure to develop data based scales for monologic spoken tasks (Turner and Upshur, 1996). The tasks used as input for those studies were of a different level of complexity when compared to the 10-minute Paired Test used in this study. For this reason, the EBB procedure was adapted in three ways:

1. **The individual familiarization stage:** This involved closely observing the 10 minute clips of peer discourse and producing Verbal Protocols.
2. **The provision of the reduced content analysis data:** The data from the protocols was transcribed and analysed. The rater comments were reduced by the researcher and presented in Tables on A4 sheets with a summary of the comments per pair of candidates made by three different raters.
3. **Consensus moderation of the EBB procedure:** There was a presentation by each team of their criterial question tree and a consensus as to which version to adopt for trial.

These three stages are each described in more detail below.

6.3.2.1 Adaptation 1: The individual familiarization stage

This adaptation was made in order to address the first concern raised by Upshur and Turner: features that did not distinguish levels not emerging as salient. In order to guard against this, in the study reported on here the 12 raters observed all the data *alone* before participating in scale development. The raters described all that they *attended* to that they considered to contribute to successful/unsuccessful peer interaction. In this adaptation of the EBB procedure scale makers work alone first and *rank* performances. In the original EBB the raters worked together at this stage. The focus here is ranking not describing what is noticeable about a performance.

The scale makers spent two hours on their own, focusing their ideas on interaction as they recorded the Verbal Protocols for Study 1 Part A, prior to coming together with their colleagues for the scale development. Raters were not guided as to what features they should consider important enough for them to make comments on.

The pre-scale development task also presented raters with the opportunity to familiarize themselves with a sample of the range of performances that would ultimately be used in the scale development. They had considered the issues at hand and the reasons for their ideas on interaction prior to the scale development workshop.

Even if a feature did not distinguish between levels, nonetheless it may have been identified by raters as observable in the discourse. In this way the concern raised by Upshur and Turner is addressed before the scale development.

6.3.2.2 Adaptation 2: The provision of the reduced content analysis data

The information gathered in Adaptation 1 was made available to all other scale makers as part of the scale development process. It consisted of an A4 Table of reduced data, one table per pair (described below). Data were drawn from the content analysis of Verbal Protocols of three different scale makers that had observed each candidate pair on video. By providing the scale makers with this information, the issue raised by Turner and Upshur regarding the effect of the scale makers on the scale was addressed. Their concern was that different scale makers would see different features as salient making different scales each time depending on the raters. By providing all scale makers with the views of three raters on each pair, each scale maker had input from others as well as their own views on what was salient about a performance.

As noted above, the content analysis data from the Verbal Protocol was reduced and summarised by the researcher. The intention was to make it manageable for scale makers to read and refer to during the scale making process. It was set out in columns per candidate pair. Separate laminated cards for each pair of candidates with one column for comments on the pair, and another for each of candidate *a* and candidate *b*. This way all comments for the each of the pairs were visible at a glance as can be seen in figure 13 below.

Figure 13: Data reduction cards for scale makers

• **Pair 1 comments:**	• **on pair**	• **left candidate**	• **right candidate**
• By Rater X	•	•	•
• By Rater Y	•	•	•
• By Rater Z	•	•	•

The aim was to give scale makers as much information as possible about each candidate before starting. Each pair of candidates had already been analysed and commented on by three different scale makers.

Rater 1 candidate 1, 2, 3
Rater 2 candidate 2, 3, 4
Rater 3 candidate 3, 4, 5 etc

Scale makers, a subset of the group of language teacher, were guided to keep the main themes from Study 1 Part A in mind as they ranked the pairs. These were:

1. To maintain text cohesion by asking relevant questions or making relevant contributions or responses to the topic
2. To respond in turns fluently and evenly without excessively holding the floor
3. To be an engaged listener by using back-channel
4. To be mutually supportive as a listener in the interaction e.g. To fill silences and gaps in language, by demonstrating comprehension
5. To use supportive gesture
6. To maintain eye contact.

6.3.2.3 Adaptation 3: Consensus moderation of the scales

The three teams developed separate but similar scales, which were presented to the group. After discussion there was consensus as to which scale to trial. Each scale is shown below in §6.4 with the differences between them.

6.3.3 Participants in rating scale workshop

The preliminary stage of data collection for Study 2 involved selecting candidate performance samples on which to base the scale. They were chosen by the researcher from a total of 17 pairs of candidates who had already taken part in Study 1 part B (Chapter 5).

In Study 1 Part A, where Verbal Protocols were elicited on successful interaction, some pairs attracted more comments. This was visible from the transcriptions of the Verbal Protocols elicited from the raters. It was assumed that the greater number of comments a pair had attracted the more salient their performance had been to the raters. This was considered when selecting eight pairs for the study: four pairs with more comments on particular *individual* candidates and four pairs with more comments on the *pair* were selected. At first glance it appeared that candidates commented on individually were not interacting as well as those pairs commented on as a pair. Also, of the eight pairs selected, four were evenly matched for linguistic proficiency and four were not evenly matched. (The matching was based on a departmental 5-point rating scale from the candidates' end of year oral performance marked by trained departmental raters, cf. Appendix 1).

In the study by Turner and Upshur (1996), in which they develop scales for monologic speaking, 12 individual performances were used. Taking into account that candidates perform together in this study, and that each sample comprises ten minutes of discourse, the number of performances was reduced to 8 pairs or 16 individuals.

The rating scale development took place at the end of semester when the teacher/rater/ language experts had marked 10 classes of paired orals, as is the norm at the end of semester. The researcher undertook the preliminary work of identifying sample pairs from the data by referring to the scores in Table 14.

Gram	3.5	2.5	3		4	4	4.5	3
Vocab.	3	3	3.5	2.5	4	4	4.5	3
Comm.	3.5	2.5	3.5	2	3.5	4	5	2.5
Comp.	3.5	2.5	3	3	3.5	4	4.5	2.5
Pair: cand	**3: 9**	**4 :2**	**5 :4**	**6 :16**	**8 :28**	**9: 15**	**11 :22**	**13 :32**
Gram.	2.5	3.5	3	2	4	3.5	4.	3
Vocab	2,5	4	3.5	2.5	4	4	4	3
Comm.	3	5	3.5	2	3.5	3	5	3
Comp.	3.5	4	3	3	3.5	4	4.5	3
Pair :cand	**3:12**	**4: 20**	**5: 34**	**6 :10**	**8 :27**	**9 :31**	**11:1**	**13:6**

Table 14: Range of candidate performance for scale development

The four criteria at the top of Table 14: comprehension, communication, vocabulary and grammar, are detailed in Appendix 1. The selection process for pairs to use as part of the input to the scale development involved many considerations. The

background of the students was taken into account, as well as how well matched they were to each other in terms of their level. In the pairs it is unusual to have 50% mismatched. In order to strengthen the tool and to challenge the raters to make the most difficult decisions they would face when rating pairs (e.g. when one candidate is ill prepared, extremely shy, or from another cultural background) half the pairs chosen were mismatched and half were evenly matched.

The next step was to decide how many raters to involve in the EBB procedure. Following Upshur and Turner, (1996:61) who recommend between "four to eight members who are familiar with the aims of assessment" in their EBB procedure for arriving at ratings, six raters participated in a workshop. The raters were experienced university Spanish language teachers, familiar with the task, the level and the rating context. They were drawn on a volunteer and availability basis from the larger pool of those who had performed the protocols on the videos of the paired candidates.

6.4 SCALE DEVELOPMENT WITH EBB PROCEDURE

The Upshur and Turner Method is a five-step process. A derived version was followed with the three adaptations. It is described in detail here to make the scale development process replicable. Scale makers followed a guide provided for them to ensure the scale development workshop followed the process step by step without the interference or influence of the researcher.

What follows is the workshop guide translated from the original in Spanish. It is the step-by-step process of developing a data based rating scale using student samples.

Workshop guide translated from appendix 6:

1. Pre-selection by researcher of 8 pairs from which to devise scale

2. A set of A4 cards with the rater protocol data reduced. These are to help you become familiar with the opinions of others on the performances of the candidate pairs involved. This is the order in which they are presented and it has no particular meaning.

3. Use preliminary findings from the rater protocols to help rank the performances
   To support through gaze and gesture
   To support mutually during the interaction
   To ask relevant questions within the topic
   To maintain cohesion from topic to topic
   To be a good listener
   To take balanced turns

4. Develop the scale. The steps for this are described below with the variations.

6.4.1 EBB step 1. A single question for the top of the hierarchy

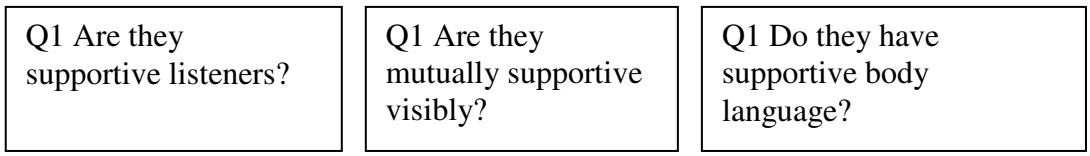
The aim of step one was to rank the performances and then to formulate a question. The ranking that takes place in the first part of the EBB is a simulated rating exercise which explains why the scale maker participants are referred to as 'raters' throughout the study even though they are not *applying* criteria but *developing* criterial questions. As raters they need to use their professional expertise to rank the performances and subsequently verbalise and formulate the criteria they are mentally using.

Criterial questions mark a boundary between levels. First rater teams watched and mentally ranked the performances without rating them but nevertheless deciding which ones were better than others. They did this by clicking on icons for the videos on a computer screen. Raters had access to multiple computers to watch and compare performances of particular pairs as needed. There was movement and discussion between the raters as they did this first task. First they all sat very quietly at their computers taking in the performances, but soon they were discussing what they thought were the best and worst performances on the screens.

Secondly, in their teams, raters discussed which particular feature of successful interaction they observed in the performances which would enable them as raters to split the sample of candidate pairs into + or – a particular feature. (The + indicated a YES response and the – indicated a NO response to a question formed by the raters.) The particular feature chosen was deemed to be the most salient attribute marking the boundary between two levels. The salient attribute was formed into a yes/no question. This question would be asked of every performance rated with the scale. The scale

makers wrote their question into a text box marked Q1. The first question proposed by the three different scale maker teams was:

Figure 15: question 1 for EBB



rater team 1 **rater team 2** **rater team 3**

As we can see in Figure 15, all three teams came to these three questions separately. Team one chose 'support as listeners during the interaction' as the most important and overriding feature that divides the set of performances in two. Team 2 focused on whether the candidates appeared to be supportive of each other in the interaction, through the word ‘mutually’. This support would be through the visible means offered by nonverbal communication. The last team also chose non-verbal communication as the first division of the candidates into two groups. Although they were working on different teams, they were all working with the same set of candidates so it is not surprising that teams 2 and 3 noticed very similar markers as the divide between the better performances and the less successful ones.

To finish step 1 of the scale development, the rater teams had to reach an agreement on which performances belonged to the group for which the response to question one was YES and likewise for the group for which the response to their first question was NO.

As a caveat, and before moving onto step 2 of the EBB process, it needs to be noted that this study did not capture the differences among the judges in terms of agreement on ranking. Therefore, although it would have been interesting to see if they all agreed on the top and the lower four, we do not have such data and cannot report on whether their rankings were the same. Because two groups selected non-verbal communication as the most remarkable difference, one could surmise that the same pairs of candidates were chosen for the successful and not so successful grouping with that feature in mind. The three teams were working independently and were not asked either where

the boundary for the upper and levels was drawn. It is not known if the same boundary was drawn for the 3 groups of judges.

With the focus on the distinction between the levels, the criterial questions were gathered at each stage by the researcher. The candidate pairs that went into each group, as selected by the different teams, were not recorded. The teams must have resolved any differences because they kept to the order set by the EBB and the adaptations set by the researcher. In order to produce one final scale based on the work of the three teams each of the scales were mapped on to each other and blended into one making the final product a consensus that was not based on a fixed choice of candidate pairs for successful communication.

6.4.2 EBB step 2. Questions for level 2 of the hierarchy

In step 2, the rater teams decided whether to work first on the upper or the lower ranked part of each sample, that is. the pairs grouped in the upper half with an answer YES or the pairs grouped in the lower half with an answer NO, to the question that divided the sample.

The scale makers ranked the performances in the section of the sample that they worked on. The scale makers wrote a question that divided the remaining performances in the sample then tested it against the candidates grouped at that level. The questions for level 2 of the hierarchy are shown in Figure 16, below:

Figure 16: question 2 for EBB

Q1 answer Yes: Q2 Supports interaction with the body? Q1 answer No: Q2 Asks questions relevant to topic?	Q1 answer Yes: Q2 Supportive listener with back-channel? Q1 answer No: Q2 Asks adequate questions?	Q1 answer Yes: Supportive listener? Q1 answer No: Q2 Asks relevant questions?
rater team 1	**rater team 2**	**rater team 3**

The questions that followed on from a YES answer on the first question from all three rater teams were either about listening or non-verbal support. The questions that followed on from the NO answer to the first question all contained 'question' as an indicator of what would move the interaction on from this point to the next level. This seems to indicate that non-verbal, or listening support, are of a higher order than asking questions, in scale makers' orientation to successful interaction. Asking and answering questions is necessary, but basic to the fine distinctions being made here by the raters. It is conceivable that one could ask and answer a string of questions without a very successful interaction. With success being categorised by listening and non-verbal communication, it is very interesting to see that the three teams working apart come to the same conclusion for the Q2 in response to NO in Q1.

6.4.3 EBB step 3. A cluster becomes a level.

The scale makers continued to rank and divide the pairs with questions that marked the boundary between levels. When the sample being considered can no longer be divided, the cluster, or group of candidate performances left without further division, becomes a level.

6.4.4 EBB step 4. Developing the EBB model

To conclude the session, each team of raters completed an overhead of their EBB model. This involved writing up the questions they had used to divide up the sample following some blank EBB model possibilities. Each team presented their model on an overhead and explained the reasons behind their question hierarchies to the group.

The three teams had many similarities in the ordering of their questions and in the design of their scales. The teams had either 5 or 7 levels. The two teams that had similar scales had 5 and 7 categories respectively. Of the three models, two models, those of rater team 2 and rater team 3, were very similar. Both of those teams had body language in Q1 of the hierarchy. Rater team 1's scale differed in that Q1 was on listening, The three different EBB scale models developed in the scale development session are presented in three figures below. In comparing the scales, it is noteworthy

that the question about cohesion distinguishes lower levels for team 2, but higher levels for team 3. This is because each question, used in the hierarchy, depends on which questions had been asked before for the previous YES or NO answer. The three scales are each shown in Appendix 10.

Teams 1 and 3 (in Appendix 10), with seven divisions, resulted in more severe scale models. That is, if the candidates were rated NO in Q1 for body language, the maximum they could then be rated was a three. Hence they would not successfully pass communicative paired interaction. It is only team 2's scale (Appendix 10), with five levels, permits candidates to reach three points after having been awarded NO for Q1. Just as with the first two teams, the questions in the Q3 group were all on interactional management.

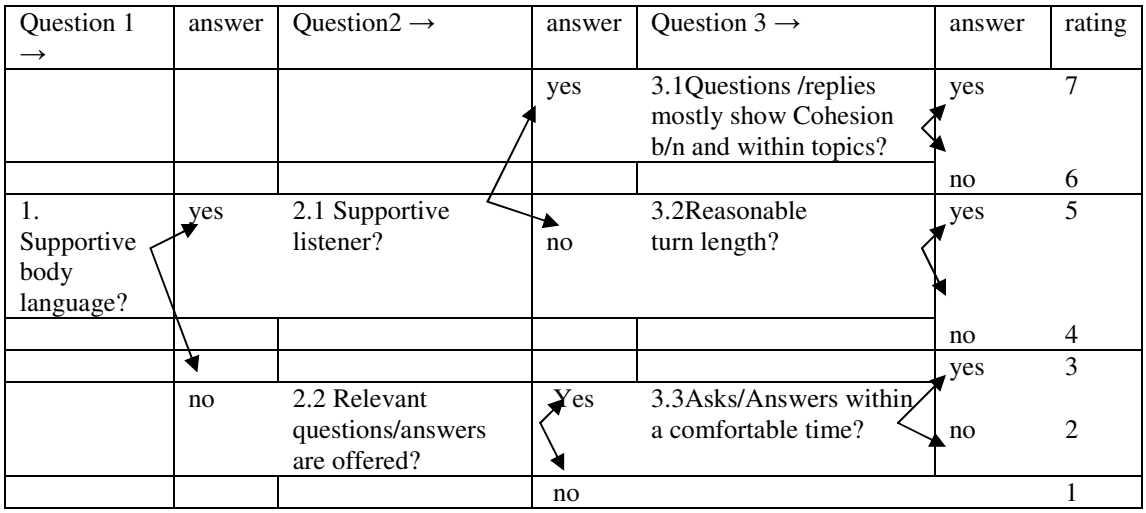
At the time of the study design and data collection the Upshur and Turner (2002) study on the effect of scale maker and student sample on the scale content was not available. In that study teams were used to create scales using the original methodology. Their findings were that the team has a minor effect on the scales produced but the student sample has a major effect. In the present study, the three teams were all looking at the same student samples. As in the Upshur and Turner (2002) study, the teams had a minor effect as seen by the three scales produced.

Fourth step: rater scale presentation for consensus on final rating procedure

It was also intended that the scale that was to be the outcome of the workshop would be more robust if it was developed based on three scales made from the data that could be combined by the consensus of the scale developer/scale makers.

By observing the similarities and differences over the three scales, the scale makers reached a consensus as to which scale to trial. The chosen scale was ‘tweaked’ before use by way of consensus moderation with the scale makers. The changes were based on evidence from the other two draft scales from the other two teams, not based on intuition. The final version is shown here:

Figure 17: consensus rating procedure to trial



Notably, the top row indicates the question order number to follow in the binary selection that channels the rater to a final rating from left to right. As seen in Figure 17, the EBB consensual scoring procedure starts with the first question *'Supportive body language?*' (YES or NO). The 'visibility' of non-verbal communication is high in the hierarchy for a successful interaction in this rating tool for interaction. It means that for someone to successfully interact they need to look at the interlocutor and signal that they are listening. This criterion, in this position, is the same as for teams 2 and 3 (Appendix 10 team 2 & team 3, respectively).

The focus for the second criterial question is on negotiating comprehension. The Q1 with answer YES, is followed by Q2.1 which asks *'Supportive listener?'*. This is once again the same as in the scales developed by rater teams 2 and 3. The Q1 with answer NO, followed by Q2.2 which asks 'Relevant Qs/answers offered?' is in the same position as in all three individual team scales.

In the final step, Q3 determines the final mark by distinguishing the level of interactional management displayed in the performance. The question asked as a result of a series of YES answers leads the rater to award a candidate the top mark. The question asks whether the candidate asks and replies to questions with cohesion within and between topics.

The highest level students are distinguished from the next highest by cohesion between and within topics. The lowest level students are marked by no body language or relevant questions or answers. Students at the intermediate levels are distinguished by reasonable turn length or speed of reply.

As the consensual version evolves it can be seen that the three teams chose the question version that had the most supporters (2 out of 3 or 3 out of 3) and included that question in the final version. There was not a great deal of discussion because to make the consensual scale the raters had moved to an intuitive plane and as raters were no longer deciding which question best separated a concrete group set within a level. The discussion moved to practical concerns such as one scale being too harsh. The result of employing that scale would result in too many fails. The other two scales out of the three was similar in their first question but one had less questions distinguishing levels. In this way other final choice out of the three was selected by the group of raters as best representing their perceptions of performance levels and practical needs. The exact wording of the questions was agreed on by the group in order for the scale to be trialled. During this final process the raters had to move to projecting onto the final scale what they thought would work from experience. But these decisions were made having developed the first three scales directly tied to the task and the performances.

6.4.5 EBB step 5. Writing a score level description

This last step involves writing a score level description to provide a picture of the trait being evaluated for score recipients such as other tutors, candidates, administrators or parents for example**.**

Due to time constraints the concrete description of each score level was compiled out of the session by the researcher. The end of the procedure consisted of writing a statement based on answers to three criterial questions to arrive at the level. These levels cannot be read in the manner that rating scales bands are normally read. They should only be seen one level at a time in isolation. Raters would only use the grid to

rate and students and other stakeholder only receive one level description as an outcome. The level descriptions, which are a list of the answers to the questions asked to arrive at the particular level, do NOT compare progressively higher and lower across levels as band scales usually do.

Level 7
Uses encouraging body language e.g. looks at speaker, smile, posture, hands head nodding
Is an audibly supportive listener e.g. really? m mm, yes yes, shows interest while the other speaks
The moves within the interaction and the responses show cohesion between and within topics

Level 6
Uses encouraging body language e.g. looks at speaker, smile, posture, hands head nodding
Is an audibly supportive listener e.g. really? m mm, yes yes, shows interest while the other speaks
The moves within the interaction and the responses do *not* always show cohesion between and within topics

Level 5
Uses encouraging body language e.g. looks at speaker, smile, posture, hands head nodding
Is an audibly supportive listener e.g. really? m mm, yes yes, shows interest while the other speaks
The turn length is balanced; it is neither too long nor too short

Level 4
Uses encouraging body language e.g. looks at speaker, smile, posture, hands head nodding
Is an audibly supportive listener e.g. really? m mm, yes yes, shows interest while the other speaks
The turn length is *not* balanced it is either too long or too short

Level 3
Body language is *not* supportive and tends towards visibly negative signals
Relevant questions and answers are given
Questions or answers are offered without too much hesitation

Level 2
Body language is *not* supportive it tends towards visibly negative signals
Relevant questions and answers are given
Questions or answers are *not* offered without a lot of hesitation

Level 1
Body language is *not* supportive it tends towards visibly negative signals
Relevant questions and answers are *not* given

This is not meant to be a descriptive scale in the normal sense, because the elements are not used to describe all the levels, but only to distinguish adjacent levels. The result is that when all the levels are placed together, features drop out or appear at different levels. Each criterial question depends on where one arrives in the procedure with the questions before. This is why the criteria are not used to build a descriptive scale in the usual sense. A particular level can be looked at in isolation, but the series of levels is not available to raters or to other stakeholders such as the candidates or the institutions.

6.5 DISCUSSION: RATING CO-CONSTRUCTED PERFORMANCE

The findings of the study are twofold. The first finding, in §6.5.1, is that the scale maker teams focused on three areas in descending order of importance regarding decision making for co-constructed performance. This is demonstrated by the hierarchy of Q1 through to Q3. The second finding, in §6.5.2, is that the EBB enables raters to neatly address the separability issue in rating co-constructed performance.

6.5.1 Decision making for co-constructed performances

The feature at Q1, (*Supportive body language?)* that is most salient for the first division of the candidates is outwardly visible signs of interaction, that is. interpersonal non-verbal communication. This is the first area of focus in rating paired interaction.

At the Q2 level there are two options for the scale makers (Q1 YES > Q2 *'Supportive listener?'* and Q1 NO > Q2 *'Relevant Qs/answers offered?'*). In other words, listening is the second focus in rating paired interaction, separating candidates who manifested signs of interactive listening from those who failed to do so.

The alternate path that follows Q2.2 does not involve listening directly, but asking and answering questions. The relevance of listening in asking questions is of more importance to the rater than in answering questions. Answering questions requires constructing a response that makes sense in the context by having listened.

Finally at the Q3 level, there are three pathways for the scale makers, determined by the answer to the preceding two questions. The element used to distinguish between levels is one of interactional management. At the highest level it is a question of cohesion. The middle level refers to turn length, while the lowest level focuses on fluency expressed as the time taken to respond.

In order to separate the last two levels, the scale makers' focus was on the very fine details of peer interaction. The salient details after non-verbal communication and interactive listening were features of the mutual support and signals of engagement

between speakers, which were demonstrated by observable interactional management skills.

We look first at the lower end of the hierarchy, which leads to awards of between 1 and 4. At the lower end (i.e. Q2-NO) there may be an audible breakdown e.g. hesitation, inefficient turn-taking, inappropriate response or initiation. If, despite the communication problem, questions and answers are provided, then YES achieves a 3 on interaction. If not, then a 2 is awarded. If for Q1 there is NO body language and there is NO relevant initiation or response - just random offerings for Q2, the rating is 1 for interaction.

An examination of the higher end of the hierarchy that leads to awards of 4, 5, 6 and 7, at the higher Q2-YES the interlocutor is now audibly as well as visibly supportive, providing back-channelling, initiating and responding appropriately with ease. If the answer is YES to Q2 the candidate engages and contributes to the development of the discourse, which moves to Q3 on cohesion, where the candidate is awarded 7 if cohesion between and within topics is sustained. However, if it is inconsistent, candidates are awarded 6.

If the answer is a NO to Q2 there may be evidence of some discourse management problems, insufficient initiating or over length responses. In this instance, candidates are awarded 5 for observing turn-taking conventions, but 4 if they are either silent for too long or conversely speak for too long.

The problems in rating a Paired Test are caused by the interaction of many different factors such as listening, speaker engagement and non-verbal communication, which have been captured and represented on this scale. To recapitulate, the raters first pay attention to the most obvious and salient feature that separates the group, then at each level the next most salient feature, then last of all they attend to the fine details. The ordered nature of salient features is from left to right, horizontally across the questions at each level, rather than an ordering represented by a vertical scale involving a set of levels

The most important findings of this evidence-based scale development are twofold: the particular features that were focused on and the order in which the criterial questions appear in the EBB.

Firstly, the elements that were found to make up the construct in Part 1 of the study mark separate levels on the scale. These are nonverbal communication, interactive listening and interactional management. This is a very interesting and significant finding, which needs highlighting. The question is whether the raters, when left to follow the rating scale workshop procedure, took very much notice of the preliminary findings presented as a list of six possible features of interaction that they as a group had found to be salient. Whether the raters stopped to consider these features from the earlier stage of the research study, and how much emphasis the rater teams put on them is unknown. The raters were not prompted or encouraged to use the list beyond a support for starting them off on their first ranking task. It seems likely that as they proceeded through the steps they followed each step without revisiting the first.

Secondly, what is most striking is the similar hierarchical order in which the elements in the criterial questions were set out. Two teams placed body language first and listening second and all teams selected interactional management as the third level of questioning. This is very strong evidence that the scales confirm the three criteria that later arose from the Verbal Protocols. To further tease out the levels, the listening construct needs to be explored (for mutually dependent interactive contexts). As noted earlier, while observing paired interaction the scale makers were aware of: the physical signals the partners emit; the listening and comprehension of the partners and the reliance on each others oral text cohesion and interactional management for the next thing they say. These all require further in-depth exploration.

These findings call into question the effectiveness of other rating criteria for ‘communicative interaction’ and ‘discourse management skills’ at least as far as they concern candidates taking tests that include a collaborative task and a discussion. Raters may observe or attend to non-verbal skills in peer interaction or displaying skills in interactive listening. Hypothetically, if raters' orientation remains below the

level of awareness, this could inadvertently affect rating. If scale makers for the Spanish beginner Paired Test noticed body language and the effectiveness of the listener, then possibly scale makers in other contexts may also attend to these factors. The study by Orr (2002) reports similar findings with regard to body language and eye contact.

The goal of empirically based scale development is to improve the quality of assessment by grounding it in student performance and the features that scale makers notice as being important to the performance. When the raters were deciding the criterial questions they were not prompted to think of the three specific categories that had emerged from the think-aloud section of the study. At the time of the scale development, there were only draft findings available and the intercoder check had not been carried out. The workshop outline offered six categories presented in alphabetical order to aid in the separation of the performances as successful or otherwise. As a result it can be claimed that the criterial questions *independently* confirm the salience of the three features oriented to Study 1, Part A and *independently* confirm the salience of the three features oriented to in the students' own reflections in Study 1 Part A.

6.5.2 The separability issue

The separability issue in co-constructed performance is not one that the raters had read about or had even considered. It would be contradictory to claim that interaction is co-constructed, and then proceed to give different marks for participation. The findings for pair dynamics as found by Storch (2001), Galazci(2004) and May (2006) demonstrate that there are different kinds of relationships in pairs whether in the SLA context or the testing context.

In order to improve the validity of rating for this test discourse, the staff adopted the EBB scale for interaction. It is currently being used in addition to already existing analytic scales for grammar, vocabulary and pronunciation (see Appendix 1). In the analytic scales candidates are awarded a separate score whereas on the newly included interaction scale they are awarded the same score.

The key point made above is that by including in a rating scale for communicative interaction, the features language experts orient to the features are thus scaleable. This means that what was previously 'intangible' (van Moere 2006) in interaction between peers has now been observed, described and placed on a functional scale. This rating method, because of its design, also allows the separability issue to be addressed in a concrete manner.

6.6 MAPPING STUDY 2 ON THE SCALE AND STUDY 1 ON THE ORIENTATION

In Chapter 5 the two sets of protocols were mapped to show how both performers and observers concentrated on three features: interpersonal non-verbal communication, interactive listening and interactional management. The candidates were not involved in the scale development but their participation in the study contributed to the validity of the scale.

The fact that three teams of raters separately developed similar EBB scales is evidence that the scale study places the conceptual categories in a fixed order of importance. The two categories of features that were new to rating scales and have not been empirically researched were the importance of the listener and the non-verbal communication. The conceptual categories for making judgments about face-to-face interaction were placed in a similar order by the three teams of raters.

Previous research using discourse analysis showed the importance of peer interaction for interactional management (Lazaraton 2002, Weir 2005). Raters included cohesion, turn length and fluency in the criterial scale questions to mark level boundaries. It is not surprising that raters placed such importance on the interactional management category (48.7% of raters' comments) in Study 1 part A. It was no less important for the candidates, for whom 28.3% of the comments were on interactional management.

6.7 CONCLUSION & CHAPTER SUMMARY

This study was motivated by a practical need to comprehensively rate peer interaction. The EBB scale developed by the scale makers has confirmed the salience of the features identified by raters through Verbal Protocols in Study 1 Part A of the study and in previous research on interactional management in peer tasks. The scale was achieved by focusing on the salient features of peer interaction, which included interpersonal non-verbal communication and interactive listening in addition to interactional management.

Recall **research question 3,** which asked 'Can candidate peer performance samples from a paired test form the basis for developing a rating procedure for interaction?'. It would be reasonable to claim that candidate peer performance samples can form the basis for the development of a judgment procedure for interaction. The scale development reported on in this chapter is based on a sample of paired candidate discourse, and has avoided the problems with criteria encountered in other scale development methods. In particular,

- The criterial questions are relevant to the task and content
- The criterial questions separate levels and group performances in clusters by moving from the large picture of interpersonal non-verbal communication to fine-grained interactional management
- The criterial questions do not include relative wording, such as ‘listens more’ or ‘listens less’ to differentiate between performances at level boundaries

The findings show how scale makers developed a scale to incorporate what is salient to them. As a result the process has responded to Pollit and Murray’s (1996) questions:

- Should comprehension be assessed as part of oral proficiency?

Yes, comprehension should be rated in a paired-peer task, because raters attend to candidates’ interactive listening skills.

- Should a proficiency battery test language production or language interaction or both?

In a peer –peer task both production and interaction can now be tested and rated analytically.

- Should the oral test be one of communicative success or linguistic ability?

Communicative success in this context equals yes to the three hierarchical criterial questions using the EBB method. Firstly it means that interpersonal nonverbal communication is the most noticeable indicator of successful communication to the raters in the Pair Task for beginner level Spanish. Second in importance is the need for candidates to be successful in negotiating comprehension as interactive listeners. Finally, and incorporating the third salient feature candidates need to maintain a high level of cohesion in their discourse displayed in the manner that they have tight control over interactional management. Based on these three points, which define peer task communicative success in this context, communicative success can now be analytically rated separately from linguistic ability through this method.

To conclude, this study has shown what is salient about peer interaction, and how it can be included on an EBB scale. It revealed the extent to which scale makers can determine what constitutes interaction, based on student performance of the Paired Test for Spanish beginners. The raters combined 'interactive speaking and listening' in the construct, which has significant implications for the validity of the oral assessment criteria currently being used in this context.

The scale, which includes speaking and listening, as well as non-verbal communication, has serious implications for current criteria for interactive communication. Interactional management, currently included in peer assessment criteria, is the tip of the iceberg. It has been shown how raters operationalize the paired speaking construct. Differences in severity, inconsistency and the use of non-criteria observed by Orr (2002) or the 'intangible' (Van Moere 2006) could be explained by the current use of 'conversation management' to mark peer assessment tasks. It appears that other elements are attracting raters' attention, with nowhere in the current scales to register what they observe.

The development of this rating procedure has implications for the notion of interactive listening during speaking, non-verbal interpersonal communication and demonstration of speaker engagement through interactional management. These have all come to the fore as part of the peer interaction construct operationalized in a scale by trained raters in this particular context.

Chapter 7: DISCUSSION

7.1 CHAPTER OVERVIEW & INTRODUCTION

7.1.1 Chapter overview

In this chapter the two studies and the three main research questions are reviewed and synthesized. In the first study it was demonstrated that candidate and rater views on peer interaction provide a mutually confirming representation of the interaction construct in the peer test format. The second study built on the theoretical basis of the first study to provide an empirically based scale. This chapter maps the data sets and discusses the implications of the findings in the studies.

7.1.2 Introduction

The main aim of this research, put forward in the introductory chapter, was twofold: to examine the manner in which the construct 'peer interaction' is operationalised from the perspective of both raters and candidates in a pair format task, and to use this as the basis for the development of a data-based rating scale for peer interaction.

The study emerged from a practical need to improve fairness in rating pairs by focusing on *what* raters and candidates consider successful interaction in a paired task. The study also rose from a gap identified in the discourse studies that had examined speaking scale development and validation using data from proficiency interviews and monologic and information gap tasks. Earlier discourse studies on pair tasks had demonstrated that paired peer test discourse displayed features that were different from those produced in other types of oral tasks. Prior to this thesis, there were no studies of rater or candidate orientation to the peer construct, or on the development of evidence-based scales to reflect peer interaction on an open task such as the one used in these studies.

7.2 REVIEW OF THE TWO STUDIES AND RESEARCH QUESTIONS

The two studies that were undertaken for the this thesis mutually confirm the salience of the features identified by raters and candidates through Verbal Protocols because three teams of unguided raters included them and ranked them equally in a hierarchical rating procedure.

The first study was reported in two separate parts. Chapter 4 explored the definition of peer interaction in a speaking test from the perspective of raters. The second part of the orientation to interaction study was reported in Chapter 5, which explored the definition of peer interaction in a speaking test from the perspective of candidates.

The second study reported on an adaptation of the Empirically-based, Binary-choice, Boundary-definition (EBB) rating procedure (Turner & Upshur, 1996). There were three stages to the adaptation: the individual familiarization stage, the provision of the reduced content analysis data and the consensus moderation.

7.2.1 Review of Study 1

7.2.1.1 Study 1 Part A: rater Verbal Protocol

In part one of the first study 12 raters observed videos of paired candidate performance and performed Verbal Protocols. The raters commented on the communicative interaction between the pairs assigned to them, from a total of 17 pairs of candidates. The pairs had self-selected into the study and had varying types of performance. For each pair, the verbal report was made by three different raters. Conversely, each rater made a report on three different pairs.

Each verbal report consisted of two parts. After watching the 10-minute performance the language specialists were asked to make a summary comment of their initial response to the success of interaction between the pair. The initial response was audio taped. The second part was a think aloud, also audio taped, where the raters made appraisals on the manner in which the peer interaction was unfolding while watching the videotape.

The content analysis of the Verbal Protocols provided insight into rater orientation to peer interaction. The raters noted non-verbal interpersonal communication, interactive listening and interactional management to be the salient features of peer interaction as they attended to the paired test performance.

7.2.1.2 Study 1 Part B: candidate Verbal Protocol

In part two of the first study, the participants were 25 individual candidates who had performed in the pairs that had been observed in part one. The candidates individually watched a video clip of their own paired test performance. The video clips were the same ones that the raters watched in Study 1. While individually watching their paired performance the candidates produced a retrospective stimulated verbal recall on how they managed their peer interaction. Their stimulated recall was audio taped.

In the analysis of the candidate protocols I looked for parallels between the features attended to by the raters in the first part of the study. The features of peer interaction that were also found to be prominent in candidates' orientation toward interaction with their peers in the Paired Test were: non-verbal interpersonal communication, interactive listening and interactional management. Hence candidates' awareness of the construct 'peer interaction' as part of a pair in a paired task reflected what raters attended to while observing the same performance.

7.2.2 Review of Study 2

7.2.2.1 Evidence-based scale development

In the second study the intention was to develop an evidence-based rating scale validated by the findings of the content analysis of the Verbal Protocols provided by candidates and raters. To achieve this, the scale needed to incorporate the three features salient to raters and candidates: non-verbal interpersonal communication, interactive listening and interactional management. The scale was developed using a data based method, using observations of live test candidate performance. The data

based scale used videos of candidates' test samples as data input. The scale aimed to distinguish the sets of pairs or individuals within pairs with regard to interaction.

A subset of six, from the 12 raters that participated in Study 1, part A participated in the scale development. The discourse sample used as input for this data based scale development was taken from a reduced set of pairs selected from the total of 17 pairs available. The 16 candidates selected, making up 8 pairs, displayed distinct types of paired performance - based on the comments provided by the raters in Study 1. In the Verbal Protocols, the raters had classified four of the pairs as mismatched in interaction, while the other four were perceived as having equivalent levels of interaction.

The second study built on the first in that the features of interaction that had been identified by candidates and raters guided the development of a data based scale for peer interaction. A summary of the comments made by each of three raters on the 16 candidates in Study 1 was collated to use as a prompt for the scale development in Study 2.

A scale development workshop was held. Following a step-by-step guide, three teams of raters were paired to develop a scale each, independently of the other scale developers. The Empirically-based, Binary-choice, Boundary-definition (EBB) scale (Turner & Upshur, 1996) was followed. Three very similar draft scales were developed separately by each team. A final scale based on all three scales was agreed on by consensus from the raters who had developed the three draft scales.

7.2.3 Research questions

The first study explored two related questions concerned with examining two perceptions of the reality of peer interaction: that of raters and of candidates. The research into the raters' perception was focused by the following question:

Research question 1: What features of peer interaction do raters attend to in paired task test performance?

This question was addressed by a content analysis of verbal reports on paired performance elicited from raters. On the basis of the content analysis of the verbal reports non-verbal communication, interactive listening and interactional management were all noted as salient features for the raters.

Research into the candidates' view of interaction in the reality of a paired oral was investigated with the question:

Research question 2: How do candidates view interaction in a paired oral?

Research question two was addressed by a content analysis of a transcribed retrospective stimulated verbal recall from the candidates. Based on the content analysis of the verbal recall, non-verbal interpersonal communication, interactive listening and interactional management were all noted as salient features for the candidates.

Study 2 was concerned with empirically developing a rating scale, focused by the question:

Research question 3: Can candidate peer performance samples from a paired test form the basis for developing a rating procedure for interaction?

Research question 3 was addressed by adapting and employing the Upshur and Turner (1996) EBB scale development procedure. This procedure involved dividing paired discourse samples by developing hierarchical criterial questions. Each question defined boundaries between levels of candidate performance. Initially, the discourse samples were used as input for three draft rating scales, which were developed by three separate teams of raters. The hierarchical criterial questions were found to vary only slightly between those three scales. The individual draft scales that had been produced in teams were then combined to enable a final scale to be developed through rater consensus.

Study 2 established that candidate discourse samples *can* be used as input for empirical rating scaled development. The scale that was devised combined the features non-verbal interpersonal communication, interactive listening and

interactional management into a single rating instrument, with an individual mark attributed to each candidate.

7.3 SUMMARY OF STUDY 1

7.3.1 Part A. Orientation to peer interaction: Rater Verbal Protocols

In this section the results for Study 1 are summarised. Firstly interpersonal non-verbal communication is presented; followed by, interactive listening ending with interactional management skills.

7.3.1.1 Interpersonal non-verbal communication

In this category, raters included two subcategories: gaze and all other body language including gesture. This category was visible even if the sound was turned off and the clip was watched in silence because evidence of non-verbal communication does not require sound.

The raters found body language or non-verbal interpersonal communication to be a contributing feature to the success, or lack thereof, of interpersonal interaction. In the content analysis 17.4% of comments were on this conceptual category. 'Gaze' has been commented on by raters in scale validation research (e.g. Orr, 2002), but it does not generally appear in the literature on paired test "performance". It is not yet widely studied in Second Language Acquisition either as is discussed in detail in Lazaraton (2004).

7.3.1.2 Interactive listening

Listening as part of successful interaction was divided into two subcategories: comprehension and 'supportive listening'. The first subcategory refers to a means of showing engagement, of giving encouragement for the speaker to continue or demonstrating comprehension. It can mean raters notice that candidates are filling a silence, asking for clarification or manifesting comprehension. In the content analysis

of the verbal reports the conceptual category 'interactive listening' accounted for 17.6% of the comments. This type required evidence of comprehension by the listener. The second kind, back-channel, did not require comprehension, just audible support with sounds. Both types of interactive listening support the interaction and were attended to by the raters.

In paired interaction, as in other dialogues, there are two roles to play: that of listener and of speaker. The findings of Galazci (2004) were confirmed here, as raters found two extremes: listening as a feature of interaction either failing to work, or at the other extreme, working successfully whereby candidates moved between the role of listener and speaker. This is where, in a pair, listening plays a crucial role in advancing speaking. If both candidates carry off both roles successfully then 'pair fluency' prevails.

Raters inferred from paired candidate performance that a beginner listener is able to keep supporting the speaker with 'listening noises' to go on until they do understand and then can demonstrate this with linguistic signs of comprehension.

7.3.1.3 Interactional management

After the calculation, 48.7% of the total comments made by the raters in the content analysis were on interactional management. Tests that have paired tasks such as the Cambridge suite already recognize the need for criteria that are worded to incorporate conversation or interactional management. The two subcategories turn taking and topic cohesion were not completely new in this context because they had been found in candidate discourse studies for scale validation, such as Galazci (2004)

In the rater Verbal Protocols there were few comments using the word 'fluency' between peers' interaction. Comments instead detailed the manner in which candidates demonstrated what 'pair' fluency was at the interactional management level. The finding that 'turn taking' and 'topic cohesion' were important is consistent with previous findings from analysis of candidate discourse such as Lazaraton (2002) for example. Turn taking and topic management could be seen as an elaboration of

fluency *between* two people, which is the essence of co-constructed dialogue (Jacoby & Ochs, 1995) compared to fluency in a monologue.

7.3.2 Part B. Orientation to peer interaction: Candidate stimulated verbal recall

In Part B of Study 1 the candidates were not guided by the researcher and were left to observe and comment on their behaviour with their partner. The findings show that candidates are able to talk about interaction when they focus on their own performance. They verbalized their experience of speaking to the other person in a test situation and commented on:

- Visual cues: through nonverbal communication
- Interactive cues: through interactive listening
- Contextual cues: through dependency in interactional management

Each of the three conceptual categories coded for in the candidate Verbal Protocols was made up of subcategories, which are listed below. These were not individually quantified but nevertheless combined to make up the category as a whole.

Visual cues: through interpersonal non-verbal communication

- Gaze and gesture
- Laughter
- Body position
- Facial expression

Interactive cues: through interactive listening

- Comprehension
- Not listening
- Listening to predict an interactive move

Contextual cues: through dependency in interactional management

- Topic change
- Turn organization
- Turn length: Flow/ Silence/ Speed

It could be inferred from the candidates' comments under the conceptual category of interactional management that they were expressing different types of dependency between themselves and the other candidate as they managed topics and turns in the

interaction in the paired task. These findings of dependency were interpreted to fall into three categories:

- Co-dependent
- Inter-reliant
- Inter-dependent

Co-dependent candidates focused on being 'an individual' who had to survive the paired interaction while at the same time considering, but competing against the partner. This is a negative situation.

Inter-reliant candidates focused on their partner's skill as an interlocutor. The inter-reliant individuals came to the paired interaction with the intention to attempt to interact. During the test the candidates were aware of the impact that their partner's shortcomings had on them.

Inter-dependent candidates demonstrate, through their comments, how one party intends to build on the other for success in interaction. In this case the individuals are totally subsumed into the pair. It is a positive situation. This inseparable co-construction of dialogue lies at the heart of communicative interaction. They work at the interaction together on the assumption that it will be a success for them though it may turn out to be an unsuccessful, as viewed by raters, though co-constructed interaction.

Candidates were not only conscious of the cues from the context that they need to change turns or change topic, but while they perform they are acutely aware of each other. Candidates were at different stages: the isolating 'what will I do next' such as in the negative co-dependent situation; or the stage of noticing what the partner is doing wrong in the inter-reliant situation. The best interactional management requires inter-dependent candidates willing to negotiate comprehension together for success.

The examination of pair-interaction during the orals revealed that candidates were conscious of the effects of interactional management on their performance. The word

fairness was not mentioned in a single protocol. Fairness is however, the overriding factor in success for the pair. To be successful they both should take part evenly by contributing equally.

It was argued in this study that candidates are aware of the process of interaction they participate in. Their ability to verbalize the importance of each candidate's participation in the performance has implications for the definition of the construct being tested in paired interaction. Candidates focus on negotiating comprehension and on managing interaction when talking. The comments showed that this allows them to further the dialogue they are co-constructing. Candidates attended to similar conceptual categories as the raters, but for them the categories are cues for moving the interaction forward. The categories are:

- Visual cues: through body language, which relate to the nonverbal category
- Interactive cues: through the listening interactive listening category
- Contextual cues: through topic and content interactional management

The candidates' perspective on what they attend to in performance has been previously unknown to testers. They commented on how closely they focused on their partner and how their partner's performance impacted on them as they co-constructed the performance. Despite the co-construction they are separate entities. While recognizing that it is a 'display' task in a test, the Paired Test nevertheless allows an authentic representation of conversation management. Candidates are acutely aware of this, with regard to cues for topic and speaker change.

Galazci (2004) researched levels of interaction in paired orals and suggested including candidates (as well as raters) as experts to contribute to the test validation process. This study reveals the extent to which candidates can determine what interaction 'is' for them based on their own performance of paired Spanish proficiency test.

7.4 SUMMARY OF STUDY 2

In this section the results for the development of the rating scale are summarized. A practical application, the rating scale makes concrete the manner in which raters and candidates operationalized the construct 'peer interaction' in a paired task. The scale was modelled on both the candidates' paired test performances and the two sets of verbal reports elicited from candidates and raters in Study 1. The verbal reports are put forward as evidence of the validity of the scale. The reports detail how the salient features of peer interaction were identified

It is not known whether the criteria in use for rating peer interaction on large-scale tests reflect what is salient to raters while they are observing candidate performance. It is also not known whether candidates performing the 'interaction' that is expected of them in the task are aware of features of paired performance during performance that may be salient to raters. Currently, validated rating criteria reflect findings from research into peer discourse - not into feature saliency to raters.

The features of the peer construct focused on by candidates and raters arrived at in Study 1 were shared with raters, and used as input into the scale development process in Study 2. Following this, each of the three teams of raters developing scales were informed of the views of the six language specialists on the construct.

The Upshur and Turner (1996) EBB scale had previously been used for monologic story retelling and closed information gap speaking tasks for English as a Second Language (as well as for writing tasks). The scale levels reported on here are for rating a dialogic peer spoken interaction open task for L2 Spanish beginners.

7.4.1 Scale levels

The five levels for 'interaction' were defined in terms of three conceptual categories: non-verbal interpersonal communication, interactive listening and interactional management. The ordering of the levels is surprising in many respects because criteria seem to suddenly cut out. This is a function of the method used. It is imperative to point out that the raters do not see level descriptors in the way set out below. Such

paragraphs would be seen in isolation as a reporting device, without the other levels being present. This is because the EBB - the criterial method, with its tree shaped diagram used to arrive at a score - not a set of strung together level descriptors as appear below:

At Level 5, the candidate uses encouraging body language eg looks at the other speaker, may smile, has a posture that shows engagement, uses hands and head expressively. The candidate is an audibly supportive listener e.g. *really? m mm, yes yes*, and shows interest while the other speaks. The moves within the interaction and the responses show cohesion between and within topics.

At Level 4.5, the candidate uses encouraging body language eg looks at the other speaker, may smile, has a posture that shows engagement, uses hands and head expressively. The candidate is an audibly supportive listener e.g. *really? m mm, yes yes*, and shows interest while the other speaks. The moves within the interaction and the responses *do not always* show cohesion between and within topics.

At Level 4, the candidate uses encouraging body language eg looks at the other speaker, may smile, has a posture that shows engagement, uses hands and head expressively. The candidate is an audibly supportive listener e.g. *really? m mm, yes yes*, and shows interest while the other speaks. The turn length is balanced; it is neither too long nor too short

At Level 3.5, the candidate uses encouraging body language eg looks at the other speaker, may smile, has a posture that shows engagement, uses hands and head expressively. The candidate is an audibly supportive listener e.g. *really? m mm, yes yes,* and shows interest while the other speaks. The turn length is not balanced it is either too long or too short

At Level 3, body language is not supportive and tends towards visibly negative signals. The candidate is not an 'audibly' supportive listener so does not use back-

channelling. The candidate asks relevant questions and appropriate answers are given. Questions or answers are offered without too much hesitation.

At Level 2, body language is not supportive and tends towards visibly negative signals. The candidate is not an 'audibly' supportive listener so does not use back-channelling. The candidate asks relevant questions and appropriate answers are given. Questions or answers are *not* offered without a lot of hesitation.

At Level 1, the candidate's body language is not supportive and tends towards visibly negative signals. The candidate does not ask relevant questions and answers are not given.

7.4.2 Discussion and interpretation of Study 2

The key point is that observable features of interaction were scaleable using the EBB tool. The features 12 raters oriented to, while observing peer interaction in Study 1 A, have been included by a subset of 6 raters on the evidence-based scale reported here. This could mean that elements that previously were "intangible' in interaction between peers (Van Moere, 2006), and that were a possible source of variability in rating candidates "relaxing" in pairs (Foot, 1999) may now have been observed, described and placed on an evidence-based functional EBB scale.

This has been achieved because raters working as test developers using an adapted EBB methodology focused on what is most salient in peer interaction. The scale development procedure also asked the scale developers to create a hierarchical order in which criteria are included on the EBB scale. This was done by asking the scale developers to define the boundaries between levels with criterial yes/no questions. The primary task of the scale developers was not to describe levels of interaction *per se*. The reported level descriptions were arrived at as a *consequence* of the scale making process.

7.5 INTERPRETATION AND DISCUSSION OF FINDINGS

The two studies and their findings were reviewed and discussed above. What follows is a discussion in two parts.

7.5.1 Mapping of the protocol data sets

The data from the two verbal protocol studies show that raters notice three salient features of interaction as they observe paired performance, and candidates retrospectively recall being aware of the same set of features. The candidate use of these categories was entirely independent as they were not prompted to think in terms of these specific categories. While coding, the researcher focused on these categories to test whether they overlapped with the same orientations that had been identified for the rater protocols.

The conceptual categories were: interactive listening, interpersonal non-verbal communication and interactional management. It would have been of concern if raters noticed a set of features that candidates did not consider to be important. Conversely, if candidates put concerted effort into producing a set of features that were not salient to raters this would be equally problematic. It can be inferred from this match in saliency that in terms of interaction, the attributes that candidates are aware of contribute to raters' evaluation of their performance. These features found *a priori*, not in a *post hoc* validation exercise, are strong empirical foundations upon which to build a data based scale.

7.5.2 Mapping the scale study and orientation

The dovetailing, described above, of the perceptions of the construct 'peer interaction' from both the perspective of candidates and raters offers two sources of validity evidence for the rating procedure. An additional source of validity comes from the scale being developed from the same candidate discourse sample independently by three teams of raters. These three teams of raters developed separate, but similar EBB scales. The set of criterial questions for making judgments about face-to-face interaction were placed in the same hierarchical order by the three teams of raters.

The teams focused on three areas: visual (as in the body language); interactive (as exemplified by listening); and contextual (exemplified by the fine detail of text cohesion in interactional management). The salient features were subsequently focused on in the manner of a camera zoom. Each question asked with the intention of determining level boundaries was increasingly more fine-grained. This is demonstrated by the hierarchy of Q1 through to Q3 in the YES/NO binary question tree that led to the final rating for interaction, as set out in the previous chapter. The feature that first divided the candidates' paired performance was the outwardly visible sign of interaction: interpersonal non-verbal communication. At the Q2 level, signs of interactive listening focused the decision. This deeper level of *interactional engagement* was judged second. Finally, at the Q3 level, the salient details were features of mutual support and signs of engagement between speakers demonstrated by *interactional management*. The visual and the audible characteristics of the performance came first in the criterial questions, followed by the cognitive load of oral 'text handling' which appears last and most finely details a candidate's level of interaction.

7.6 IMPLICATIONS

The findings of the studies reported on in this thesis are a challenge for current tests that include peer tasks. The fact that raters find interactive listening and interpersonal non-verbal communication both to be salient features of interaction have not been recognised in the testing literature. Lazaraton (2002) suggests that conventional SLA research needs to take non-verbal communication into account and by implication this should also apply to second language testing. The fact that interactional management is also salient supports the claims already made by post hoc test validation research (Taylor, 2001; Lazaraton, 2002; Galazci, 2003; ffrench 2003) regarding the ability of candidates to display different types of interactional management when paired with peers compared with when paired with an interviewer/examiner.

Taking into account what we normally rate as peer-peer oral proficiency in testing, we could consider inclusion of listening and body language in criteria. We should look at this different kind of listening because mutual orientation between speaker and hearer is basic to interaction and is implicated in all types of social interaction beyond

testing. Even in a speaking test, listening binds the mutual orientation between the pair.

Teachers may argue that *interaction* is not taught, and therefore why should it be 'tested'? This is a valid question that expresses a desire for fairness and natural justice for the candidates. However, it seems that raters are oriented to this feature, whether they are aware of it or not. Recall that research has shown raters may rate beyond the criteria (Brown 2000). Having somewhere 'to park' what they notice during interaction may help them focus better on the other four or five remaining criteria. The interaction construct could be precisely defined for a task and a testing context. Therefore raters, instead of being distracted by body language, faulty topic changing or a lack of mutual support demonstrated by poor listening, can address interaction directly and consciously. In this manner raters can set down what they notice and continue marking other interactional or linguistic skills.

By broadening and defining the construct, we accept that interpersonal communication is affected by elements beyond what is taught in a textbook or by the majority of classroom teachers. The elements identified and presented above deserve their place alongside other traditional criteria, in order to aid the rating process to become more transparent and representative of a paired test task reality.

As put forward in the literature review, there have not been many studies into the cognitive processes employed in rating oral proficiency. So far it has been limited to rater orientations (Pollitt and Murray, 1996; Brown, Iwashita, McNamara 2005) and scale validation (Meiron, 1998, Brown 2000). The fact that raters have been able to self generate features of peer interaction relevant to rating suggests this study parallels previous findings by Meiron (1998) who reported that raters self generate criteria not in the scoring rubrics and by Brown (2000) who reported that rater focus is broad ranging and particularly on communicative skill, lies beyond the scales.

The main implication of this study is that interactive communication should be put into our rating scales. This would enable raters to focus their attention on rating topic

management and turn taking along with accuracy and range of vocabulary, grammar, pronunciation etc

7.7 CONCLUSION & CHAPTER SUMMARY

7.7.1 Conclusions

This chapter has highlighted how unless there is a clear and common understanding of the construct, the rating system cannot work as it is intended to and tests are not fair.

The studies have added a clear and common understanding of the construct by complementing what was previously known about peer performance. Study 1 Part A added the perspective of raters by asking the raters what aspects are salient when they operationalize the construct. Study 1 Part B showed that candidates are aware of the features that are salient to raters during their paired performance.

With regard to the rating system working as it is intended, the issue is plainly addressed by (North 2003:40) with his reminder that "what should be assessed is not the performance itself but beyond that what one can deduce from the performance of the learner's ability in relation to the constructs of communicative competence underlying the performance." The study examined performance-based contextualised scales and in particular the performance of peer interaction. An agreement on the construct is necessary for the rating system to work as is intended but also for test to be fair for all stakeholders. In order to improve the validity of rating for Spanish language in this test, where the pressures of accountability are great, the staff adopted and supported the implementation of new criteria by way of a procedure for assessing interaction in addition to grammar, vocabulary, pronunciation and fluency.

This scale and the rater and candidate orientations particularly to the listener in the pair have implications for tests that employ the paired test as part of a battery for oral proficiency beyond this context. These qualitative findings are transferable and call for serious consideration of this expanded construct.

7.7.2 Chapter summary

The two principal findings of the studies that make up this thesis are that:

- Interactive listening, interpersonal non-verbal communication and interactional management are perceived as salient features of successful performance in peer test discourse samples by both candidates and raters
- These same features, observed in samples of paired candidate performance, *can be used as the basis for a rating procedure* that consists of hierarchical criterial questions each defining differences between levels of interaction in peer test discourse, as follows:
- Visual cues: through non-verbal fluency
  Interactive cues: through interactive listening
  Contextual cues: through dependency in interactional management

Chapter 8: CONCLUSION

8.1 INTRODUCTION

This chapter concludes the thesis. There is a review of the differences between this study and others, and then the limitations of the studies and the methodology are presented, followed by practical implications, which result from the findings. Finally, some suggestions are offered for further research.

8.2 SUMMARY OF TWO STUDIES AND FINDINGS

This thesis explored how 'interaction' in a paired live test is described by raters and candidates observing videoed performances. It was also concerned with whether interaction was assessable and if so whether an empirical data based scale development method would be suitable for incorporating levels of 'interaction'. As a result of the research carried out for this thesis our understanding of the phenomenon of paired interaction has been broadened and an empirical assessment procedure has been developed.

The thesis consisted of two studies. One put forward what 'interaction' in a paired oral means from two perspectives: that of raters and candidates. This first study investigated interaction in the paired oral in detail, through Verbal Protocols, to enable better understanding of the nature of the interaction between adult beginner Spanish language students in this type of test. The other study described the procedure for Upshur and Turner's (1996) Empirically-based, Binary-definition, Boundary-bound (EBB) assessment procedure methodology. The EBB data based procedure was adapted for the site and test context. The resulting scale focused on interaction and was developed based on sample of video clips of test-discourse drawn from oral tests of the candidate pairs.

Study 1 was carried out in two parts with research questions that focused on describing the construct of paired oral interaction as viewed by candidates and by raters. Verbal reports were used to examine rater and candidate orientation to paired interaction. The use of verbal reports in this study was motivated by an intention to

avoid the shortcomings of previous studies on peer test tasks. Previous studies had focused on discourse analysis of candidates' test performance. Those studies had not sought clarification whether the features that were noticeable, and made evident through discourse analysis, were equally salient to raters or candidates while they were observing the performance.

A set of salient features of interaction was found to be in common between the orientation of the candidates and the raters: these were interactive listening, interpersonal non-verbal communication and interactional management. Of these, two had been identified in previous literature. Earlier studies had noted that raters had focused on interpersonal non-verbal communication by going beyond available rating criteria while marking orals (Orr 2002). Interactional management had also been found to be a function made evident in peer dialogue in a paired task (Taylor 2001, Lazaraton 2002, ffrench 2003)

Interactive listening had not previously been identified in the literature as an element of speaking. However, the study in this thesis identified and illustrated the importance of interactive listening on the part of the listener in the peer pair. Interactive listening was salient to candidates and raters so it can be claimed to form part of the construct of peer interaction. The findings have provided us with a greater understanding of the interactive nature of peer performance and the integral role of each of the participant as both speaker *and listener* during the paired task.

In Study 2 an empirical set of criterial questions to rate interaction was developed and validated. It was developed using the EBB empirical scale development methodology, (Upshur and Turner, 1995). The methodology involved applying existing empirical scale development techniques to a new area in order to create rating criteria on a new type of scale for interaction. Study 2 demonstrated the usefulness of engaging raters to focus on the *difference* between levels in order to separate out salient features within rating bands. The scale study, while supporting earlier studies for scale development in both writing and speaking with this methodology, shows, in addition, how adapting the development procedure can aid in overcoming shortcomings which the scale developers themselves had pointed out.

The findings of the study challenge the current conceptualization of speaking within paired tasks in language tests to be constructed *without* 'listening' or 'non-verbal communication'. They suggest we need to broaden the construct of peer interaction in test development and implementation. The findings point to methodological and practical implications, set out below, for more transparent data based rating scale development.

8.3 DIFFERENCES BETWEEN THIS STUDY AND PRECURSORS

There are four differences between this study and its precursors, which mark its importance.

1) It is a *beginner* level class room assessment; in contrast much of the research in paired or group orals up till recently has been at the intermediate level or higher.

2) Candidates perform a paired *Spanish* L2 dialogue compared with most of the research which is on candidates speaking English.

3) Candidates' awareness during speaking test performance and raters' orientation to what was salient to them in the construct being tested, in paired tasks, *had not yet been empirically elicited* prior to this study. Candidates' post-hoc verbal reports on videos of their own speaking test clarify and validate what was actually occurring between the pair. The same observed discourse is later used by raters to describe salient components of the paired interaction construct.

4) A construct should be described to enable it to be the target of testing and to allow it to be measured after a process of scale development. Before this research speaking scales for peer interaction as a construct had been developed and but not *validated empirically from observed performances.* They had been validated from conversation analysis of transcriptions where the *visible* dimension of listener behaviour and the non-verbal interpersonal communication of both participants were not available to the researchers.

8.4 METHODOLOGICAL IMPLICATIONS

Claims based on the findings in this thesis resulted from three sets of evidence: two Verbal Protocols and an empirical scale to triangulate the findings and lay the foundation for claims. These findings and the three sets of evidence to make claims had important methodological implications.

Content analysis of Verbal Protocols provided two examples of qualitative methodology in which two sets of participants, candidates and raters, observed the same set of peer performances and came to similar conclusions. The conclusions were that successful peer interaction requires interpersonal non-verbal communication skills, active partner listening skills and interactional management skills. The results of the independent content analysis of both sets demonstrated that candidates and raters have parallel focus on interaction in speaking performance. The conclusions drawn from the orientation study and those from the scale development study appeared to support each other. From a methodological perspective two sets of protocols have been used to make the same parallel claims. These claims validate rating criteria in a scale for a similar set of performance features of interaction.

Any protocol study will raise issues of methodology, and certain issues arose in this case. In Study 1 when the language experts and the candidates were focusing on candidate behaviour on the 10 minute test discourse clips during the verbal reporting, the question *directly* asked them to focus on the success or otherwise of the paired interaction. Candidates or language experts so directed may have modified their behaviour. To minimize this possibility, future studies could possibly ask for a general comment on 'success' without pointing out the focus on interaction; however this would also pave the way for greater irrelevant and unfocused comment in the verbal reporting, and orientation to interaction may have remained implicit to a certain extent.

It is clear from the study that it would have been insufficient to make an audio recording when working or researching with paired interaction. Much of the

information that was commented on would not have been possible without the visual channel. This has practical implications for the delivery of large scale international tests where a cassette recording, which lacks the visual component, is used for double marking.

8.4.1 Implications for the field

Finally, the thesis redefines the construct tested in paired interaction within the context of a Paired Test to include interactive listening, interpersonal non-verbal communication and interactional management.

It has also shown that candidates are capable of making appropriate and meaningful comments on their own performance in order to contribute to interpretation of test discourse and the development of rating procedures. With such specific feedback from candidates and raters on interaction, the implications for notions of the speaking construct in mutually dependent interactive contexts need to be further explored.

The degree to which successful paired interaction can explain a component of a language test score for a peer task seems reasonable since, for instance: (1) raters orient towards features of interaction when rating and (2) candidates are aware of the fact that they monitor, plan and evaluate their interaction with the other candidate in a pair during a test.

8.5 LIMITATIONS OF THE STUDY

Although interactive listening, interpersonal non-verbal communication and interactional management are profoundly implicated in the development of a procedure for rating for interaction the following limitation needs consideration when considering the findings of this study.

The present study has not examined the nature of peer interaction in other types of paired tasks such as role-plays or group discussion. In the PT the candidates are being themselves ‘students in a test’ displaying to the rater what they can talk about together

within the time constraints of the test and on the topics for the task. It makes it more artificial as a task when there are no set roles to take. The implications are that the type of discourse elicited is not transferable to a real life situation out of the testing procedure. The inferences made from such a performance would not be as generalizable to situations beyond the test compared to a role-play. It would be useful to understand in which task types test-takers are more successful in interaction, and how interaction varies among task types. Further analyses at these levels are needed.

Interaction is a necessary but not a sufficient condition for successful L2 test performance. The construct definition and the scale development were concerned with interaction but another possible limitation to consider is that if the linguistics levels had been the same with only the interaction varying perhaps the outcome would have been more replicable or robust because the language experts and the scale developers' focus would not have had other linguistic distracters. Hence, an issue to consider for future research is to develop a scale from a student sample of pairs where the only difference between the individuals in the pairs is their performance of degrees of successful interaction. The choice of candidates for the data set involved many different linguistic levels although the topic sets were the same. Closer linguistic levels would have been impossible to control for in this study (they were already limited to beginner level) because the pairs volunteered and self-selected their partners. Had it been possible it would have eliminated the manner in which salient features of linguistic performance impact on degrees of interaction.

8.5.1 Limitations of participants in the sample

In Study 2 the raters who worked on the scale development and the sample of candidate test discourse that was selected were intended to represent the target population. However by setting up the study data set using only pairs that came back for feedback perhaps the findings were compromised because there might not have been enough representation of the average student. Instead, those who came for feedback and to participate in the study may have been the very conscientious students or the ones that were struggling.

Then there is the question of sample size. The fact that it would have been very time consuming to include more pairs may have been a limitation to the scale. The single candidates that made up the pairs in the sample may have been limited only to candidates who were of a particular type such as high achieving wanting improvement or anxious low achievers who came back to view their performance and receive feedback.

To control for that and to make the rating procedure more robust the paired candidates were very carefully chosen to represent an equal spread of the best and the worst possible scenarios in peer interaction from within the sample that volunteered to take part in the study. So this possible limitation was dealt with in the best way possible.

In the case of the raters, perhaps only those who felt confident in their rating skills were those who returned to develop the scales. By replicating the study with: another 12 language experts, not from beginner level, providing construct input; a different set of 6 raters developing the rating procedure; and taking a random language sample, not just the sample of those interested/anxious participants who made up the data set only then could the findings indicate whether they can be applied to other similar test contexts.

Nevertheless, taking into account these limitations, the study has shown that the change of position from listener to speaker during peer interaction is intrinsically connected to the production and perception of cues based on interactive listening, interpersonal non-verbal communication and interactional management. These features need to be taken into account in rating paired tasks in fairness to all candidates because these features are salient to raters and candidates are aware of them also. The suggestion is that a broader representation of the peer interaction construct be adopted in the testing context to greater reflect its complexity.

8.6 FURTHER RESEARCH

One consideration is whether the features salient to *these* scale makers incorporated in *this* rating scale for paired interaction have resulted in a practical tool for rating peer interaction beyond this particular language, context and task. Reports from scale trials

in other Spanish and other language beginner programs or will shed light on this point.

Evidence-based scale development is a time consuming exercise. A quantitative comparison by examining two scales for discrimination between raters and candidates would be useful. In the comparison, a departmental intuitively developed scale could be measured against an evidence-based scale for its discriminatory powers. Only then could claims be made for using the evidence-based scale development style over the intuitive scale development procedure.

Finally, the findings from the two studies in this thesis (the rater orientation study and the candidate awareness study) and the scale development all have implications for our understanding of the construct of effective interaction in paired candidate speaking tests, and for the development of appropriate rating scales in the future.

LIST OF REFERENCES

Afflerbach, P., & Johnson, P. (1984). Research methodology on the use of verbal reports in reading research. *Journal of Reading behaviour, 16*, 307 - 321.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149-164.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Berry, V. (1995). *A qualitative analysis of factors affecting learner performances in group oral tests.* Paper presented at the 17th language testing research colloquium, Longbeach,CA.

Berry, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In A. K. Huhta, V. Kurki-Suonio and Luoma, S. (Ed.), *Current developments in language testing*. Jyvaskyla: Jyvaskyla University Press.

Berry, V. (1998). *Personality and oral test score variability.* Paper presented at the TESOL conference, Seattle ,WA.

Berwick, R., & Ross, S. (1996). Cross-cultural pragmatics in oral proficiency interview strategies. In M. Milanovic & M. Saville (Eds.), *Performance testing, cognition and assessment: selected paper from the 15th Language Testing Research Colloquium*. Cambridge: Cambridge University Press.

Bonk, W. J., & Ockey, G. J. (2003). A many facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Brindley, G. (1998). Describing Language Development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge: Cambridge University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt: Peter Lang

Brown, A. (Ed.). (2000). *An investigation of the rating process in the IELTS speaking module* (Vol. 3). Sydney: ELICOS.

Brown, A., & Hill, K. (Eds.). (1998). *Interviewer style and candidate performance in the IELTS oral interview* (Vol. 1). Sydney: ELICOS

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English-for-academic purposes speaking tasks* (No. MS 29): ETS.TOEFL.

Brown, A., & Lumley, T. (1997). Interviewer variability in specific-purpose language performance tests. In V. Kohonen, A. Huhta, L. Kurki-Suonio & S. Luoma (Eds.), *Current Developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 137-150): Jyvaskyla: University of Jyvaskyla and the University of Tampere.

Brown, A., & McNamara, T. F. (2004). The devil is in the detail. *TESOL Quarterly, 38*(3), 524-538.

Buck, G. (1991). The testing of listening comprehension. *Language Testing, 8*(1), 67-91.

Cafarella, C. (1994). Assessor accommodation in the VCE Italian oral test. *Australian Review of Applied Linguistics, 20*, 21-41.
Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.), (pp. 333-342). Rowley :MA: Newbury House.
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.
Celce-Murcia, M., Dornyei, Z., & Durrell, S. (1995). Communicative competence: a pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5-35.
Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-33.
Clark, J. L. D. (1978). *Direct testing of Speaking Proficiency: theory and application.* Princeton, N.J.: Educational Teaching Services.
Creswell, J.W. & Plano Clark, V.L. (2007) *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage
Csepes, I. (2002). *Measuring oral proficiency through paired performance.* Unpublished Ph D dissertation, Budapest.
Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.
Douglas, D. (1994). Quantity and quality in speaking assessment performance. *Language Testing, 11*, 125-144.
Egyud, G., & Glover, P. (2001). Oral testing in pairs: A secondary school perspective. *ELT Journal, 55*(1), 70-76.
Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal reports as data (Rev. Ed.)*. Cambridge, MA: MIT Press.
ffrench, A. (2003). The development of a set of assessment criteria for Speaking Tests. *UCLES research notes, 13*, 8-16.
Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal, 30*, 156-167.
Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 36-41.
Fulcher, G. (1987). Tests of oral performance: the need for data based criteria. *ELT Journal, 41*(4), 287-291.
Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13*, 123-151.
Fulcher, G. (2000). The communicative legacy of language testing. *System, 28*(4), 479-482.
Fulcher, G. (2003). *Testing second language speaking*. London New York: Longman.
Fulcher, G. (Ed.). (1997). *The testing of speaking in a second language.* Amsterdam: Kluwer academic publishers.
Galaczi, E. (2004). *Peer-peer Interaction in a paired speaking test: the case of the First Certificate in English.* Unpublished PhD dissertation, Teachers College, Columbia University, New York.
Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.
Gass, S., & Varonis, E. (1985). Task Variation and Non-native/Non-native Negotiation of meaning. In S. Gass & C. Madden (Eds.), *Input in Second Language Acquisition* (pp. 149-161). Rowley, MA: Newbury House.
Gibbons, P. (2002). *Scaffolding language, scaffolding learning*. Portsmouth NH: Heinemann.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. New York: Aldine.
Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.
Gruba, P. (1999). *The role of digital video media in second language listening comprehension.* Unpublished University of Melbourne, Melbourne.
Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex Publishing Corporation.
Hatch. (1992).
Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Newbury House Publishers.
Hatch, P. (1992). *Discourse and language education.* Cambridge: Cambridge University Press.
Hildson, J. (1991). The group oral exam: Advantages and limitations. In J. C. Alderson & B. North (Eds.), *Language testing in the 90's: the communicative legacy*. London: Modern English Publications and the British Council.
Humphry-Baker, A. (2000). Speaking tests: students' Perception and Performance, (Vol.). Manchester: University of Manchester.
Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.
Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: selected readings*. Harmondsworth: Penguin.
Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-65.
Iwashita, N. (1999). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 8*(2), 51-66.
Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171-183.
Jenkins, J. (1997). Testing pronunciation in communicative exams. In M. Vaughan-Rees (Ed.), *A special issue of speak out: Bringing together the interests of the IATEFL Pronunciation and Testing* (pp. 7-11).
Johnson, & Tyler. (1998).
Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the oral proficiency interview.* New haven, CT: Yale University Press.
Katona. (1998). Meaning negotiation in the Hungarian oral proficiency interview. In R. Y. a. A. W. He (Ed.), *Talking and Testing: Discourse approaches to Assessment of Oral Proficiency.* Philadelphia: John Benjamins.
Kormos, J. (1999). Simulating conversations in oral proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*(2), 163-188.
Lazaraton, A. (1992). The structural organization of a language interview: a conversation analytic perspective. *System 20*, 373-386.
Lazaraton, A. (1996a). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing, 13*(2), 151-172.
Lazaraton, A. (1996b). A qualitative approach to monitoring examiner conduct in the Cambridge assessment of Spoken English (CASE). In M. Michael & N. Saville (Eds.), *Performance testing, cognition and assessment: selected papers*

from the 15th Language Testing Research Colloquium. Cambridge: Cambridge University Press.
Lazaraton, A. (1997). Preference organization in oral proficiency interviews. *Research on Language and Social Interaction, 30*, 53-72.
Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: UCLES/Cambridge University Press.
Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A micro-analytic inquiry. *Language Learning, 54*(1), 79 - 117.
Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language testing, 17*(1), 43-64.
Long, M. (1983). Native Speaker/Non-native Speaker Conversation and the Negotiation of Comprehensive Input. *Applied Linguistics, 5*, 177-193.
Lumley, T. (2002). Assessment criteria in a large scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 332-336.
Luoma, S. (2004). *Assessing speaking*. New York: Cambridge University Press.
Matthews, J. (1990). The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *ELT J,, 44*, 117-121.
May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated recall. *Melbourne Papers in Language Testing, 11*(1), 29 -51.
McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.
McNamara, T. F. (1997). 'Interaction' in second language performance assessment: whose performance? *Applied linguistics, 18*(4), 446-466.
McNamara, T. F., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242.
McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*, 140-156.
McNamara, T. F., & Roever, C. (2006). *Language Testing: the Social Dimension*. Malden USA: Blackwell Publishing.
Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes.* Unpublished master's thesis. University of California Los Angeles.
Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing 13*, 241-256.
Milanovic, M., Saville, N., & Shen, S. (1996). A study of the composition behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing and Research Colloquium.* (ed., Vol. 3, pp. 92-114). Cambridge England: Cambridge University Press.
Morrison, D. M., & Lee, N. (1985). Simulating an academic tutorial: A test validation study. In Y. P. Lee (Ed.), *New directions in language testing* (pp. 85-92). Oxford: Pergamon Institute of English.
Morrow, K. (1979). Communicative language testing: revolution or evolution? In C. K. J. Brumfit , K. (Ed.), *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.
Morton, J., Wigglesworth, G., & Williams, D. (1997). Approaches to the evaluation of interviewer performance in oral interaction tests. In G. Brindley & G.

Wigglesworth (Eds.), *Access: Issues in English test design and delivery* (pp. 175-196). Sydney: NCELTR.
Nakatsuhara, F. (2004). *An investigation into conversational styles in speaking tests.* Unpublished Masters thesis, University of Essex, Essex.
Nakatsuhara, F. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language and Linguistics, 9*, 83-103.
North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445-465.
North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*: TOEFL Monograph 24.
North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.
Norton, J. (2005). The paired format in the Cambridge Speaking Tests *ELT Journal* (59), 287 - 297.
Nunn, R. (2000). Designing rating tasks for small group interaction. *ELT Journal, 54*(2), 169-178.
O' Sullivan, B. (2002). Using observation checklists to validate speaking test pair tasks. *Language Testing, 19*(1), 33-56.
O'Loughlin, K. (2001). *The equivalence of direct and semi-direct tests of speaking.* (Vol. 13). Cambridge: Cambridge University Press/UCLES.
O'Loughlin, K. (2002). The Impact of gender in oral proficiency testing. *Language Testing, 19*(3), 277-295.
Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143-152.
Perret, G. (1990). The language testing interview: A reappraisal. In J. de Jong & D. K. Stevenson (Eds.), *Individualising the assessment of language abilities* (pp. 225-228). Philadelphia: Multilingual Matters.
Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to? In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: selected papers from the15th Language Testing Research Colloquium* (pp. 74-91). Cambridge England: Cambridge University Press.
Pollit, A., & Hutchinson, C. (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72-82.
Porter, D. (1991). Affective factors in the assessment of oral interaction: gender and status. In *Language testing in the 1990's* (pp. 32-40). Modern English Publications in association with the British Council: London: Macmillan.
Reed, J., & Halleck, G. B. (1997). In V. Kohonen, A. Huhta, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: proceedings of LTRC 1996* (pp. 225-238). Jyvaskyla: University of Jyvaskyla and University of Tampere.
Reves, T. (1981). The group oral examination: A field experiment. *World language English 1-2*(4), 259-262.
Ross, S. (1996). Formulae and oral interviewer variation in oral proficiency interviewer interaction. In R. Y. a. A. W. He (Ed.), *Talking and testing*.
Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition., 14*, 159-176.
Savignon, S. J. (1983). *Communicative competence: Theory and classroom practice*. Reading: Addison-Wesley.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53*, 42-51.
Scott, M. L. (1986). Student affective reactions to oral language tests. *Language Testing, 3*, 99-118.
Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal, 40*, 212-220.
Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan & M. Swain (Eds.), *In researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167-185). London: Longman.
Spence-Brown, R. (2003). *Authentic assessment: the implementation of an authentic teaching and assessment task.* University of Melbourne, Melbourne.
Spolsky, B. (1990). Oral examinations a historical note. *Language testing*.
Spolsky, B. (1995). *Measured words: the development of objective language testing*. Oxford: Oxford University Press.
Stansfield, C. W. K., D M. (1992). The development and validation of a simulated oral interview. *The Modern Language Journal, 76*, 129-142.
Storch, N. (2001). *Role relationships in dyadic interactions and their effect on language uptake.* Unpublished Ph D Thesis, The University of Melbourne, Melbourne.
Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275-302.
Taylor, L. (2001). The paired Speaking test format: recent studies. *UCLES Research Notes*, , from http://www.cambridgeesol.org/rs_notes/rs_nts6.pdf
Taylor, L., & Jones, N. (2001). Revising the IELTS speaking test. *Research Notes, 4*, 9-11.
Turner, C., & Upshur, J. (1996). Developing rating scales for the assessment of second language performance. *Australian Review of Applied Linguistics, Series S*(13), 55-79.
Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1 Spring).
Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 84-111.
Upshur, J. A., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3-12.
van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.
Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411-440.
Weir, C. (1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall.
Weir, C., & Milanovic, M. (Eds.). (200). *Continuity and innovation: Revising the Cambridge Proficiency in English*. Cambridge: Cambridge University Press
Weir, C., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English*. Cambridge: Cambridge University Press
Wigglesworth, G. (2005). Current approaches to researching second language learner processes. *Annual Review of Applied Linguistics, 25*, 98-111.

Wilds, C. (1979). The measurement of speaking and reading proficiency in a foreign language. In K. P. a. H. Jones, V. (Ed.), *Testing kit: French and Spanish* (pp. 1-12): Department of State: Foreign Services Interview.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah. NJ: Laurence Erlbaum Associates.

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*(1), 3-42.

Young, R., & Halleck, J. B. (1998). Let them eat cake or how to avoid losing your head in cross-cultural conversations. In R. Young & A. He (Eds.), *Talking and Testing: Discourse approaches to the assessment of oral proficiency* (pp. 355-382). Amsterdam, Philadelphia: John Benjamin.

Young, R., & Milanovic, M. (1992).

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

APPENDICES

Appendix 1 Trial task rating criteria (before devising rating scale for interaction)
Criteria for rating communicative interaction

Mark out of 20: one criteria per section.

Comprehension: understands the question and generally answers logically.

1) It is necessary to repeat slowly and to expand on and clarify what is being asked
2) The question can be repeated slower without rephrasing.
3) Understands well but appears to pay too much attention in order to comprehend
4) Replies immediately and the conversation continues
5) The question is understood effortlessly in a relaxed manner and as naturally as someone who understands everything

Communication: response to requests; complexity of structures and questions

1) Very few questions (i.e. ¿Y tú?), many errors (i.e. un-conjugated verbs) or they do not understand each other
2) The questions are basic with various errors, short responses searching for words; it is an effort to respond.
3) The questions are correct but not very varied. Responses are unconnected utterances even though an effort is made to combine phrases making a first step towards a natural oral text.
4) Different types of questions asked. Answers well, joining utterances.
5) Wide range of questions and responses, there is fluency and ease of expression with high text cohesions

Vocabulary: A variety of vocabulary showing levels of agreement

1) Minimal communication, difficult to rate
2) Very Basic, repeats in order to try to communicate. Articles are missing as are prepositions, no agreement of number or gender.
3) Adequate vocabulary but erratic agreement.
4) Adequate vocabulary for the topic and the level with some errors of agreement but with some self correction.
5) Wide range of vocabulary with errors in agreement in number or gender.

Grammar verbs: agreement/ tense/ irregular forms / changes of stem/ expressions with the infinitive / after prepositions

1) Minimum communication difficult o rate.
2) Very basic: repeats to try to comunícate, no pronouns, no agreement; verbs in the infinitive (yo dormir) o dos conjugados/ infinitivos (me\yo gusta duermo/ gustar dormir)
3) Poca variedad de verbos pero adecuado para el tema; mezcla las personas: la concordancia es errática
4) variety of verbs for topic and level. Concordancia casi siempre, con autocorreción
5) Error free range of structures.

Appendix 2

Task trial Spanish oral transcription with English gloss
Male and female
Beginner level students 12 weeks intensive Spanish at university level
Task:
to introduce and maintain conversation on three topics each in a conversation 10 minutes long

Topics: Sport and Exercise — Celebrations
Leisure activities — favourite Season
Family — Weekend activities

J has studied French and S has studied Italian

1. S: ¿Qué vas a hacer esta tarde?
2. S: What are you going to do this afternoon?

3. J: Ah, esta tarde vas, voy a mi casa y a hacer
4. J: Ah, this afternoon I'm going home to do my

J; mi tarea y porque necesito , necesito hacer una una escrita* de historia
5. J; my homework because I need, I need to do a an essay for history

6. S: !Qué aburrido! Risa
7. S: How boring

8. J: Es un poco aburrido y yo quiero, pause , necesito leer
9. J; It is a bit boring and I want….I need to rea

10. J; mucho es, es, es sobre nacionalismo y esa una especialidad
11. J; lot it is ,it is on nationalism and that is for my major

12. J; eh… pienso es muy muy interesante
13. J;eh I think that it is very very interesting

14. S: Sí, sí um voy a tomar café con mi hija después
15. S:Yes yes mm I am going to have coffee with my daughter

16. J:Ah
17. J:Ah

18. S:Porque mi hija trabaja en la universidad
19. S:Because my daughter works at the university

20. J:Si. Tienes muchas hijas?
21. J: Do you have many daughters?

22. S: si tengo tres hija hijos, dos hijas,pero vive solamente
23. S:Yes I have three daughters, children, but they live alone

24. J:¿ cómo se llama?
25. J:What are their names?

26. S:mi hija mayor es Fiona y ella es esposa.ah, su esposo se llama Ben mm
27. S;My eldest daughter is Fiona and her wife ah her husband name is Ben mm
28. S;y vivo en Sydney .viva en bondi beach y trabaja sí
29. S;and they live in Sydney . they live on Bondi Beach and work yes

30. J;Ah
31. J;Ah

1. S:Si Bondi Beach si. Kate es mi hija mayor.
2. S;Yes Bondi Beach yes Kate is my eldest daughter

3. J;Vas a viajar con Kate? si?

4. J;You are going to travel with Kate? Yes?

5. S: Si en tres semanas voy a ir a , um en europa con mi hija kate,
6. S;Yes in three weeks I am going to go to um to Europe with my daughter Kate
7. S;y viajar en europa
8. S;and I am going to travel with my daughter

9. J: no este fin de semana?
10. J;Not his weekend?

11. S: No, no
12. S;NO ,no

13. J: tienes planes para este fin de semana?
14. J;Do you have any plans for this weekend?

15. S: este fin de semana uh, uhm pienso que mm ir a la playa en portsea porque
16. S;This weekend uh ,hm I think I am going to the beach in portsea because
17. S;mm mi casa necesita muchas um trabajo porque durante el invierno um prefiero
18. S;mm mi house needs lots of um work because during the winter um I prefer
19. S;um trabajar en el jardin
20. S;um to work in the garden
21. S;Y no
22. S;And not

23. J: invierno es tu estación favorita?
24. J:Winter is your favourite season?

25. S: no,no
26. S:No no

27. J: porque?
28. J;Why

29. S: me gusta esquiar en el invierno pero… gusto mucho el verano
30. S:I like to ski in winter but I really like Summer

31. J: también me gusta verano. Muy bien nadar en la mer.
32. J:I also like Summer . It's great to swim in the sea.
33. J:hacer tiempo en el jardin
34. J:Or to spend time in the garden

35. S:Si,si
36. S:Yes yes

37. J:Si es verdad
38. J:Yes it's true

39. S: tienes hijos
40. S:Do you have children?

41. J:No
42. J:No

43. S: o Hermanos?
44. S:Brothers and sisiters?

45. J: no no tengo hijos
46. J:I don't have any children

47. S: risa
48. S:Laughter

49. J: Pero ,y uh, tiene tengo tengo dos hermanas
50. J:But, and uh, I do have I have two sisters

51. S:Si
52. S:Yes

53. J:Si se llaman Romy y Tania las dos se, mm las dos son
54. J:Yes their names are Romy and Tania both mm I don't

55. J:mi no divertidas
56. J:think either of them is much fun

57. S:Si?
58. S:Yes

59. J:Si
60. Yes

61. S:Estas,estan estudiando?
62. S:Are they are they studying?

63. J: Ah, si estan estudiantes en la escuela segundo
64. J:Ah yes they are students at secondary school

65. S:Si, si
66. S:Yes yes

67. J: Si mm, y um tambien, tambien mi familia
68. J:Yes mm also , also my family
69. J:tengo uh tengo uh muchos animals
70. J:I have a lot of animals

71. S: as si animales. Perro?
72. S:A yes animals. A dog?

73. J :L si tengo un perro grande se llama tess . es su una llama de shakespeare
74. J:Yes I have a large dog her name is tess. It is a shakesperean name

75. S:Ah si
76. S:Ah yes

77. J:Y um
78. J:and Um

79. S:Que divertido
80. S:How funny

81. J: mis hermanas tienen dos gatos y dos uh piscos y uh y no tiene mas. Risa
82. J:My sisters have two cats and two fish and they don't have any more laughter

83. S: ah si ehh Y que te gusta hacer durante las vacaciones?
84. S:Ah yes ehh and what do you like doing during the holidays?

85. J : mm durante las vacaciones? Tengo dos meses uh de vacaciones y voy oh voy
86. J:Mm during the holidays I have two months of vacation and i'm going oh going

87. J:a salir con mis amigos y leer mucho y escuchar musica y
88. J: to go out with my friends and read a lot and listen to music and
89. J:yo quisiera voy a esquiar ah y se esquiar pero no puedo esquiar pero es cuesta mucho
90. J:I would like to go skiing ah and I know howe to ski but I can't ski beacuase it costs a
91. J:lot dinero
92. J:of money.

93. S:Si
94. S:Yes

95. J:Si
96. J:yes

97. S: si es verdad pero es muy divertido
98. S:Yes that is true but it is great fun

99. J:Entonces, entonces, si es muy divertido entonces yo pienso
100. J: then, then yes it is great fun and then I think
101. J:que no es posiblemente esquiar estas vacaciones.
102. J:That… it is not possible to go skiing these holidays

103. S: si
104. S:Yes

105. J s
106. J:Yes
107. J: y tu? Tienes planes por las vacaciones?
108. J:And you? Do you have any plans for the holidays?

109. S: suspiro si despues viajar en Europa

110. S Sigh yes after I travel to Europe

111. J: ah
112. J: ah

113. S es dos o tres semanas umm tengo tengo ganas de esquiar pero no lo se puedo
114. S :In two or three weeks mm I want I wasn't to ski but I don't know if I can

115. J: tengo ,tengo tambien tengo ganas hacer mi tarea
116. J:I feel I feel like also I feel like doing my homework

117. S: oh risa que aburrido
118. S: Oh laughter how boring

119. J: es aburrido pero yo necesito
120. J: It is boring but I need to
121. J: mucho um y si no tengo ganas uh no no es posiblemente, hacer
122. J Very much um if I do not feel like it uh it isn'y isn't possible to do
123. J: mi tarea.
124. J: my homework

125. S:No me quiero hacer tarea durante las vacaciones
126. S:I don't want to do homework during the holidays

127. J: no mi quiero pero uh no uh necesito hacer
128. J:I don't want to but uh no uh I need to do it

129. S: Si si
130. S: Yes yes

131. J: si Celebras la navididad? La navidad
132. J: Yes, Do you celebrate Christmas

133. S: si si siempre con mi familia a la casa
134. S: Yes yes With my family at home

135. J: si Y y um tambine con tu hijos , tu hijo que vive en bondi??
136. J: Yes and and um also with your sons, your son that lives in Bondi?

137. S: Si. Mi hijo que vive en bondi siempre ricorgere a mi casa por la navidad
138. S: My son that's living in Bondi always come back to my house for Christmas

139. J: Por la navidad
140. J: Fo Christmas

141. S: Si si si
142. S: Yes yes yes

143. J ah Muy bien
144. J: That's good

145. S: si
146. S: Yes

147. J no celebra la navidad en el momento
148. J:I don't celebrate Christmas at that moment

149. S: Porque
150. S: Because

151. J: Tambien uh si porque yo celebro el januca
152. J: Also uh yes because I celebrate Januca

153. S: A si cuando es el januca
154. S: Ah yes when is januca?

155. J : en el tiempo u h el tiempo de la navidad normalmente si y um
156. J: At the time uh Christmas time normally yes and um

157. S: solomente por un dia?
158. S: Only for one day?

159. J ocho dias
160. J: Eight days

161. S: ocho dias
162. S: Eight days

163. J oh, ocho noches
164. J Oh eight nights

165. S: ocho noches
166. S: eight nights

167. J : si es um muy divertido
168. J Yes it is um great fun

169. S:Es um un um semana um divertido o serio?
170. S:Is it um a um fun week or a serious week?

171. J Le es muy divertido es una celebra celebro de vivo
172. J: It is great fun it is a celebration, I celebrate living

173. S: Ah
174. S: Ah

175. J: Celebra de vive
176. J: You celebrate life

177. S: Si
178. S: yes

179. J: Con las gentes
180. J :with poeple

181. S: Con regales
182. S: With presents

183. J: Ah, si uh no con muchos regalos pero con mucho mucha comida
184. J:Ah yes uh not with many presents but with a lot of food

185. S: si si comidas especial?
186. S: Yes yes special food?

187. J: si,ah,no con mucho mucho comide de leche
188. J:Yes ah no with a lot of dairy products

189. S: a si
190. S:Ah yes

191. J: Si especial con mis primas.Tienes una familia grande?
192. J: Yes especially with my cousins. Do you have a large family?

193. S: no no tengo una familia grande porque nacio en inglaterra eh muchas
194. S:NO NO I don't have a large family because I was born in England eh a lot of
195. S: mi familia en Inglaterra
196. S:my family in England

197. J: en inglaterra ah en inglaterra tienes una ciudad favorita?
198. J: In Engalnd ah in England do you have as favorite city?

199. S: si si Me gusta mucho Steeton. Es una ciudad pequeno pero es vecino
200. S:Yes yes I really like Steeton It is a small city but it is next
201. S;de colina, vecino de yorshire dales
202. S:to the hills next to Yorkshire dales

203. J: en inglaterra , no viajar a Inglaterra pero en Inglaterra mi
204. J in England I haven't been to England but in England my
205. J:ciudad favorita es Manchester
206. J:favourite city is Manchester.

207. S:Manchester (risa) no es una ciudad muy bonita es muy industrial
208. S:Manchester laughter it is not a very nice city it is very industrial

209. J: si pero me gusta manchester united
210. J:Yes but I like Manchester united
211. J: pero
212. J:But

213. S: ah si si si! Risa

214. S:Ah yes yes yes laughter

215. J: pero
216. J:But

217. S: te gusta deporte?
218. S Do you like sport?

219. J: me gusta ver pero no me gusta jugar Risa pero levanto pesas
220. J:I like to watch but I don't like to play laughter but I do weight training
221. J:but I don't…
222. J:pero no

223. S: oh que fuerte que fuerte!
224. S:Oh how strong how strong.

225. S J Risa
226. S JLaughter

227. J:Gracias gracias pero no es verdad pero no si no tengo mucho tiempa mucho
228. J:Thanks thanks but it is on true but I don't if I don't have much time.
229. J:tiempo libre jugar al futbol pues.
230. J:free time I play football so.

231. S: ah si Juegas al tennis?
232. S:Ah yes do you play tennis?

233. J: no no no juego nada
234. J:Nono I don't play anything.

235. S: levantas pesas si?
236. S:You do wight trainging?

237. J : ah si
238. J:Ah yes

239. S:En un gimnasio
240. S:In a gym

241. J: Si en el gimnasio en la universidad No cuesta mucho dinero
242. J:Yes in a gym at university. It does not cost much money

243. S: es verdad
244. S:That's true

245. J: si es muy bien
246. J:Yes it si very good

247. S:mi hija Fiona se gusta mucho
248. S:my daughte Fiona likes it a lot

249. J: a tu hija/
250. JYour daughter?

251. S: si si pero
252. S:Yes yes but

253. J : ah se trabaja en la universidad
254. J:Ah yes she works at university

255. S: si si si
256. S:Yes yes yes

257. J: ah
258. J:Ah

259. S ahora si
260. S:Now she does

261. J: es muy bien trabaja cerca de deportes favoritos
262. J:It is good to work near your favourite sport

263. S: si si si es verdad
264. S:Yes yes yes it si true

265. J: si
266. J:Yes

267. S:gracias muchas jo
268. S:thanks a lot jo

269. J Y tu muchas gracias
270. J:And thanks to you

Appendix 3 Final band scales to which paired interaction grid is added: Criterios para calificar la comunicación oral

Expresión: Complejidad de las preguntas y comunicación de la información pedida**.**

1. Las preguntas son mínimas (ej. ¿Y tú?), tienen muchos errores (verbos en infinitivo) o no se entienden y responde con pocas palabras.
2. Las preguntas son básicas con varios errores y responde con frases sueltas y se expresa con esfuerzo bastante marcado buscando las palabras. Hay silencio entre palabras en las frases.
3. Las preguntas son correctas pero no muy variadas. Contesta con frases no todas unidas aunque se oye esfuerzo para ir combinando las frases hacia el primer paso en un texto oral natural. Es correcto porque es muy limitado.
4. Hace varios tipos de preguntas. Contesta ampliamente uniendo algunas frases. Si no sabe la palabra puede parafrasear para expresar lo que necesita.
5. Pregunta y contesta ampliamente, hay fluidez y facilidad de expresión. Demuestra alta cohesión de texto. Responde con rapidez y fluidez. Si se para es para pensar y buscar la palabra exacta, y no por problemas de comprensión o de expresión.

Variedad de vocabulario: Tiene variedad de vocabulario con un nivel de concordancia.

1. Un mínimo de comunicación difícil de calificar.
2. Muy básico: repite lo mismo para intentar comunicar, faltan artículos y preposiciones. Sin concordancia de número o género.
3. Vocabulario adecuado para expresar básicamente lo esencial, para hacerse entender, pero la concordancia es errática.
4. Vocabulario y estructuras son adecuadas para el tema y el nivel con algunos errores de género y concordancia pero con auto corrección.
5. Vocabulario y variedad de estructuras amplias y los errores de género o concordancia no causan confusión o malentendidos.

Gramática: Verbos (concordancia/tiempo/irregulares/expresiones con infinitivo), concordancia de género y número, uso de preposiciones, etc.

1. Un mínimo de comunicación difícil de calificar.
2. Muy básico: repite lo mismo para intentar comunicar y faltan pronombres, sin concordancia alguna. Verbos en infinitivo (yo dormir) o dos conjugados: me \ yo gusta duermo/ gustar dormir.
3. Poca variedad de verbos pero adecuado para el tema, mezcla las personas, la concordancia es errática.
4. Verbos variados para el tema y el nivel, concordancia casi siempre con auto corrección.
5. Sin errores. Mucha variedad.

Pronunciación:

1. Habla con una pronunciación que causa confusión y mucha dificultad para el oyente.
2. La pronunciación no es clara y hay problemas en comprender lo que dice sin poner mucha atención.
3. La pronunciación es clara pero marcadamente no nativa. Comprende frases simples y cortas.
4. La pronunciación no es perfecta, pero se aproxima a un nivel en que aunque se detecta algo no causa problemas de comprensión.
5. La pronunciación no causa ningún problema de comprensión.

Interacción: Left to right

<table>
<tr><th>Question 1</th><th>answer</th><th>Question2</th><th>answer</th><th>Question 3</th><th>answer</th><th>rating</th></tr>
<tr><td rowspan="6">Supportive body language?</td><td rowspan="4">yes</td><td rowspan="4">supportive listener?</td><td rowspan="2">yes</td><td rowspan="2">questions /replies mostly show Cohesion b/n and within topics?</td><td>yes</td><td>5</td></tr>
<tr><td>no</td><td>4,5</td></tr>
<tr><td rowspan="2">no</td><td rowspan="2">Reasonable Turn length?</td><td>yes</td><td>4</td></tr>
<tr><td>no</td><td>3.5</td></tr>
<tr><td rowspan="3">no</td><td rowspan="3">relevant questions/answers are given?</td><td rowspan="2">yes</td><td rowspan="2">Asks/Answers within a comfortable time?</td><td>yes</td><td>3</td></tr>
<tr><td>no</td><td>2</td></tr>
<tr><td>no</td><td colspan="2"></td><td>1</td></tr>
</table>

Appendix 4

ethics permission

PLEASE READ CAREFULLY.
PROJECT:
PAIRED ORAL ASSESSMENT IN A BEGINNERS' LANGUAGE TEST IN SPANISH.
SIGN BELOW IF YOU ALLOW YOUR TAPE TO BE USED FOR RESEARCH.

Ana Maria Ducasse, from the Spanish Program, LaTrobe University (tel.9479 2437), the sole researcher in this project, requests the voluntary use of your Spanish oral assessment tape for research. The research which is to be conducted looks at the development of language tests. Samples from three universities as well as morning and evening groups will be used so your privacy is protected by the numbers in the study.

I understand that the recording made of the Spanish oral assessment may be used for research purposes. My name or level will not be disclosed and I will be able to read any material published or presented on request. It makes no difference to my mark whether I choose to volunteer the tape for research into language testing.

I understand that my oral assessment MAY be used for research and I give my permission.
signed dated

Appendix 5

La Trobe University Ethics clearance

RESEARCH AND GRADUATE STUDIES OFFICE

Ms Ana Maria Ducasse
Spanish Program
Room 203
Old Arts Building
The University of Melbourne
Victoria 3010

29 October 2003

Request for Ethics clearance to recruit La Trobe University students in a research project

Dear Ms Ducasse

Thank you for submitting a copy of your ethics application, advertisement and Plain Language Statement for the research project "Interaction in a Beginners Spanish Oral Proficiency Test for the Degree of Doctor of Philosophy in Applied Linguistics (Reference: HREC 030552)" which you are undertaking at the University of Melbourne under the supervision of Professor Tim McNamara.

On behalf of the La Trobe University Human Ethics Committee, the Acting Chairperson has reviewed and supports your proposal to recruit 25 students from the Spanish Department at La Trobe University for the above research project on the condition the following points are observed:

1. The procedure outlined by Jacky Angus, in her correspondence dated 22 October 2003, with respect to dependency (2.5) should be followed.
2. Please notify the Head of School, Historical & European Studies at La Trobe University of this research project.
3. Subject to the approval of the University of Melbourne HREC, please indicate on the recruitment flier that you are a student of the University of Melbourne and your contact details.
4. Recruitment cannot commence until the researcher receives final ethics approval from the University of Melbourne.

Please contact me by telephone on 9479 1443 or by e-mail m.junge @latrobe.edu.au should you wish to discuss this matter further.

Yours sincerely

Ms Mira Junge
Secretary, La Trobe University Human Ethics Committee

Appendix 6 Workshop guide used in Spanish

Guía del taller: Elaboración de una escala de interacción

1 Selección de 8 parejas para elaborar la escala:

En el CD miramos hoy las parejas números
3 4 5 6 8 9 11 13
Hay buenas, malas, regulares y bastantes desiguales para que las escala se pueda adaptar a todo.

2 familiarización

Tenéis unas tarjetas con los comentarios hechos de la pareja y de los individuales
Miramos como empieza cada pareja en el CD ROM
Para tomar apuntes al decidir el orden de las parejas de 1 a 8 de mejor a peor

Pareja 1	izquierda	derecha
2		
3		
4		
5		
6		
7		
8		

3 Resultado de los comentarios sobre interacción

Seis puntos que resaltan de los comentarios de los profesores.
Para que resulte la interacción hace falta, y no en este orden que es alfabético….

1. Apoyar visual y corporalmente. Ej. mirar al hablar usar las manos o la cara
2. Ayudarse mutuamente Ej. llenar silencios, dar la palabra que falta
3. Hacer preguntas adecuadas Ej. que significa que se ha escuchado
4. Mantener cohesión del texto Ej. cambiar oportunamente de tema
5. Ser buen oyente Ej. asentir con si, si y comentarios y entonación y postura que muestra interés y que se sigue la conversación
6. Turnarse de una manera equilibrada y fluida pero no hablando demasiado.

4 Elaboración de la escala

1. **Primer paso ….**
    - **Se ven como pareja al principio, en la mitad o al final? Nunca?**
    - En pares *separad en dos grupos* los pares

Tened en cuenta como profesores qué característica valoráis primero al observar una pareja de estudiantes hablando?
Primer pregunta:

Se formula la pregunta y se dividen los 'si' y los 'no'

Llevan gafas? O han desayunado?
Es lo que más resalta sin demasiado detalle

2. **Segundo paso**

Dividir los del ´´si´´ con otra pregunta que necesita fijarse en más detalle

3. **Tercer paso**

Dividir los del ´´no´´ con otra pregunta que necesita fijarse en más detalle

Seguir hasta que estéis seguros o contentos de que funcione. Hay tablas en blanco para rellenar

Probamos para ver si funciona alguna versión mejor que otra para quedarnos con una aunque sea modificada al final para usar en octubre

Al final si hay tiempo probaremos las escalas con cualquiera de las parejas que faltan
1, 2, 7, 10 ,12, 14, 15, 16, 17,

Appendix 7

The three Verbal Protocols from Rater 4 on pairs 7, 8 and 9

Rater 4 pair 7 transcription of verbal protocol divided into ideas units

1. Hola anamaria pareja numero 7
2. 1 think the interaction in this pair does work, it does work especially because he makes a lot of effort to make it work/
3. There are a number of differences between him and her between the way they speak and the way they communicate with each other./
4. You can see that he has his legs crossed and she doesn't in the beginning this doesn't look like an important thing. But it really shows his attitude towards the ten minutes speaking there./
5. You can see that he actually uses a lot of gestures /
6. his intonation is a lot closer to the real thing /
7. and he makes an effort to have constant visual contact with her he is actually looking for her eyes a lot more than she does his eye /
8. Umm 1 think he is not um, he doesn't seem interested in passing an exam as I think she is or interested in communicating with her through the ten minutes /
9. you can see that he is constantly looking for ways to get his message across /
10. he laughs he moves his hands a lot he is looking at her /
11. she hardly ever laughs /
12. her visual contact is all over the place /
13. and there are actually a couple of moments where he laughs for the first time and she makes another joke he laughs and then she doesn't laugh ever again /
14. Again at the beginning of the exam, she is not making any gestures with her hands then he starts using his hands and you can see her starting to use her hands and in her somehow she is following the path he is marking /
15. Also the way they are sitting you can see that he is sitting in an upright position /
16. and speaks with quite a lot of self confidence though her Spanish is quite good she doesn't appear that self confident in the way she is leaning on the chair it looks as though she is not that confident with what she is going to say /
17. In a way 1 think the way they use correction again stresses stresses what 1 am saying . he hardly ever uses self correction for himself he prefers to continue speaking even though when he uses sentences where he is making mistakes he um realizes he is making a mistake and he um in the next sentence he corrects himself /
18. whereas she makes a huge effort to speak accurately in every sentence that makes her look very hesitant. She stops suddenly and her answers are not that fluent it stops and limits the interaction actually./
19. For example when he says when he is asked how long he will have to study to be a historian he says ' por lo mas poco tres anos' which is completely incorrect but it works he says and he continues talking /
20. in her case on more than one occasion she thinks and looks for the right word, the right tense 1 think or for um something um then the interaction um/
21. 1 did like though how they changed their topics and how they went from one to the next one. 1 think that they waited until they had exhausted on topic. They made quite a few questions questions on every topic and when they move onto the next one it looked as though they had said enough in the previous one. /
22. And it looked quite fluent it did not seem to be a problem as they changed from one to the next one /
23. maybe what was missing was a bit of connection between one topic and the next one /

24. they were talking about something and soon they were talking about the next thing in between what they were saying /

25. In any case to sum up his interaction works because he is communicating her interaction fails because she is trying to pass an exam!/

26. Rater 4 pair 8

27. Me again now we are going for pair number 8 /

28. actually it is quite interesting to see two pairs. Once 1 have seen this one 1 think 1 have to change some of the things 1 said before for the previous pair /

29. for me the most important thing for this pair in pair number 8 was to see how specially him couldn't care less about what she was saying he just was showing off /

30. well first there was a very clear umm difference in the sense that he is asking and she is answering and each one of them accepts that rule ah then they change it and he asks and she answers it doesn't seem to be much interaction in the sense that one person is asking and replies and urn 1 didn't think that in the other way half way through their answer in a sense they have very clear the roles are very clear and divided in the sense that there is not really and interaction. One is asking and one is replying /

31. And um when it is his turn to ask he did not seem really interested in what she said he couldn't careless what she said or where the conversation was going urn he seems to be more interested in what his next question was going to be he is specially clear when he is asking , he is talking about schools he asks something she replies and the next question that he asks has nothing to do with what she has said and a third question again has again absolutely nothing to do with what he had been saying/

32. Rater 4 pair 9

33. At one stage they think they have done enough talking and they move on to the next topic without interrupting at all /

34. about asking He was asked she replied (…)(they use some vocabulary that was not all that good 1 thought it was quite bad) they think they have done enough And when they are talking about buying the next topic she pushes him a bit more and she asks him one question which is related to why he has replied on the third question when again she asks about what he has replied so there if SOME connection between the questions /

35. um 1 think sometimes the questions say a lot more than the answers are they related to what the other person has as is or are they out of the blue and have nothing to do with what the other person has asked? In the first minute or twenty minutes later it is at the moment when they are asked that 1 realize now that it is quite important /

36. Um in terms of the physical interaction between the two of them um well 1 think it is very telling them she has her legs crossed and one of her arm or hands is between her leg as if she is not trying to move that hand that much maybe 1 put a lot of emphasis on the movement of the hands being Spanish but a 1 think it is very telling. In the first pair number 7 he had a very successful interaction because he was using his hands and he was really into it /

37. but um these two 8 are really not very interested in communicating

38. um Urn He is again very hesitant in the way he speaks/

39. there is some self correction but again he is not too good and he is thinking too much about what he is saying instead of about how he is saying it/

40. his intonation is very bad in the sense that I think he is he felt very very flat intonation he mentions sometime sometimes he is a bit better. /

41. Also all through the exam both of them are extremely serious. 1 mean you don't have to laugh right through an exam but um smile very so often recognize and make a gesture to your partner that you understanding what the other person is saying/

42. Um they don't know they' don't say yea they don't move their body they don't communicate physically so that they are understanding 1 don't know in general terms I would rate them as as very poor interaction /

43. and one last thing is she chewing gum

Appendix 8

Candidate stimulated retrospective Verbal Protocols

pairs 7, 8 and 9 the same pairs rater comments on above

Pair 7 candidate 5

1. V really has a thing with numbers I really shouldn't; have asked her she got stuck with that in the first oral
2. I try to concentrate to see we do understand each other but with the tenses we both know the conjugations but we have to figure out what each verb is
3. We had had this long conversation before but when we practice we went off on tangents we talked about art a bit so we had discussed this
4. I am just waiting and thinking how am I going to answer that one
5. I don't sound really good there I had never really thought about that question in English so what could I say
6. There was a bit of a pause as she looked like she wasn't going to ask me another question
7. So, we had already talked about that
8. beforehand yeah no but I think we'd talked about ah no I'd mentioned it before but ah I guess she had to ask it she had it on her thing so she had to ask that
9. I was wondering was I meant to answer that? I know that, her birthday I would have had to think about a bit more
10. I didn't really know where we were going but it was good 'cause you now I didn't really know what to say but I generally tried to say something
11. Yeah yeah well it's a we always jump on the tenses and trying to figure it all out but we'd done it we had done this bit
12. Yeah we had talked about Wilson's Prom already
13. Yeas no I just was seeing if they were
14. It was a bit hard to pick up when to talk to each other I thought there was more of a flow actually
15. I thought it flowed a little better I guess it was alright it was harder than I remember to put it all together it was stopping and starting a lot it wasn't that bad
16. You had to more just ask a question and ask a question back when we were practicing we talked a lot generally I ma not sure it was that cohesive she would talk about Frieda kahlo and she wanted t be an artist then talk about Frieda kahlo we went off the topic more something we had in common we did that generally
17. then we had to get through a certain amount of topics in a certain amount of time so you spend your time trying to figure out how are we going to work this in
18. I think if there had been more topics we might have been pushing it Virginia did as well we both did ok on that frame
19. You normally never know what you sound like at all the pronunciation is interesting
20. You hear the grammar more than you think
21. It was more relaxed in the practice it was good though in class you never have a natural conversation maybe we are meant to laughter

pair 8 candidate 27

22. I am very nervous obviously as you can see I am fidgeting with the paper/
23. I am just trying to think of the right tense to say what I was doing/
24. I am thinking of the next question not really listening while she is talking./
25. I just can't get the words out like normally it is fine but I can't get the words out/
26. I did that hand wave for her to ask the next question/
27. I got a bit more confident after this stage I felt a lot better/
28. I was sort of thinking at the same time and I knew which one I was going to ask next/
29. I should have helped her out there/
30. The problem is that I wasn't really wasn't listening to what she was saying so should I have helped her out?/
31. We laughed because the person she lives with doesn't clean at all/
32. Yeah she wasn't ready but I think I didn't have any more questions left so there was nothing I could do I had asked my three/
33. we had sort of prepared this if this question came up we had this sort of scenario so I was trying to remember what we had prepared/
34. I was trying to think of we eat together but I knew it in French but I couldn't think of it in Spanish I didn't like the question so I moved on pretty quickly so I changed ti to what wanted to /
35. I' ve got anther question and I didn't realize/
36. I talk about it remembering the names of the foods was not my specialty so
37. I went pretty mush as I thought I went normally we go really well in class I think B was a bit let down

38. it is not the camera
39. me in particular I hate exams more in the practical sense like essay and the normal interaction
40. A lot of it was prepared and that was part of the problem we were trying hard to remember the script so we had to remember the tenses we were more confident being more confident so worried that you are under the pressure of getting it right undermines your confidence

pair 8

41. I am just so nervous and my hands are so sweaty/
42. What is he asking I am thinking I am afraid that I won't understand his question/
43. I laughed' cause I didn't know what to say I don't know him I have known him/
44. I am trying to decide which one to choose and to formulate the question and think I have to get this right I have to get this right/
45. I was just trying to work out what to say thinking the right way to say it and I am confused in my brain I think/
46. I always do that when I am nervous /
47. Iam kind of listening but I am trying to think of what to say next/
48. That 's French/
49. I am thinking now it is your turn/
50. I am thinking what more can I say what more can I say finding words I can remember/
51. I am laughing probably because am nervous/
52. I am thinking now I am going to ask if he wants to go shopping/
53. we were trying to do a do you want to go shopping dialogue but I can't remember how it goes/
54. I am probably thinking what to ask which topic to choose /
55. I didn't get to ask about the secondary school because I didn't know how to ask about it so I tried to avoid it as much as I could/
56. I am thinking like should I help him or should I let him keep on going /
57. it is confusing if someone helps when it is coming out you just want o keep on going/
58. I don't think I realize what i am doing

59. I think I yeah that I was very nervous in the beginning /
60. en it got better after a while I knew I as thinking and formulating and thinking too much in my head/
61. It would have been enter to do some mistakes and do um ah um ah it was stupid really because we practice so much before we came but for some reason we got really nervous before we came but that s life

Pair 9

62. I remember at the very start I was a bit frozen I couldn't remember how to say where did you go to school so I changed the question to something else like did you like school
63. Um I felt it was good but K was a bit softly spoken and that sometimes I didn't know if she didn't understand so I tried to lead her in
64. I was waiting for her to introduce one of her topics so I jumped back in with one of mine
65. I found that a bit difficult some times because I was waiting for her to ask questions back but she didn't so I had to fill in
66. When I said that I didn't know if Lebanese was the right way to say it but I said it anyway
67. I think that I had covered all my topics so I thought I would give her a chance but she didn't
68. Um I don't think she understood me so then I answered
69. Just before we came in we met up to practice for a little bit but I can't remember which of our questions we had already looked at
70. I don't know if that is
71. I forgot what I was saying then I tried to think of how to say how to rent videos but I couldn't so then I had to say something else to get around it
72. A lot of the time when she spoke I couldn't understand what she said but I just picked out the key words and then put them together
73. I didn't quite know if she said if I spent much time at home so instead I just started saying all the things I did at home
74. She said something I didn't understand so I said no so she wouldn't say any more
75. I thought it was new but then I said old I got confused and I corrected myself
76. It felr better than it looked I didn't speak as clearly as I thought there was more psuing that I thought
77. Um just trying to remember all the grammar I know what I wanted to say tehn I had to think I had to think of all the things I wanted to say sometimes had to think of it in English first then I had to pause I can think in Spanish but I cant talk straight away because I leave out the person or the plural so I took extra time to
78. We practised most of the topics I knew all the vocab and all the stuff it was a matter of just getting it all together Karen was not sure of herself so I told her so I might say some her questions just in case so she could prepare herself

Appendix 9

Pairs 7, 8 and 9 rough transcription as a guide with no gloss, not used in study.
ONLY videos used in study

Same set of pairs that the raters comment on in appendix 7 and from the candidate study appendix. 8

Both listen both talk

1. R Como estás
2. V ah como estas. bien y tu
3. R bien gracias y como se llama
4. V Me llama me llamo virginia
5. R um donde vives
6. V vivo en Preston en una casa ah con mi novio ah y tu donde vives
7. R um Vivo en moonee ponds en los barrios oeste de melbourne y um vive con mi familia. Ah cuantos pisos tiene su casa¿=
8. V un piso e solamente. vive en una casa o un apartamento?
9. R Si si vive en una casa
10. V si tu casa tiene una o dos pisos
11. R tienen un pisos vivo en una casa es una casa viejo construyo en 1910 y su casa viejo o joven
12. V Mi casa es vieja también
13. R Sabes en que a;o construyo?
    V ah no se exactamente pero pienso que construye en mil um mil um
14. R Novecientos_
15. V Si Si empieza del sigo. Quieres Um quieres um la que quieres um carrera seguir
16. R No se exactamente me pero me interesa la historia y la cultura de mejico generalmente asi tal vez me gustaría ser un historiador y tu que carrera quisiera seguir?
17. V Quiero pinto pintar me gusta pintar los retratos en particular y un una dia quisiera entrar en el archibald prize
18. R si _ impresionante
19. V para cuantos años necesitas estudiar para ser historiador?
20. R Por la la mas poco um tres años así tengo dos años mas estudiar pero se puede hacer master de universitario y phd pero no no se que que hac hago este año.uh Que tipo de restaurante es su favorito
21. V Me gusta um restaurantes italianos porque me gusta comer um pasta con copa de vino tinto y tu_
22. R Me gusta los restaurantes mejicanos porque me gusta mucho la comida mexicano especialmente tostadas y quesadillas pero pienso que la comida mejicana es diferente en México que en Australia
23. V Tienes uno favorito_
24. R No
25. V Vas a taco bil
    R si si es bueno me gusta vas a taco bil
26. V Si una o dos veces iba a taco Hill.
27. R Y te gusta
28. V Si tiene grande nachos
29. R Si si y grande margarita.
30. V Um Vives con la familia_
31. R Si vives con mi padres y mis dos hermanos mi hermana mayor y mi hermana menor.
32. V Cuantos anos tienes tu madre_
33. R No le gusta te decirte pero tiene 52 dos años
34. V 52
35. R 52 si y mi padre es el mismo edad. Pero vive solo con su novio
36. V Pero mis padres tiene sesenta y pocos años llevas bien con tus familia
37. R Uh si nos llevamo sbastante bine como la mayoria de las familias sabes._ y cuando eras niña a dode ibas para sus vacaciones_
38. V Cuando era niño mi familia y yo íbamos a lugares diferentes usualmente pero íbamos a lugares muy cerca de la playa y tu donde cuando eras niño dónde iba ibas de vacaciones
39. R Tomaba vacaciones con mi familia también y todos. también íbamos a lugares diferentes pero todos los anos íbamos a la peninsular mornignton porque mis abuelos tenia n una casa de vacaciones allí. También íbamos a wilsons promontoria Gippsland del sur
40. V mis abuelos Vivian allí Nosotros íbamos a fosterÍbamos a foster a veces para cenar
41. R Tiene un salon

42. V Y la comida muy interesante no mucho vegetariano
43. R No son nadie en mi familia so no es no era uno problema para nosotros. Si y Um
44. Y me gusta wilsons promontro mucho porque Ok necestat partir
45. V oh Ciao hasta luego hasta.

Pair 8 they do not listen to each other

1. K Hola me llamo Karoline y tu
2. D Hola Me llamo David David Deary. Comos estas
3. K Muy bien y tu?
4. D bien. Primero. Cuando eras niña uh que haces qh que haga cuando niña
5. K cuando era niña um jugaba con mis amigos y caminaba en el parque con mis abuelos. U m Yo nadaba en la pis piscina y si y tu
6. D uh Cuanto hace cuanto tiempo hace que tu ah tu tener tu primer beso cuanto tiempo hace hace doce años con qui con quien
7. K con amigo con risa.
8. K trabajo
9. D Mi trabajo?
10. K trabajas
11. D ah si yo trabajo en calender bridge richmond y trabajo es una lugar donde se sirve la comida vegetariana
12. K Trabajas jornada media
13. D Jornada media porque es necesita que yo estudie n en la universidad_
14. K que es que te aspectos que mas gusta en su tu te trabajo
15. D a veces me gusta los clientes pero otras veces los cliente ah puede muy grosero pero ah me gusta ah cocinar y y los otros employees- si
16. K Quieres ser un cocinero_
17. D No quiero porque um porque no me gusta mucho cocinar mucho si
18. K Que quieres hace que quieres ser cuando tu terminar los estudios?_
19. D Tal vez quiero ser un diplomático porque puedo viajar todos La monde el mondo mundo perdón es francés .puedo conocer las culturas nuevas
20. D en tu que a haces para limpiar?
21. K um cuando es necesario desempolvar y pasear la aspirador y yo lavo los platos y la ropa y si mmm
22. D y con que frecuencia tu limpias la casa_
23. K uh risa limpio la casa una vez a la semana
24. D en tu casa quien limpia_
25. K yo
26. D si
27. K Si
28. D y también por las diversiones en tu casa te gusta lee televisión
29. K si me gusta ver la televisión y vista bar con amigas y pero no me gusta limpiar me gusta jugar cartas no me gusta … mucho es muy aburrido quieres hacer la compras este fin de semana
30. D quería hacer la compra pero fue al supermercado ayer y compre los tomates y las comidas y todas los siento
31. K porque no me llamas vas a comprar
32. D porque no te gusta cuando yo hacer hago la compra
33. K pero quieres comprar los zapatos
34. D ah si tal vez porque cual días tu
35. K lunes si tal vez_ que te gusta comer en las casa

36. D en mi casa? todas las comidas porque en mi casa cmomimos ya con t con nos. Es mas y muy importante tener cenar con la familia y cuando nos nosotros comemos tenemos um pasta y con carne bistec y um tal vez tacos y con arroz y enchiladas verdes si
37. K Si
38. D Es todos um tus clases en la universidad cual es tu favorita esta universidad
39. K Mi favorita es la política y si
40. D y que aspecto de esta clase de política el es uh mejor que los otras clases_
41. K um la clase
42. D adios

Pair 9 one does not talk

1. El Um uh um que que clases que clases te gustaba en la escuela secundaria
2. Ella mm clases escuela secundaria me gusta ingles
3. si
4. si
5. y matemáticas
6. El Si porque
7. Ella Mm porque um gusta leer libros y novelas y especial novelas
8. El tenias muchos amigos en la escuela secundaria_
9. Ella ah si um tie tenia muchos amigos
10. Si
11. si
12. El y ah que te gustabas hacer tu tiempo libre
13. Ella ah
14. El tu tiempo libre ah te gusta
15. Ella ir a vena con mis amigos
16. El y es grande tu familia
17. No es grande es pequeno
18. El vives con tus padres
19. Ella vivo con mi familia
20. si
21. y si
22. El tienes hermanas o hermanos
23. Ella no
24. No
25. y tu
26. El Si tengo una hermana una hermana se llama briony y una hermana que tiene 18 anos
27. Ella 18 años
28. El 18 anos Te llevas bien con tus padres
29. Yep si lleva bien con mis padres uh especialmente con mi madres
30. si
31. si
32. El um um Quien te pareces en tu familia a quien
33. Ella no se con mi padre
34. Porque
35. Porque llevamos lentes
36. Si
37. Que tipo de restaurante te gusta?

38. El um me gusta restaurantes chino y y restaurantes italiano um ah mi restaurante favorito se llamo cantoman es restartuartne chinesn pero um voy a
39. Ella donde es
40. Es en Richmond
41. Tiene banquetes muy ricos pero solo voy a veces porque es caro.
42. Ella con que frecuencia tu comer?
43. El dos veces por ano y tu que clases de restaurante
44. umm
45. prefieres_
46. Ella me gusta la comida japonés
47. si
48. si tines un restaurante favorito
49. ELLA si tienes restaurante favorito es se llama blue train y es en la ciudad
50. EL es caro o barato comer?
51. Ella oh
52. El um cual es tu clase favorito a la universidad?
53. Ella mm ingles
54. si
55. si
56. Por que
57. Porque me gusta leer novelas
58. si
59. y escibir estudiandolas
60. El es tu especialidad? En al universidad ingles?
61. Si
62. Si
63. si y tu
64. el mi clase favorita es español porque me gusta me gusta prender otro idiomas y el próximo ano voy a estudiar a chino también y los dos son mis especialidades
65. mm
66. y um tienen tiene un empleo tienes tienes un empleo
67. ella no
68. el no
69. Ella y tu
70. El si tengo dos trabajo en un restaurante um um atendo atendo mesas y
71. mezclo bebidas y trabajo en videoteca sirvo a los clientes y limpiar
72. El limpio
73. Ella que que vacaciones de nino uh viajo mm viajar las vacaciones de pequeño
74. de nino
75. El um de nino uh de niño no viajaba mucho
76. ella mm
77. el en el verano iba a la playa con m con mi familia
78. \ell mm
79. el y y pero no no viajaba mucho. y tú
80. Ella si viajo mucho
81. si ¿
82. con mi familia
83. El Dónde

84. Ella Mm china
85. a si
86. y Japón
87. si
88. ella malasia
89. el si
90. ella si me gusta mucho viajo
91. el si um
92. ella mm te gusta mucho tiempo en tu casa?
93. El um si um en mi casa ah ,m me gusta leer y ver videos y jugar con mi perro
94. y y tu?
95. Ella mm te gusta leer libros
96. elv si
97. ella y me gustas computadoras y dormir mucho en mi casa
98. el si me gusta usar las computadoras también
99. Ella hay un centro de negocio en tu vecindario
100. El si el centro comercial mas cercano a mi casa es Box Hill central
101. si es bueno porque Hay muchos tiendas que vende venden comida chino
102. ella oh si
103. el y hay muchos bien restaurantes y tu
104. mm
105. cual es?
106. Ella no hay centro comercial
107. si
108. cerca mi casa o vecindario pero hay uno cafe es muy grande
109. El si
110. ella y siempre toma el cafe
111. el y hay mucho mucho parques cerca de mi casa
112. ella ah si
113. el y um es viejo yo tu vecindario
114. Ella no mm es nuevo
115. el nuevo
116. Ella porque es en universidad
117. el vives en residencia estudiantil
118. ella si vivo en residencia estudiantil
119. El hay muchas cosas en tu casa
120. Ela silencio
121. El tienen muchos muebles en tu casa
122. Ella si
123. el Ok si (to rater)
124. EL Ah bye
125. ella bye

Appendix 10

Rater team 1

Question 1		Question 2		Question 3		level
					Yes	7
			Yes	Q3.1 expands and Changes Topics?	No	6
	Yes	Q2.1 Supports the interaction with the body?	No	Q3.2 Speaks for too long or holds the floor too long?	yes	5
Q1 Are they supportive Listeners?					no	4
					yes	3
	No	Q2.2 Asks questions relevant to topic	Yes	Q3.3 Responds fluently when asked?	no	2
			No			1

Rater team 2

Question 1		Question 2		Question 3		level
					Yes	5
			Q2 yes	Q3.1 takes even turns	No	4
	yes	Q2.1 Supportive listener audibly?			yes	3
Q1 Are they mutually supportive visually?			Q2 no Q2 yes	Q3.2do they maintain topic cohesion?	no	2
	No	Q2.2 asks adequate Questions?	Q No			1

Rater team 3

Question 1		Question 2		Question 3		level
					yes	7
			yes	Q3.1 do questions/ Replies show cohesion b/n and within topics?	No	6
					yes	5
	yes	Q2.2 supportive listener?	no	Q3.2 reasonable turn length?	no	4
Q1 Supportive body language?			yes	Q3.3asks/answers within a comfortable time?	yes	3
	No	Q2.2 Asks relevant questions?			no	2
			no			1