



UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

DEPARTAMENTO DE PSICOLOGÍA Y SOCIOLOGÍA

Assessing Speaking Skills in English:
An Approach to Test and Rating Scale
Design in a University Context

Tesis Doctoral

Susan Cranfield McKay

Las Palmas de Gran Canaria

2007

SERVICIO DE INVESTIGACIONES

TERCER CICLO

REGISTRO

Nº 123

FECHA: 23/03/07

EL FUNCIONARIO



Nº 6 Curso 2007/08

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
SUBDIRECCIÓN DE TERCER CICLO Y POSTGRADO

Reunido el día de la fecha, el Tribunal nombrado por el Excmo. Sr. Magfco. de esta Universidad, y finalizada la defensa y discusión de esta tesis doctoral, los señores miembros del Tribunal, emiten la siguiente calificación global:

SOBRESALIENTE CUM LAUDE
POR UNANIMIDAD

Votos favorables:

CINCO

Las Palmas de Gran Canaria, a 5 de noviembre de 2007

La Presidenta: D^a. Leslie Bobb Wolff

El Secretario: D. Richard Clouet

El Vocal: D. Plácido Bazo Martínez

La Vocal: D^a. Patricia Arnaiz Castro

La Vocal: D^a. Laura Cruz García

La Doctoranda: D^a. Susan Isabel Granfield Mckay

**D/D^a MARIA EUGENIA CARDENAL DE LA NUEZ SECRETARIA
DEL DEPARTAMENTO DE PSICOLOGÍA Y SOCIOLOGÍA DE LA
UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA,**

CERTIFICA,

Que el Consejo de Doctores del Departamento en su sesión de fecha 11 de julio de 2007 tomó el acuerdo de dar el consentimiento para su tramitación, a la tesis doctoral titulada "*Assessing speaking skills in English: an approach to test and rating scale design in a university context*" presentada por la doctoranda D^a Susan Cranfield Mckay y dirigida por los Doctores Gina Oxbrow y Marcos Peñate Cabrera.

Y para que así conste, y a efectos de lo previsto en el Artº 73.2 del Reglamento de Estudios de Doctorado de esta Universidad, firmo la presente en Las Palmas de Gran Canaria, a 11 de julio de dos mil siete.


UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
DPTO. PSICOLOGÍA Y SOCIOLOGÍA



UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA



**Programa de Doctorado: Formación del Profesorado
Departamento de Psicología y Sociología**

Assessing Speaking Skills in English: An Approach to Test and Rating Scale Design in a University Context

**Tesis Doctoral
Autora: Susan Cranfield McKay**

Susan Cranfield



Directores: Gina Oxbrow

Marcos Peñate Cabrera

Gina Oxbrow

Marcos Peñate

Las Palmas de Gran Canaria, julio de 2007

Acknowledgements

If I were to include here the names of all those to whom I owe my gratefulness for the part they have played in my life during the creation of this dissertation, there would be as many pages as in any of the chapters it includes. I cannot express enough appreciation to all those of you who have helped and supported me through so many difficulties, moments of desperation and despair, and who believed enough in me to know that with your combined affection and firmness, this project could finally come to fruition. Mostly, you know who you are. In the continuing ebb and flow of life, some people have come and gone in the time it has taken me to get to this point. Despite their absence today, they are not, and never will be, forgotten.

I would like to offer my special and sincere thanks to the following people without whom this project would never have reached its conclusion:

My two directors, Gina Oxbrow and Marcos Peñate, for their guidance, support (moral, spiritual and statistical), hours of proof-reading, and constructive observations; but most of all for knowing when and how to apply just the right amount of pressure to get me working again, and when to keep a distance so I could cope. I have learned more from you than you imagine.

Rosario Blanco, for her eternal patience with my failure to renew library loans on time and her incredible efficiency in procuring articles from journals our library does not possess and has never heard of.

All my students from *Lengua BII* in the year 2003/04 who underwent not one, but two oral tests in order to collaborate with this study; by research standards you are a small sample; to me you will always be the greatest group of students I have ever had the pleasure of sharing a classroom with (and I have you on film!). Thank you.

The untiring rater/interviewers, Richard Clouet, Laura Cruz, Goretti García and Manuel Wood, who disinterestedly gave up their time and effort to conduct and assess a total of over one hundred oral tests. I owe you more than I can say.

Ana María García, for her exceptional translation and discourse structuring skills that brought about the creation of a coherent Spanish summary of this work.

And finally, I ask forgiveness of my children, for every time I was too busy to play pirates or watch *Scooby-Doo*, too preoccupied to listen to a problem, too irritable to go surfing on the beach, too tired to read a bed-time story. Your sacrifice and patience has been your contribution; your love, unconditional throughout.

Las Palmas de Gran Canaria, July 2007

CONTENTS

I.	INTRODUCTION	1
II.	THEORETICAL BACKGROUND	10
II.1	The Historical Development of the Testing of Second Language Speaking Skills	10
II.2	Defining the Construct	16
II.2.1	The Link between Construct Definition and Rating	25
II.2.2	Strategic Competence	26
II.2.3	Structuring Speech	28
II.2.4	Turn-Taking	29
II.2.5	Adjacency Pairs	30
II.2.6	Openings and Closings	32
II.2.7	Rules of Speaking and Pragmatic Appropriacy	32
II.2.8	Doing/Being Things with Words	33
II.2.9	Interactional Competence	35
II.3	Evaluating Second Language Speaking Tests: Constructing a Validity Argument	35
II.3.1	From Testing Approaches to Curriculum Implications	38
II.3.2	Validity	40
II.3.3	A Validity Model	44
II.3.4	Alternatives to the Individual Proficiency Interview: Testing Speaking in Pairs or Groups	46
II.4	Test specifications and validity	51
II.4.1	The writing of test specifications	54
II.5	Tasks for speaking tests	56
II.5.1	Test Task Characteristics	59
II.5.2	Task Difficulty in Pedagogy and Second Language Acquisition	62
II.5.3	Types of Task in Speaking Tests	70
II.6	Rating scales	72
II.6.1	Approaches to Rating Scale Design	76
II.6.2	Intuitive Rating Scales	77
II.6.3	Rating Scale Terminology	79
II.6.4	Data-Based Scale Development	82
II.6.5	Empirically-Derived, Binary Choice, Boundary Definition Scales (EBBs)	82
II.6.6	Scaling Descriptors	84
II.7	Raters and Rater Training	86
II.7.1	Rater Reliability	87
II.7.2	Rater Training	88

II.7.3	Interlocutor Training	91
II.8	Test Administration	93
II.8.1	Administration	93
II.8.2	Environment	93
II.8.3	Test Accommodations	94
II.9	Conclusions	95
III.	RESEARCH METHODOLOGY	96
III.1.	Research questions	100
III.1.1	Test Format	102
III.1.2	Scoring Procedures and Rating Scales	103
III.1.3	Self-Assessment	104
III.2	Participating Subjects	106
III.2.1	Learning Context	106
III.2.2	The Global Context	107
III.2.3	The Micro-Context	109
III.2.4	The Learners	110
III.2.5	The Influence of Translation	112
III.2.6	Teacher and Learner Expectations	114
III.3	Testing Procedure	117
III.3.1	Student and Examiner Test Preparation	118
III.3.2	Data Collection	120
III.3.3	Introductory Phase	125
III.3.4	Materials Packs and Test Tasks	125
III.3.5	The Test Environment	127
III.4	Rating scales	128
III.4.1	Defining Features of Speaking for Assessment	128
III.4.2	ARELS Marking Key for the Higher Certificate Examination in Spoken English and Comprehension	128
III.4.3	Trinity Grade Examinations in Spoken English for Speakers of Other Languages	131
III.4.4	University of Cambridge ESOL Examinations	135
III.4.5	Rating Scales for <i>Lengua BII</i> Speaking Tests	137
III.4.6	Designing the Rating Scale	138
III.5	Questionnaires	140
III.5.1	'Individual Oral Proficiency Interview': Student Perspective	141
III.5.2	'Individual Oral Proficiency Interview': Interviewer Perspective	144
III.5.3	'Group Speaking Test': Student Perspective	149
III.5.4	'Group Speaking Test': Interviewer Perspective	153

III.6	Mark sheets and Student Self-assessment Sheets	158
IV.	RESULTS	159
IV.1	Individual Oral Proficiency Interview	160
IV.1.1	Scores Obtained	162
IV.1.2	Data from Questionnaire 1 (Student)	174
IV.1.3	Data from Questionnaire 2 (Interviewer)	194
IV.2	Group Speaking Test	195
IV.2.1	Scores Obtained	198
IV.2.2	Data from Questionnaire 3 (Student)	207
IV.2.3	Data from Questionnaire 4 (Interviewer)	227
IV.3	Discussion	229
IV.3.1	Test Format	229
IV.3.2	Scoring	238
IV.3.3	Self-Assessment	244
IV.3.4	Empirical Evidence Regarding Self-Assessment	249
V.	CONCLUSIONS	253
V.1	Summary of Research Results	253
V.1.1	Test Format	253
V.1.2	Rating Scales and Scoring	256
V.1.3	Self-Assessment	259
V.2	Practical Implications of the Study	262
V.3	Contributions and Limitations of our Study	265
V.4	Possible Areas for Future Research	267
	BIBLIOGRAPHY	270
	APPENDICES	
	RESUMEN DE LA TESIS EN LENGUA CASTELLANA	

I. INTRODUCTION

Assessment is a key feature and frequently the culminating point of the majority of foreign language teaching programmes, therefore it seems obvious that its analysis and validation should constitute an important and interesting area in the field of Second Language Acquisition research. As test users and designers, we need to be able to justify the implementation of tests and the validity of their scores since we have a responsibility in responding both to the challenges of defining language constructs and to developing validity arguments that we can apply to testing practice. The tests we use need to be appropriate to the context in which they are used and provide scores that are as accurate a reflection as possible of learners' linguistic performance, underlying language ability, and of the principles behind the approach and content of the teaching and learning programme.

It is precisely this concern with making true and meaningful statements about performance and underlying ability that brings into play the design of rating scales as a fundamental issue. Most of us, in our capacity as teachers and examiners, spend time writing tests for the programmes we have taught, focusing our attention on the tasks and topics we choose, and the questions we require our students to answer. The rating scale is rarely, if ever, a cause for concern; it already exists, it has been presented to us 'ready-made'; we place our students on a scale of 0 – 10 according to judgements that are necessarily unstable because they do not correspond to a statement of criteria or to a definition of construct that describe the attribute we wish to measure. This almost certainly leads us to compare student performance so that it is the performances themselves that guide the internalised scoring criteria applied during any one testing session. If we approach assessment in this way, we can see that, in fact, the test score can only

give information about the single instance of performance on the test and cannot be generalised to underlying ability because the rating scale does not contain any fixed definitions of what it intends to measure (construct definition) or show to what extent these features have been demonstrated through performance.

In the case of speaking tests, this failing is further accentuated through the nature of the skill we wish to measure. Little is known about the rapid, complex cognitive processing which takes place in speech production and this necessarily limits our ability to describe the elements that actually compose oral competence. Added to this difficulty is the ephemeral essence of spoken language; unless it is recorded, it remains only as an idea or memory in the mind of the speaker and of the listener(s). Unlike the written word, which remains static and unchanged, if we were asked to repeat a spontaneous utterance of more than fifteen or twenty words, our overriding tendency to focus on the meaning of what is said rather than on the form would make it unlikely that either listener or speaker would be able to reproduce, word for word, what has just occurred.

In order to be able to measure speaking competence then, it is first necessary to attempt to describe just what speaking involves, that is, to provide a construct definition so as to be able to develop meaningful rating scales, and tests that will allow us to generalise beyond an instance of performance to underlying language ability. In doing so, we will at the same time be constructing a validity argument that supports the link between a test score and what we claim it means. We should also aim to define constructs for speaking tests in ways that are meaningful for the learners for whom the test is designed.

This definition should attempt to describe related, but distinct, features or components that make up the construct of speaking founded on certain theories of

competence. Two of the major bases for this description are *strategic competence* and *interactional competence*. Strategic competence (Canale and Swaine, 1980; Canale, 1983; Bachman, 1990; Bachman and Palmer, 1996) refers to the cognitive capacity to manage communication, but in order to be able to include it in a construct definition we also need to consider what observable features of speech would provide us with evidence of its use. These would need to be described in the rating scale so that raters and test-users were aware of them as distinct, measurable features of the construct. There is still no consensus on the definition of interactional competence; some researchers, generally those concerned with definitions of communicative competence, claim that, as part of the latter, it is ability within an individual rather than a feature of interaction (e.g. Bachman, 1990).¹

Currently, researchers in educational and social science fields are moving away from theories that embody all encompassing representations of constructs that are meant to hold true in infinite situations, arguing that such theories cannot provide rich and meaningful representations of socially-mediated constructs. In line with this trend, Chaloub-Deville (2003) proposes that we expand the concept of interactional competence to include the context in which the interaction takes place as part of the construct definition. This approach assumes that although the essential abilities of language users are internally stable or unvarying, unless we develop a theory that takes contexts into account, we will not have enough evidence to make generalised inferences about abilities and performances across contexts. She claims that we need to better understand the complex interplay of

¹ This contrasts with the view of McNamara (1997:447) who sees it as a social/behavioural construct which refers to the way in which speech is co-constructed by participants in the interaction.

both the stable and the more variable interactional systems in order to be able to make informed judgements about underlying competence.

It can be seen then, that construct definition is a complex and on-going task which is inexorably linked to validation theories. Even though we may not yet be close to achieving a comprehensive construct definition for speaking, it is essential to be aware of the implications and importance of attempting to include it in our testing procedures in order that our tests, although not infallible, might be based on a validity argument that is supported by theories of language description, second language acquisition and of the measurement of language skills.

As instruments for the collection of evidence of oral proficiency, our speaking tests should permit data to be obtained in a systematic way by means of tasks or other elicitation techniques that can be replicated with different candidates and in different testing sessions. The summary of the evidence (the score) should provide us with information about the construct as we have defined it, and should also allow us to make inferences about the learner's performance in non-test contexts. This also presents us with the challenge of designing adequate test tasks and materials that will permit us to collect a speech sample with the appropriate characteristics and size required for assessment in our given situation.

It can be seen, then, that it is of paramount importance to set our testing procedures within a theoretical framework if we are to respond to the need to be systematic, coherent and purposeful in our task of developing language tests. This issue has long been a personal concern due to the life-changing consequences that assessment has on its primary users: the students who take our tests. In the present study we are concerned with students who are in the second year of an undergraduate degree programme in 'Translation and Interpreting' at the

University of Las Palmas de Gran Canaria in Spain and who are receiving instruction in the subject *Lengua BII* in English as a foreign language. Our marks and scores directly affect them in areas such as being able to continue with their degree programme, the continuity of government grants, access to parallel or higher level study programmes such as Master's degrees or doctorate programmes, and participation in national and European student exchange and work experience programmes. They are also highly likely to affect them in the emotional areas of personal development, concept of self-worth and their general outlook on life. The overall mark for their degree will probably have a bearing on career choices and their incorporation into the world of work. It is in this intertwining of the human and social aspects of testing procedures and the meaning and interpretation of measurement scales, that our interest and sense of responsibility in assessment lie.

Our aim in this study is to explore in greater depth some aspects of the issues mentioned above in order to be able to propose possible changes to our current testing procedure based on empirical evidence. Our research questions will centre on three main areas of testing speaking skills: **(i) test format, (ii) scoring and rating scales and (iii) the role of self-assessment** in teaching and learning.

Within the first area of test format, we will focus on contrasting the use of an **individual oral interview** involving one candidate and an interviewer (with an independent rater for control purposes), with a **group oral test** where students are examined in groups of three and will interact with each other during the test. In the latter, we have an interlocutor who is responsible for initiating and managing the test, and an objective rater who is not involved in the interaction and is responsible only for scoring. Our intention here will be to try to discover whether the group test produces less anxiety in students because they are accompanied and supported

by their peers, and also to see whether it is easier to rate the interaction produced in such a test from the point of view of an objective observer (rater), while the test is managed and guided by an interlocutor. According to Fulcher (2004: 186), an unpublished UCLES² study showed that in the paired test format candidate turns were increased and the amount of talking time attributable to the interlocutor was substantially reduced in contrast to an individual oral interview format. We presume that this will also be the case for a group speaking test, and that a wider range of language and language functions can be elicited in this situation. The negotiation of meaning that takes place in group interaction and which may promote second language acquisition is more likely to be produced in a group speaking test than in a one-to-one interview situation. Swain (2001: 274) states that “dialogues construct cognitive and strategic processes which in turn construct student performances, information which may be invaluable in validating inferences drawn from test scores.” In other words, the pair or group format may generate language performances that allow us to assess much more complex constructs than a traditional one-to-one interview.

We will also be concerned with our second area of interest, **scoring**, and the problem that lies in defining constructs which recognise the co-construction of discourse and meaning where the speech sample to be assessed has been produced through interaction, and possibly peer collaboration and support. In an attempt to take this into account, we will propose a **rating scale** that is based on a construct definition which considers interaction as one of its components while, at the same time, trying to retain features of description that allow an individual performance to be scored within a group situation. In the implementation of this scale, we will be

² University of Cambridge Local Examinations Syndicate. This name has since changed to ‘Cambridge ESOL’.

interested to see how raters apply the descriptive criteria in carrying out their assessments and to what extent they either remain objective or internalise the scores in order to use individual interpretations of them to score student performance.

Our third area of interest lies in how useful the measurements our assessment and scoring procedures are to our students in the context of the current marking system (the universal 0 – 10 scale) and whether we might be able to have an impact on learning and motivation by allowing them to participate in a process of **self-assessment** using the same descriptive scales themselves as we provide for the raters in the speaking test. Here, we will attempt to discover whether students feel that, through the provision of a description of the way in which raters attempt to assess their speaking skills, self-assessment can be a useful tool in language learning and improving language proficiency. We will also address whether they feel their self-assessments are objective and accurate enough to be included in their final mark for the subject *Lengua BII*. Additionally, we elicit teacher/examiner opinions on these two aspects of student self-assessment to see whether they coincide or differ, with the ultimate goal of comparing self-awarded marks with the scores received from the rater during the testing sessions. Our aim here will be to discover whether there is a correlation between the scores that suggests that there is a justification for the introduction of self-assessment in our current study programme and, if there is, what preliminary steps may need to be taken in order to begin to implement it in future teaching and assessment programmes.

Having given a brief outline of our research concerns and questions in this introductory chapter, which are fully described in Chapter III, in the following chapter, 'II. Theoretical Background', we will provide a review of the relevant

research literature which has informed our study. In order to provide a context for our present situation with regard to testing and assessment, we will begin with a short consideration of some of the historical concerns of testing second language speaking that have given rise to current practices (Section II.1). We will then proceed to look at the complex issue of construct definition (Section II.2) and how this is related to rating scale design, with particular emphasis on the different facets of speaking that we may wish to take account of, and therefore include, in our scale descriptors. The relationship between the considerations of construct definition and rating scale design naturally gives rise to a discussion of the construction of validity arguments and this will be explored in Sections II.3 and II.4, which focus on the way in which tests can be evaluated and their validity thus supported through a theoretical framework. This is followed by a brief discussion of test specifications and how these relate to the validity argument.

Section II.5 will explore some of the characteristics of test-type tasks and the ways in which these affect the sample of language collected during a test. Here, we focus on such issues as kinds of task, the language structures and functions they elicit, and the factors that influence theories of difficulty in test task design. A further section (II.6) explores some of the different approaches which may form the basis for the design of a rating scale, followed by a consideration of rater characteristics and rater training (Section II.7) and their effects on speaking test implementation. Finally, in Section II.8, we look very briefly at some of the aspects that are external to the testing procedure itself, such as the environmental concerns such as the furniture positioning, temperature and lighting conditions, which we need to take into account when administering tests due to the possible impact they may have on candidate performance.

Chapter III will set out our research objectives, design and method. The first section will enumerate and justify the research questions we shall attempt to answer in the current study, followed by a description of the learners whose speaking skills in English we test, and their language learning context in Sections III.2 and III.3. Sections III.4 and III.5 provide an account of our experimental design whose aim, as we have seen above, is to compare the implementation of two different types of speaking test (the 'Individual Oral Interview' and the 'Group Speaking Test') and two different types of rating scale (holistic and analytic). Specific details of the tests that our students took in this study are included here (Section III.5), along with a reasoned description of our own rating scale which was designed for use in this project and the rationale underlying its implementation. This chapter concludes with a discussion of the rationale behind the questionnaires that were used for data collection and the considerations that were taken into account in their design.

In Chapter IV we present our results in the form of interpreted graphs, indicating where statistical significance is found in relation to test scores and to item responses on the questionnaires that were completed by the students and examiners participating in this study. These results are then discussed with a view to confirming or discounting our original research questions as set out in Chapter III and to ascertain whether we have been able to answer them. Finally, in Chapter V, as well as recognising the limitations of our study, we will try to draw some conclusions from our findings in the main areas addressed in the present research project, that is **test format, scoring and self-assessment**, and show how these may provide sufficient evidence for implementing some changes to our current teaching and learning syllabus and assessment procedures.

II. THEORETICAL BACKGROUND

In order to provide a sound empirical basis for the research questions guiding our own investigation (which we will address in Chapter 3), in the following chapter we will survey and critically appraise the relevant literature dealing with the most important concerns in the testing of spoken language in foreign language contexts to date, with particular emphasis on the aspects of *construct definition* and *validity* which are two of the major, and most complex, issues in the area of oral testing and which are at the forefront of current debate and investigation and are therefore worthy of special attention and analysis. We will also address other relevant areas of interest such as *task difficulty* and *rating scales*. However, in order to gain a more informed insight into the current situation, we shall begin by exploring some of the historical factors that have given rise to recent concerns in language testing.

II.1 THE HISTORICAL DEVELOPMENT OF THE TESTING OF SECOND LANGUAGE SPEAKING SKILLS

Language testing is one of the youngest fields of research and practice in the discipline of Applied Linguistics, with the assessment of spoken language proficiency only becoming a focus of interest during the Second World War (Fulcher, 2003: 1). In order to better understand the development of modern speaking tests, it is important to first outline the close connection between the development of speaking tests and political and/or military needs, since these have had a deep impact on the format and scoring procedures of many modern speaking tests.

Before 1939, the overwhelming concern in oral testing was with achieving reliable scores, ensuring that tests were consistent over a number of administrations,

and relatively little attention was paid to test *validity* (the concern that the test actually measures what it is intended to measure). Today, reliability (providing consistent comparable scores over a number of test administrations using different raters) along with practicality (viability in terms of financial cost in development and administration, time, and the number of personnel required to administer the test) are still important issues. They are the major drives behind research into semi-direct tests of speaking carried out in a language laboratory where candidates respond to recorded instructions and prompts, with their speech being recorded and rated from the recording.

The first true speaking test used in the United States was the ‘College Board’s English Competence Examination’ which was introduced in 1930 for overseas students applying to study in US universities and colleges. Apart from test scores, the examiner was also asked to record whether the candidate was shy, showing an early sign of interest in the individual differences that may be a threat to the valid interpretation of test scores. The speaking test was scored according to the criteria of:

- fluency
- responsiveness
- rapidity
- articulation
- enunciation
- command of construction
- use of connectives
- vocabulary and idiom

These aspects were graded on a three-point scale of ‘proficient’, ‘satisfactory’ and ‘unsatisfactory’. The design of the grading procedure for this test therefore reflects an

interest in trying to define those aspects of the speaking skill that are important in this early attempt at defining the *construct* by means of enumerating the key features for assessment and considering at least one factor that may affect scores which is not related to the construct (construct irrelevant variance, in this case shyness).

During the Second World War, there was a sudden, pressing need for military personnel to be able to speak and understand (as opposed to read and write) foreign languages and this led to the introduction of language instruction programmes in the US Army Specialized Training Program (ASTP) that focused on speaking, with an aim to “impart to the trainee a command of the colloquial spoken form of a language and to give the trainee a sound knowledge of the area in which the language is used.”¹ This shift in pedagogy towards the teaching of speaking skills implied a leap from the assessing of grammatical knowledge in traditional written language assessment to the ability to perform in communicative contexts and the ASTP was the precursor to the Foreign Service Institute (FSI) ‘Oral Proficiency Test’ (OPT) which has had a tremendous influence and bearing on the development of all oral tests which have followed it.

The FSI test was also the first one of its kind to recognise the importance of inter-rater reliability by attempting to train examiners to interpret the meaning of assessment criteria established by the examining board over a period of time and through practice (see Appendix 1). The FSI was aware that some kind of standardisation was necessary if scores were to be consistent, fair and meaningful. However, the development in 1952 of the FSI rating scales for the selection of Civil Service personnel during the Cold War was not based on any kind of definition of the

¹ Angiolillo, 1947: 32, with reference to the US Army Specialized Training Program (ASTP).

speaking construct. There is no evidence available concerning the choice of six bands as opposed to any other number, nor is there any explanation as to why Band 4 was the minimum requirement for diplomatic personnel. No attempt was made to define separate components of language proficiency as in the first example from 1930, resulting in an intuitive 6-band holistic rating scale with weak descriptors only for the lowest and highest bands (0 = no functional ability; 6 = equivalent to an educated native speaker).

In 1958, the FSI testing unit modified the 1952 rating procedure by adding a checklist of five factors, each on a 6-point scale: (1) accent; (2) comprehension; (3) fluency; (4) grammar; (5) vocabulary (Adams, 1980). This was another early step towards developing multiple trait rating, (even though the components were to be interpreted with a single holistic score). Although it was claimed that this rating procedure was a highly accurate predictor of a person's speaking ability,² it was also acknowledged that a limitation of the scale was that it did not measure "effective communication" (Sollenberger, 1978: 7-8). So it can be seen that from the earliest days in the design of modern rating scales for speaking tests, both the roles of linguistic competence and communicative ability were already issues of concern for test developers.

These early developments in the testing of speaking skills generated interest in the effectiveness of holistic versus multiple trait rating and in the distinction between linguistic and communicative criteria for rating and reliability. The focus was almost

² Confidence in the new testing procedures developed by the FSI was so high that in the 1960s they were adopted (and adapted) by the US Defense Language Institute, CIA and Peace Corps. These diverse agencies came together and produced a standardised version of the test which is still in use today.

exclusively on the design of bands or rating scales and their descriptors, or the rubrics for the test (task design and the role played by tasks in tests was a concern that was to arise at a later date). Until recently, the actual components which are chosen for criteria in a rating scale, and the importance of each of those components, was an area of research in which little progress had been made, apart from the contribution by Adams (1980).

The high status held by the FSI speaking test has meant that, despite being poorly defined, its rating scale and some other test concepts (such as that of the 'Oral Proficiency Interview' which we will discuss later in Section II.3.2, , have provided a model for, and been adopted into, many other tests and rating scales (e.g. the American Council on the Teaching of Foreign Languages (ACTFL) and the Inter-agency Language Roundtable (ILR)). However, these are still problematic in that the descriptors produced by the scale-writers come from intuitive judgements about how language competence develops, and how this competence is used in a test performance (this will be discussed below in Section II.6 on rating scales). They also contain a combination of linguistic and non-linguistic criteria, along with undefined degrees of accuracy to be achieved for each level in each scale, mixed with references to the types of task or the situation in which the student would be expected to operate outside the test situation. If the purpose of the speaking test is to provide a sample of language from which performance on a wider number of non-test situations can be predicted, then it is reasonable to expect the rating scale to contain descriptions of those skills or abilities which underlie successful performance across a range of situations, or for the test itself to specify tasks or situations which could be demonstrated to allow generalisation to other tasks or situations. Yet this mixture of linguistic and non-

linguistic criteria has been viewed as a confusion which makes validation studies very difficult (Bachman & Savignon, 1986; Matthews, 1990).

As we have seen above, the descriptors in the FSI absolute rating scales (see Appendix 2) are notably vague. The scores they represent on the scale range from 0 to 5, with the scales assuming linear development from zero to “perfect native speaker” speech (5). The concept of NS proficiency has affected most rating scales since the FSI, even though, arguably, there is no such thing as a ‘perfect native speaker’, since all native speakers have a different level of competence (Chipere, 2001). The bands show an increase in the accuracy of the language used: e.g. the ‘Grammar’ scale hinges on the progression of modifiers from *constant* → *frequent* → *occasional* → *few* errors. Nowhere is it suggested what kinds of errors typify each band, and there is no indication that the scale is linked to any consideration of a sequence in which learners acquire certain grammatical forms. The reference to “major patterns” of grammar in Band 2 of the ‘Grammar’ scale seems to suggest that the authors had a notion of the grammatical forms that they expected to occur in earlier and later bands, but these are not listed or described.

In the ‘Fluency’ rating scale, a similar situation occurs: Band 1 states that “conversation is virtually impossible”, while in Band 5 conversation is “as effortless and smooth as a native speaker’s”. The concepts that dominate the bands between the extremes are those of speed of delivery, hesitation and “unevenness”, modified by ‘very’, ‘frequently’ and ‘occasionally’ in Bands 2 to 4. These concepts, however, are not defined. Speed of delivery may vary considerably among non-native and native speakers of a language, and “unevenness” seems to be related to “rephrasing and

groping for words”, something which is frequent in native speaker speech (Fulcher, 1987). The nature of hesitation and its causes have not, to date, been investigated.

It can therefore be seen that in early tests of speaking skills, the ‘construct’ – what is being measured – is defined within the rating scale. The bands or levels of the rating scale are intended to describe levels of language proficiency that ‘exist’ in the real world, whether these are expressed in terms of language elements or functional ability. However, the detailed consideration of what the speaking skill actually involves did not come to be seriously questioned until the 1970s.

II.2 DEFINING THE CONSTRUCT

The exploration of the second language construct is of vital importance in terms of its impact on many aspects of test design, validity concerns and the formulation of theories of measurement of linguistic competence. Douglas (2000: 25) states that, “language knowledge is multicomponential; however, what is extremely unclear is precisely what those components may be and how they interact in actual language use.”

The question of how we might define the construct of speaking is a major concern for all those involved in the testing of speaking skills. Since it is not possible to observe *ability* directly, a measurement of it can only be based on observations of performance, in other words, what students do in the classroom or how they perform in a test situation. Thus ‘ability’ here is defined in terms of the observable behaviours that are of interest in a particular learning or testing context.

A construct may be said to be a concept that is deliberately defined for a special scientific purpose. Kerlinger & Lee (2000: 40) argue that a construct differs from a

concept in two important ways: firstly, it is defined in a way that can be observed and measured, and secondly the relationship between different constructs constitutes a theory. There are currently several different models of the L2 construct to which language test developers may attend which we will discuss below, and important research into extending these models has been undertaken by authors such as Chapelle (1999) and Chaloub-Deville (2003).

Early interest in attempting to define the speaking construct was embodied in the 'trait-theory' approach to construct validity (Lado, 1961). He argued that testing the ability to speak a foreign language was the least developed in the language testing field due to "a clear lack of understanding of what constitutes speaking ability or oral production" (Lado, 1961: 231). Lado's concern was to make a speaking test purely a language test, and to avoid non-linguistic variables such as 'talkativeness' or 'introversion' that might be confused with the construct.

A later model was that of 'Communicative Language Ability' (Bachman, 1990; Bachman and Palmer, 1996), which has recently found recognition in the general measurement literature (Bachman, 2002a). This model recognizes that the ability to use language communicatively involves both knowledge of, and competence in, the language and the capacity for implementing this competence. The framework is consistent with previous models such as that of Canale and Swain's model of communicative competence (1980), and Canale's later version (1983), where communicative competence was extended to include grammatical competence (knowledge of the rules of grammar), sociolinguistic competence (knowledge of the rules of use and of discourse), strategic competence (knowledge of verbal and non-verbal communication strategies) and discourse competence (verbal and nonverbal

communication strategies that may compensate for breakdowns in communication due to performance variables or insufficient competence). However, Bachman's model is innovative in its attempt to represent the processes by which the various components interact with each other and also with the context in which language use occurs:

Communicative language ability consists of language competence, strategic competence, and psychophysiological mechanisms. Language competence includes organisational competence, which consists of grammatical and textual competence, and pragmatic competence, which consists of illocutionary and sociolinguistic competence. Strategic competence is seen as performing assessment, planning and execution functions in determining the most effective means of achieving a communicative goal. Psycho-physiological mechanisms involved in language use characterise the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented. (Bachman, 1990, quoted in Weir, 1990 :8)

While this is probably not the definitive solution to the complex problem of defining language proficiency (and Bachman does not propose it as such), for some years it has been the most operational one available. In order to create a theoretical basis for language testing (as opposed to description), Bachman subsequently incorporated categories of test method facets that may have an influence on language performance. The following general structure of the model is outlined by Skehan (1991: 9) and shown in Table 1 below):

<ul style="list-style-type: none"> • Trait Factors : <i>Competences</i> <i>Language Competence</i> organisation competence grammatical textual pragmatic competence illocutionary sociolinguistic <i>Strategic Competence</i> assessment planning execution • Skill Factors psycho-physiological mechanisms mode (receptive/productive) channel (oral/aural ; visual) • Method Factors language use situation amount of context distribution of information type of information response mode
--

Table 1

All these factors combine to make a framework which extends beyond previous models of communicative competence and performance to include the essential

component of context, thus making the model applicable to real-life language use. As Skehan (1991: 9) clearly points out:

The Bachman model contains within itself, as it were, a concern with the competence-performance relationship. Clearly, the basic *competences* are concerned with generalised abilities. However, the *skill* and *method* factors connect up with real language performance in a way that is integral to the whole model. So it is part of the model to make statements about actual performance as well as underlying abilities. In this respect, it is striking to see the inclusion of *method* issues, that is, ways in which the format of the test may intrude and cloud the measurement that is being made.

The fact that this model has built into it the consideration of the effects of test design on the conclusions reached about learner language in test situations makes it a more valid and usable framework. It implies that testing is not infallible and that a test result may sometimes tell us more about the testing method or format than about underlying language ability. Therefore, in using or designing tests we should at the very least be aware of the effects that a testing method may have on the results we obtain, and if these seem to be systematic, make an effort to avoid them.

Bachman's CLA model, however, continues to give an essentially cognitive/psycholinguistic representation of second language use in context and more recently there has been a tendency to admit that sometimes, when contextual factors clearly impact upon discourse and test score, it may be more appropriate to include contextual factors in the construct definition (Chapelle, 1999). In these cases, a full description of the target language use domain forms the basis for test design, as the

inferences made from the scores are not of *speaking ability*, but to ‘speaking ability in one or another context’ (e.g. in tests for air-traffic controllers). In other words, the test purpose drives the definition of the construct, its range, and its generalisability.

Bachman has also revised and developed his original model (2002a; 2002b), promoting both a construct-based and task-based approach to test design which gives tasks (or contexts) equal prominence in test design and interpretation. However, he continues to distinguish between the abilities targeted and the context in which they are observed, showing interaction to be individual-focused and largely a representation of a cognitive, or ‘within language-user’ construct. Diverging from this position is that advocated by proponents of *interactional competence*, who view the language-use situation primarily as a social event in which ability, language users, and context are intertwined and inseparable. Chaloub-Deville (2003) describes her interactional competence model as ‘ability – in language user – in context’. The ability components interact with situational facets in order to change them as well as to be changed by them. The situational aspects of the context the language user attends to dynamically influence the ability features activated, and vice versa. Thus, ability and context features are intricately connected.

We can see here that although Bachman’s CLA model and Chaloub-Deville’s interactional competence model both take into consideration a learner’s ability as well as contextual features, they still represent fundamentally different perspectives on language use. In contrast to Chaloub-Deville’s interdependent interactional model of ‘ability – in language user – in context’, Bachman’s cognitively based CLA model contends that the separation of construct and task/context is both feasible and desirable. Chaloub-Deville’s interactional competence model portrays a link between

'ability – in language user' and 'the context' considering them to be two important, yet separate interacting entities, with some interactional competence researchers even treating them as a single interacting structure. According to Chaloub-Deville (2003: 373):

The social interactional perspective compels language testers to address two fundamental challenges:

- amending the construct of **individual** ability to accommodate the notion that language use in a communicative event reflects dynamic discourse, which is co-constructed among participants; and
- the notion that language use is **local** and the conundrum of reconciling that with the need for assessments to yield scores that generalise across contextual boundaries.

Evaluating the performance of test-takers according to the social interactionist perspective offers a serious challenge to language testers in terms of the generalisability of scores. If internal attributes of ability are inextricably intertwined with the specifics of a given situation or context, then any inferences about ability and performance in other contexts are questionable since the idea of transfer of conceptual schemes is at the heart of the issue of generalisability. It implies that learners are capable of applying knowledge and skills in situations other than those in which they were developed. The issue of how to document the connections learners make in order to transfer relevant knowledge and skills is still in its very early days of research and whether or not it would be possible to teach students strategies for knowledge transfer

is another important research question which would require some kind of taxonomy of how we perceive this to happen in individuals, based on our own experience.

The questions of how individuals connect situations and language and of how this knowledge can be applied across contexts call for further research to explore the external interactional contexts in which internal knowledge and processes are accessed adequately in similar ways in order to allow a degree of generalisation across these contexts. Chaloub Deville (2003: 377) states that discussions of the notions of familiarity and practice with a given context are also required:

...the language user has a set of preferred abilities that are typically activated in contexts with particular features. The more familiar the language user is with these ability structures-contextual features, the more efficient and fluid learners become at activating them: combining and recombining knowledge structures as needed to engage in a given situation. It is likely that language users at different proficiency levels call upon different or differentially developed abilities. Furthermore, their learning experiences would help determine the associated resources they are likely to engage in a given context.

This approach assumes that the most essential abilities of language users are internally stable or unvarying. But, by accounting for the more stable aspects of the construct, is there sufficient evidence for inferences about abilities and performances across contexts? Transfer arguments based on learning history and language-use practice lend credence to the cognitive position that the generalisability of abilities is feasible. However, an exclusive focus on stable systems that may generalise is not

sufficient; we need to gain a deeper understanding of the complex interplay of both the stable and the more variable interactional systems.

Currently, researchers in educational and social science fields are moving away from theories that promote all-encompassing representations of cognitively-based constructs that are meant to hold true in infinite situations, arguing that such theories cannot provide rich and meaningful representations of socially-mediated constructs. Also, generic cognitive theories do not afford sufficient direction or provide appropriate hypotheses to guide specifically situated investigations. Social interactional investigations, on the other hand, would be able to consider focused hypotheses of the complex interaction of linguistic and non-linguistic knowledge, as well as cognitive, affective and conative (the instinct or desire to act purposefully) attributes engaged in particular situations. Chaloub-Deville (2003: 381) concludes that she believes context to be paramount in a theory of the speaking construct and that a theory of context is now essential for our understanding of speaking abilities and their measurement.

What is clear is that different construct definitions are appropriate for different test purposes, and it may be more helpful to evaluate the usefulness of a construct definition not through its correspondence to psychological reality, but in its usefulness in allowing inferences to be made from test scores, and its value in helping us to construct a validity argument that supports the link between a test score and what we claim it means. Fulcher (2003: 20) states that:

From what we know about speaking and testing second language speaking from research, we can ... 'pick and mix' to make a construct.

All we have to do is provide a rationale and empirical evidence to support the ‘mix’ we end up with in terms of test purpose.

It is also important to define the constructs for speaking tests in ways that are relevant and meaningful for the learners themselves, or more abstractly, the test-taking population for whom a test is designed. Therefore, it seems desirable that constructs should be driven by test purpose, taking into account the needs and motivations of those who will take the test, and should also be sensitive to the requirements of score users (for example, in academic and professional contexts). In the following section we shall consider the way in which rating scales implicitly define the construct, although this attempt is often not based on a theory derived from empirical findings.

II.2.1 The Link between Construct Definition and Rating

In our survey of the history of the testing of speaking skills above, we have seen how the attention of raters has, in general, consistently been directed at the accuracy of structure and vocabulary as one component of assessment, and the quality and speed of delivery (fluency) as a separate component. This can be considered an attempt at construct definition: the operational definition of two related but distinct components that make up the construct of speaking. As we will see later in the descriptors devised for our own rating scales, other factors may be involved and hence included, but the important thing to notice is that the rating scale descriptors for any test implicitly define the speaking construct, or the test-designer’s belief as to what that constitutes. For example, the attribution of what constitutes high and low gravity errors will give an indication of the test-designer’s attitude towards language acquisition and test-taker performance. Errors in word order and omission are almost

always considered high gravity, while low gravity errors are usually those made with the tense system, since conversation is nearly always comprehensible despite errors of this kind. However, test designers should also bear in mind that while learners with a limited command of a language usually avoid complex sentence structures, particularly relative clauses, more experimentation occurs as they progress. If, as teachers, we wish to encourage risk-taking as an effective learning strategy, it needs to be taken into account in test scores. As we so often tell our learners, systematic errors are a sign of learning so they should not always be treated as negative evidence and penalised in speaking tests. What follows is a consideration of those elements or characteristics of the speaking skill to be included in our construct definition which will be shown in the following chapter to inform our test design.

II.2.2 Strategic Competence

Canale and Swaine (1980) define the component of strategic competence as the speaker's ability to cope when there is difficulty in communicating because of a deficiency in grammatical or sociolinguistic competence. Bachman (1990: 107) later extended their definition to include:

the capacity that relates language competence, or knowledge of language, to the language user's knowledge structures and the features of the context in which communication takes place. Strategic competence performs assessment, planning and execution functions in determining the most effective means of achieving a communicative goal.

In other words, strategic competence is a cognitive capacity used to manage communication. However, if we intend to include strategic competence in our construct definition, we need to be able to define what it is that we would observe in speech that would provide evidence of such strategy use. Fulcher (2003: 31-34) divides these strategies into two categories: strategies of *achievement* and of *avoidance*. Learners use achievement strategies when their knowledge of grammar and vocabulary is insufficient to communicate what they want to express. Included in achievement strategies are **over-generalisation**, or morphological creativity (inappropriate transfer of knowledge of the language system, e.g. *buyed*); **approximation** (replacement of an unknown word with one that is more general e.g. *went* for *drove*); **paraphrase** (finding an alternative way to express a structure or lexical item which is unknown in the L2); **word coinage** (invention of a new word for an unknown word, e.g. *air ball* for *balloon*); **restructuring** (reformulating the grammatical structure of an utterance); **cooperative strategies** (getting help from the listener); **code-switching** (changing from one language to another); **non-linguistic strategies** (gestures, mime, pointing etc.). Avoidance (or reduction) strategies, unlike achievement strategies, are not creative and are used by learners who try to avoid having to use language which they do not know by only communicating messages which they already have the linguistic capacity to convey. They are manifested either in the abandoning of utterances, the overuse of delexicalised words, such as *thing*, or the absence of appropriate grammatical structures. The latter is difficult to detect since the absence of a part of the language system does not necessarily mean that the learner is actually avoiding it.

Whether or not we wish to test the use of communication strategies will depend on the purpose of the test and whether we are interested in the process of producing speech as well as the product. The question also arises as to whether it is actually possible to test strategy use: how can we tell for certain if a learner is employing a particular strategy when speaking? It is also extremely difficult to attribute purpose to the use of the strategy in a test situation. It appears that testing strategy use is a high inference process, similar to that of testing the construct of 'fluency', and this is probably the reason why raters are rarely asked to score strategy use in a test.

II.2.3 Structuring Speech

Despite the misleading appearance that conversation is of a random nature, with speakers free to say what they want, when they want, and how they want, in fact, most speaking takes place in highly structured contexts. Participants usually take turns to speak, an event which involves interactional competence (the sequential organisation of speech, turn-taking and repair of communication). Recent research also considers how talk is sequenced and how turn-taking operates in situations where speakers are equal and unequal in social power (Markee, 2000). This work is potentially important for discourse studies that look at the type of language elicited by different task formats, and in particular will be a point of reference for later contrasting and advocating the different types of oral test included in this study: the individual oral interview versus the group speaking test.

II.2.4 Turn-Taking

The question of turn-taking in conversation has been addressed most fully by Sacks *et al.* (1974), who state that a speaking turn normally comes to an end at a *transition relevance place* (TRP). These places in conversation are usually easy to recognise because most speaking is structured in pairs of contributions that occur together naturally. These have been termed 'adjacency pairs' (Schegloff and Sacks, 1973, cited in Fulcher, 2003: 36) since they occur adjacent to one another and always follow a set pattern. (They will be discussed below, since they are elementary to an understanding of how speech works). The current speaker in a conversation actually possesses a great deal of power; for example, it is common for politicians to begin a speech by stating the number of points they wish to make. This, at the very least, should put others off from interrupting and if someone does attempt to speak before the list is complete, they will be perceived as rude. This is a strategy that is not usually acceptable in everyday conversation. A speaker may also select the next speaker by asking a question, or leave the floor free for anyone else to participate by ending a turn without specifying who should have the next turn.

As listeners in our first language, we are adept at recognising transition relevance places, and this accounts for why a new speaker often begins before the first speaker has completed their turn. The slight overlap between speakers is a result of the listener's ability to predict a TRP. Learners must also be good listeners if they wish to be good speakers, since they need to decide when it is appropriate to speak. It is important to consider the implications of turn-taking for second language learners, especially when they are taking face-to-face speaking tests.

Consequently, if turn-taking is considered important, it should be included as a part of the test construct. This, in turn, will have implications for the types of tasks included in a test. Saville and Hargreaves (1999) defend the use of tasks where test takers are paired and have to talk to each other. They argue that this is the best way to elicit turn-taking behaviour from test takers that is not dependent on the unnatural turn taking that occurs between a single test taker and a more powerful interlocutor/rater who manages the conversation in a sequence of questions and answers.

II.2.5 Adjacency Pairs

Adjacency pairs, as we have seen above, can be considered the most fundamental unit of conversational structure, and also the key to understanding how turn-taking works. Some examples of adjacency pairs given in Fulcher (2003: 36) are the following:

question – answer

greeting – greeting

invitation – acceptance

compliment – acceptance

request – compliance

offer – acceptance

complaint – apology

Although the first part usually predicts or expects the second part, it is also possible for a speaker to select alternative second parts, separating the adjacency pair with an inserted sequence. For example, in a ‘compliment – acceptance’ pair, we might observe the following sequence:

A: That's a pretty dress.

B: Thank you.

However, it is possible that the conversation will proceed in way similar to this:

A1: That's a pretty dress.

B2: I got it in the sale. Do you like it?

A2: Yes. That colour suits you.

B1: Thank you.

These inserted or embedded sequences can be much more complex, so adjacency pairs do not necessarily appear adjacently in the conversation, but the embedded pairs are always seen as a preliminary to the introduction of the second part. If a second part does not occur, then there must be an explicitly stated reason for its non-appearance.

Listeners must be able to understand the function of utterances such as questions, requests or offers, for example, in order to be able to predict TRPs, and they must be able to respond appropriately either with the second part of an adjacency pair or introduce an embedded sequence. The type of tasks included in a speaking test will require the recognition of different types of adjacency pairs and TRPs: in the interview format, learners will need to recognise and use almost exclusively the 'question – answer' sequence, while in other tests, such as the paired or group oral procedure, that involve a more equal power structure with candidate-candidate interaction, they will need to understand and manipulate a much wider variety of adjacency pairs (for example, in a group speaking test, candidates would need to both express and invite opinion; in an individual interview with an interlocutor/rater an invitation structure would be inappropriate for candidates).

II.2.6 Openings and Closings

Openings, and especially closings, to conversations differ according to whether social power is equal or unequal. In situations of unequal power, such as teacher/student or interviewer/candidate, one speaker has the right to bring a conversation to an end quite abruptly, without going through the formal closing routines. Many tests of speaking, especially those that follow the individual interview format, do not include tasks where the test taker is required to open and close topics and conversation. If we want to know whether candidates can structure conversation, the opportunity to do so must be presented and therefore needs to be included in the construct definition for the test.

II.2.7 Rules of Speaking and Pragmatic Appropriacy

The ability to communicate through speech involves much more than the knowledge of the grammatical and phonological system of a language. As Dell Hymes (1971: 10) states, "There are rules of use without which the rules of grammar would be useless." These 'unwritten' rules of speaking are often taken into account in tests of speaking in the use of terms like 'appropriacy', a construct concerned with the way in which native speakers instinctively use language according to social rules and pragmatic conventions of which they are hardly aware, and which second or foreign language learners may find much more complex to do automatically in the target language. In understanding learner 'errors' it is important to realise that if a grammatical error is made in speaking or a word is pronounced incorrectly, the listener is likely to be patient and make an effort to understand what is being communicated. However, if the error is pragmatic the consequences are potentially serious (for

example, using forms of address that are too informal in a situation where social distance is great, such as that of university professor/student).

Fulcher and Márquez Reiter (2003) used the notion of pragmatic scales, authority and distance, to investigate the difference between test takers from two different L1 cultural backgrounds. They recorded marked differences between English and Spanish L1 speakers in the same role-played situations (test takers performed the tasks in their L1) and that the Spanish speakers used forms of address that would be unacceptable or inappropriate in the English cultural context. In order for considerations of this type to be included in a construct definition, attention would have to be given to the features of tasks that would provide evidence for the ability of the test taker to manage the pragmatic force of utterances. If a speaking test is designed with communicative principles, including tasks that encourage or require test takers to direct questions at the interlocutor/rater, problems may arise where cultures require teachers to be treated with great respect and we may consider this a justification for the paired or group test format. In the same way, learners should be made aware of the communication norms of speakers with whom they are likely to interact, which in turn will have a washback effect on teaching.

II.2.8 Doing/Being Things with Words

It is an accepted fact that learners should know how to do things with words in the second language. Different languages do things with words in different ways and the impact that utterances have on listeners in one language will often not translate directly into the second language. Different languages perform speech acts in different ways, and this is crucial for issues such as politeness strategies. Searle (1969) draws a

distinction between the meaning of what a speaker says and the literal meaning of an utterance. For example, “It’s cold in here” probably means something like “Please close the window” or “Could you turn the heating on?” If politeness is felt to be an important factor for the contexts in which our learners will use spoken language, then we may consider including it in our construct definition and therefore in our test task design.

Although it is necessary to understand the role of the rules of speech in communication, the social context of speech is critical to understanding the aspect of appropriacy that Fulcher (2003: 43) terms ‘being things with words’. For example, the social status of different speakers will lead to a change in the directness or indirectness of speech acts. In a speaking test, if power relations between the assessor and the assessed are fixed, as in the interview situation, the context restricts test-takers in their use of language. The resultant discourse has been termed a ‘test genre’. In a group oral test there is a more equal power structure and it is therefore likely that a greater range and variety of language will come into play during the test.

Speakers also adopt roles in the use of language. In any particular context, the role the speaker is playing will have speaking rights attached: speakers of higher status have the right to initiate and close topics, and to direct the conversation, as is the case with teachers, or interviewers. So, we not only do things with words, but we are things through words in that we define our status and role through speech: it is the context and our place in it that dictates to a large extent the kind of language we use.

II.2.9 Interactional Competence

As we have seen above in Section II.2, interactional competence is defined in terms of how speakers structure speech taking into account its sequential organisation and turn-taking rules, and sometimes includes communication strategies. McNamara (1997) argues that interactional competence is a social/behavioural construct where joint behaviour between individuals is the basis for the joint construction and interpretation of performance. Unlike Bachman (1990), who sees interaction as ability within an individual, he proposes a more dynamic understanding of social interaction as performance within context. A whole range of factors can affect the quality of the performance and the test score. These factors would include interlocutor talk, the personality of the test-taker(s), and the nature of the task.

'Interactional Competence Theory' provides an alternative to the models of Communicative Language Ability with which testing has traditionally worked, giving importance to the co-construction of speech. In this model, abilities, actions and activities do not belong to the individual, but are *jointly* constructed by *all* participants (He and Young, 1998). If talk is co-constructed in this way in a speaking test, we need to consider how scores can be given to an individual test taker rather than to the pair or group.

II.3 EVALUATING SECOND LANGUAGE SPEAKING TESTS: CONSTRUCTING A VALIDITY ARGUMENT

It is doubtful that we can ever guarantee that a test score means exactly what we think it does. Validity is not an 'all or nothing' concept; it is an ongoing process to improve an argument and gather evidence to support it. In the case of the oral

proficiency interview tradition, the evidence supporting its use relies heavily on claims of 'face validity' (see II.3.2: both candidates and examiners 'feel' that it is a valid measurement of competence, but there is no empirical evidence to support this idea) along with language testers' experience of working successfully with this system of testing (Liskin-Gasparro, 1984; Lowe, 1987). Scoring procedures for such tests assume that each level of the rating scale represents a higher level of proficiency than the previous level and that the descriptors within the scale accurately describe the construct being measured. But if research into how learners acquire or use language has not been taken into account, the testing procedure can easily become the focus of criticism. In this case, the validity argument has been questioned on the grounds that the approach lacks theory and that what validity argument does exist, lacks empirical support (Lantolf and Frawley, 1985; Pienemann *et al*, 1988). Post-hoc investigation of validity is also unacceptable, since it means results cannot be related to initial hypotheses and constructs. Invented language samples, such as those used by Liskin-Gasparro (1984) produced to justify the band descriptors, cannot be used to support validity arguments. Much more sophisticated approaches to discourse analysis need to be applied to actual test performances in order to make claims about validity.

A 'non-compensatory' approach to rating, where everything in a band description must have been achieved before a test taker performance can be placed in a higher band also finds little justification in second language acquisition literature. This linear, building block view of language learning is not supported by SLA research, yet it is the basis for the ACTFL (American Council on the Teaching of Foreign Languages) speaking tests. The Functional Trisection (shown below in Figure 1),

taken from the web site <http://www.languagetesting.com/scale.htm>, assumes that each level is qualitatively and quantitatively more difficult to achieve than the previous one.

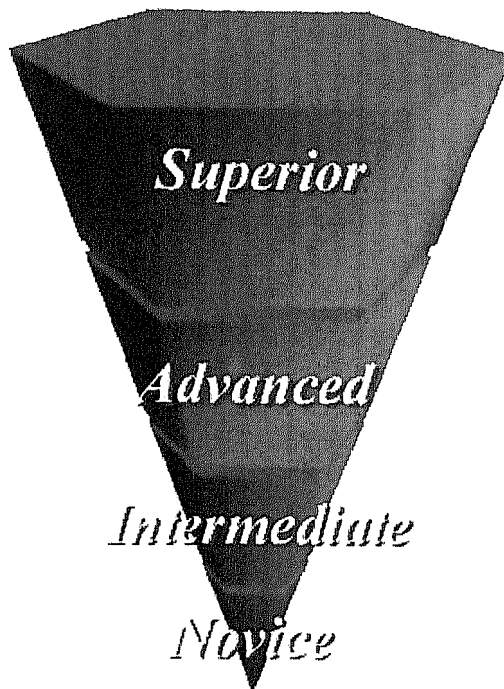


Figure 1: The Functional Trisection

Very little progress in language learning is necessary to progress at the lower levels, while comparatively much more learning is required in the later stages.³

In contrast to individual test formats, McNamara (1997) argues that the use of the paired or group format in speaking tests requires a validity argument that supports the hypothesis that individual scores can be given to candidates on the basis of discourse that has been jointly constructed. Current evidence suggests that the most plausible argument for this type of test is that the format provides an opportunity to

³ The levels are fully described on the web page cited.

test a much richer and more complex construct than is possible in the traditional interview (Swain, 2001).

II.3.1 From Testing Approaches to Curriculum Implications

Despite its problems, the model of language learning assumed by the majority of speaking tests which has become the basis for a whole approach to language teaching as well as language testing is the 'Proficiency Movement'. The wide acceptance of the principles of this movement constitutes a strong claim for the validity of the ACTFL approach to testing speaking. Lowe (1987, quoted in Fulcher, 2003: 177) claims that:

Its [the American Education Institute model of language acquisition as represented in the band descriptors of the rating scale] ultimate utility may lie beyond testing *per se* in its effect on curriculum. In this case, teaching for the test – teaching for general functional foreign language ability – is not to be discouraged.

The basic idea behind this argument is that testing drives curriculum. However, it is important to be aware that the descriptors in the rating scales do not represent the way in which language is acquired, nor do they provide informative feedback for the learner. Research has shown that the links between testing and teaching are not as clear as this and that there is no guarantee that a particular type of test or testing procedure will automatically lead to better teaching (Messick, 1996; Wall, 2000). It is also questionable as to whether the use of a descriptive system that has little theoretical or empirical support is capable of providing either a framework for teaching and learning, or a sound basis for the provision of feedback to learners.

Lantlof and Frawley (1985, 1988) are among the strongest critics of the Proficiency Movement and the rating scales on which it is based. They believe that this whole approach to testing speaking skills is flawed due to widely differing claims about the number of study hours required to reach the same level, and also due to the belief that it is easier for students to achieve greater accuracy on familiar rather than unfamiliar task topics, which is an explicit assumption of band scale descriptors on oral proficiency interview type tests. Yet little evidence exists to suggest that there is a relationship between accuracy of language production and the degree of 'abstractness' or familiarity with the topic.

The main problem Lantlof and Frawley (1985: 340) see in the ACTFL rating scales are the underlying assumptions taken with regard to how these scales represent levels of language acquisition. They claim that the approach to testing developed is 'analytical' rather than empirical, and that the analytical approach is presented as if it were based on empirical study. Speakers are grouped abstractly in levels organised from simple to complex (1 to 5), with the descriptions of these levels being produced subsequent to the grouping. It is therefore evident that the levels do not exist except in terms of the linguistic criteria which define them. The logic of the levels and their criteria is symmetric implication: X (levels) = Y (criteria), therefore Y = X. In other words, if we were to ask the question "What is a *low-novice*?" the answer would be "Someone unable to function in the spoken language". Similarly the answer to the question "What is someone who is unable to function in the spoken language?" the answer would be "A *low-novice*". This logic cannot yield a criterion-referenced test, but a criterion-reductive test: the criteria are the levels and vice-versa. The criteria, as absolutes, are converted into requirements because they are required absolutely to

define the levels. With the criteria as analytic and reductive, it is impossible to evaluate second language speakers except as they are described in the test guidelines, because the guidelines are absolute and reductive.

II.3.2 Validity

When dealing with concerns of validity in testing it is usual to distinguish between 'face validity' and 'construct validity'. Face validity, is the extent to which a test *appears to its users* to be a credible measurement of the construct (i.e. is based on intuition), while construct validity, is based on a prior definition of the construct itself and on theories of measurement.

Face validity, a common concern in the early years of communicative language testing and the touchstone for validity for many years, is concerned with the 'real-life' approach to language testing which essentially attempts to make test tasks look as though they are events which could occur in the real world by eliciting 'natural language' (Morrow, 1982; Lowe, 1987). In the case of the individual proficiency interview, this has been questioned on the basis of studies of 'interview talk' which suggest that the interview generates a special 'genre' of language different from normal conversational speech (e.g. Lazaraton, 1992; Young and He, 1998).

In addition to the fact that face validity is based on language testers' experience of successfully working with a testing system for many years which, they believe, leads to a consistent application of the criteria that should reassure critics and score-users, a further aspect to consider with regard to face validity is the presupposition that the speaking test is direct (i.e. that it directly measures speaking ability as apposed to simply the test performance). This has been severely criticised on the grounds that no

test of speaking can be direct in the sense of *direct measurement*; speaking tests usually aim to elicit a speech sample or performance which provides evidence of competence in speaking, so that score meaning can be generalised to other test tasks, and other speaking contexts (Bachman and Savignon, 1986: 382-3). All oral tests are therefore *indirect* measures of the construct, and it cannot be legitimate to rely solely on face validity to justify the use of a test, since experiential claims to validity do not constitute either theoretical rationale or empirical evidence. However, we cannot disregard face validity completely since it can also have an important effect on test performance and hence on results. If test-takers cannot understand the reasons for the tasks they have been asked to do, or feel that they are engaged in pointless exercises, this will necessarily affect the way they perform and in consequence the scores which are recorded. This is a serious problem both within an educational setting which involves assessing students or evaluating a teaching programme and for testing which is carried out for research purposes, because it limits the usefulness and application of the results.

Tests such as the U.S. Government Foreign Service Institute (FSI) oral interview rely heavily on face validity and yet are felt to be valuable measures of overall language proficiency. Clark (cited in Bachman 1990: 306) considers that “the great strength of direct speaking tests of the FSI interview type as measures of global proficiency lies precisely in the highly realistic testing format involved” and that “the direct proficiency interview enjoys a very high degree of face validity.” This type of oral test defines learners’ abilities in terms of what they are able to do in the foreign language, for example, they “can handle with confidence but not with facility most social situations including introductions and casual conversations [...],” and they “can

understand most conversations on non-technical subjects ...” (*FSI Absolute Language Proficiency Ratings*, Appendix 2). The interview is also structured in terms of language functions such as giving directions, elicited through role play, at the simplest level, and gradually becoming more demanding as the interviewee demonstrates the capacity to handle more complex situations.

The greatest claim for the validity of this approach (if we trust the FSI’s own evaluation of itself) is that both examiners and candidates have faith in the system and the scoring procedures and believe that the rating scale describes distinguishable levels of speaking ability demonstrated in the test:

This scale has become so widely known and well understood that statements like ‘The consul has an S-2 R-3 in Thai’ ... are immediately intelligible within meaningful limits of accuracy to everyone concerned with personnel assignments in the numerous government agencies who use the FSI testing facilities.

[...]

The examiners are made continuously aware that test ratings are commitments on the examinee’s linguistic capacity to perform certain functions, and it is obvious that these commitments are being met with sufficient consistency to enable many different groups to rely on them without question.

[...]

The examinees themselves have generally accepted both procedure and rating system as valid measures of their competence. (*FSI*, Appendix 2)

This kind of experiential claim to validity which has no empirical evidence or supporting rationale, should be treated with extreme caution; the fact that habitual users of a system believe in it does not necessarily mean that it is accurate or correct, and such statements as “rely on [the ratings] without question” do not contribute to a furthering of our knowledge and understanding of the measurement of language ability. We should at least be aware that the ability to “perform certain functions” as described here is pertinent only to certain contexts, and we should be wary of employing this test method as if it were a globally applicable way of assessing learners’ oral ability in any communicative situation. As Upshur (cited in Bachman, 1990: 250) points out, a test score which is seen as an indication that a person ‘is able to do X’ in the language, rather than ‘has ability to do X’, does not require a theoretical description of language ability and is therefore only sufficient as long as one is only interested in predicting future performance. We should also be aware that analysis of actual speech (either from learners or native speakers) does not yield models like the ones on which the descriptors are based. There is also an assumption present that the rating scales describe distinguishable levels of speaking ability which have been demonstrated in the test. This has been questioned on the basis of analysis of actual speech compared to band descriptors (Fulcher, 1987), the scalability of the rating scale (Pienemann *et al.*, 1988) and the theory of the method itself (Lantolf and Frawley, 1985, 1988).

In contrast to face validity we have construct validity, which is concerned with the extent to which test tasks reflect the theory on which the test is based. This concept has been extended to include the factors which it is hypothesised will affect test performance, and also the uses for which the test itself and its scores are intended.

Construct validity differs from face validity in that it attempts to take a scientific, rather than an intuitive, approach to measurement for which empirical evidence can be presented. In taking into account scores and their uses, this definition of validity also goes beyond the limitations of 'content validity' which is concerned only with the relationship between the test content and the domain of ability to be measured, without accounting for actual performance on the test. Even though the content of a test may be relevant, this does not allow inferences to be made about ability or mean that an individual's score is valid for different uses.

Consequently, construct validity is an implicit part of the authenticity of a test. Only if the test is based on a theory of language ability can it be said to be a measure of that ability. To the contrary, it may have a predictive utility, but it does not actually measure anything. It would also seem that tasks which do not elicit authentic interactions are unlikely to involve all aspects of a test taker's communicative language ability and, as such, are not valid measures of that ability. However, it is important to recognise here that the test, the testing situation, and the test taker may interact in a way which is authentic in their own context, but which does not necessarily replicate a real-life activity.

II.3.3 A Validity Model

In order to determine the plausibility or otherwise of a validity argument, it is useful to have access to a model which will help to determine what kind of evidence is required to either support or challenge an argument. Messick (cited in Fulcher, 2003: 194) proposes a generic validity model that can be applied not only to testing second language speaking, but to all types of educational testing and assessment. He defines

construct validity as “an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (Messick, 1989, cited in Fulcher, 2003: 194). The model defines six aspects of validity which can help us to decide which types of evidence should be collected to support or question the meaning of test scores in the context of test use. These are outlined below as applied to testing speaking skills:

(i) The substantive aspect: this deals with how it is possible to be certain that the processes the test takers use when responding to the tasks in the speaking test match the construct definition.

(ii) The structural aspect: this refers to the scoring procedure and how scores are reported. If the construct described is very broad, such as simply ‘speaking’, then a holistic scale may be used and a single score reported. However, if the construct is complex and broken down into different components or features, it may be necessary to use multiple- trait scales and report the score as a test-taker profile.

(iii) The content aspect: the content of the speaking test should have a strong relationship with the test construct as well as clear links to the programme of study, or content which is representative of the domain to which the scores are to be generalized.

(iv) **The generalisability aspect:** this is the extent to which we can say the scores are meaningful beyond the immediate context of the test. The test taker may also be a variable, particularly in the paired format. Generalisability is investigated in any context where it is claimed that the score obtained in one testing context would mean the same in another testing context with a different rater and different tasks (written to the same specifications). This increases the plausibility of an assumption that the test score can be generalized to non-test contexts.

(v) **The external aspect:** this concerns the relationship of the speaking test to other tests or variables outside the test. It considers to what extent a high correlation of different test results indicates that the tests measure the same construct.

(vi) **The consequential aspect:** the consequences of test use may be intended or unintended, for example an intended consequence may be to increase the importance of speaking in the classroom in a particular programme of study. Unintended consequences may be internal or external and, due to their very nature, are more difficult to measure. For example, an internal unintended consequence may be that the design of a task makes it easier for males than females.

II.3.4 Alternatives to the Individual Proficiency Interview: Testing Speaking in Pairs or Groups

Folland and Robertson (1976) were the first to recommend the use of more than one test-taker in speaking tests, primarily on the grounds that it would reduce test anxiety. There have been other reports of the successful use of such tests, although

most are experiential rather than research-based; for example, Reves, (1980, 1991); Shohamy, *et al.*, (1986); Hilsdon, (1991); Taylor, (2000a). From the perspective of test validity and authenticity, the pairing or grouping of candidates provides a more varied sample of interaction than an individual interview, i.e. candidate-candidate as well as candidate-interlocutor. In their English as a Foreign Language speaking examinations, Cambridge ESOL use a paired format at the 'Key English Test' (KET), 'Preliminary English Test' (PET), 'First Certificate in English' (FCE), 'Cambridge Advanced English' (CAE) and 'Cambridge Proficiency English' (CPE) levels; even at the 'Movers' and 'Flyers' levels of 'Young Learners English' (YLE) which use an examiner-candidate format only, the examiner and the candidate are, at one point during the test, engaged in a collaborative exchange.⁴

The Cambridge ESOL rationale for examining candidates in pairs is based on several premises. Firstly, the development of testing in the UK has been closely linked to language pedagogy and the introduction of the paired speaking test brings testing into line with classroom practice where students commonly work in paired co-operative and collaborative tasks. Also, research on the one-to-one interview format (Ross and Berwick, 1992; Young and Milanovic, 1992) has shown that the interaction in this type of speaking test was asymmetrical because of the unequal power status of the two participants (interviewer/candidate). The paired format allows a range of different interactions and also addresses the question of power structure which undoubtedly has an effect on the type, quality, and amount of discourse produced during the test. An unpublished Cambridge ESOL study (cited in Taylor, 2000a)

⁴ The first level of YLE, 'Starters', only requires candidates to understand and respond by placing picture cards in the correct place on a larger illustration or by providing minimal responses to interlocutor questions.

showed that in the paired format candidate turns were increased and the amount of talking time attributable to the interlocutor substantially reduced. Furthermore, the number of different language functions observed in the paired format was much higher than in the interview; an *a posteriori* analysis of CPE test-taker performances showed that from a list of 30 communicative language functions which characterise spoken discourse, the one-to-one format elicited just 14, while the paired format was able to elicit an average of 26 of the 30 (cited in Taylor, 2000a). These findings are in line with Saville and Hargreaves (1999), who argue that the use of the paired format should be seen in the context of test design, in which a variety of tasks are used to elicit a wide range of language. The format and range of tasks allow a broader range of construct features to be represented in the test.

There are however, several issues that need to be considered when introducing a paired or group format for oral testing with two examiners in the role of interlocutor and rater. Some of these are summarised below:

PAIRS

- Who is paired with whom? Should test takers be familiar with each other or does it matter if they are strangers?
- Does it matter if their L1 is not the same?
- Should they be at roughly the same stage of L2 or can they be at different stages?
- What is the effect of personality differences between test takers, e.g. pairing extrovert and introvert candidates.

- Does the test format result in a reduction or an increase in test taking anxiety, depending on the various types of pair combination possible?

INTERLOCUTOR AND RATER

- What is the impact of the role of the rater on the test takers?
- The interlocutor also rates the two candidates, as well as participating in the interaction. Does this enhance the validity of the rating process?
- How do the raters assign grades to each of the test takers separately when, due to the differences there may be brought about by all the candidate variables listed above, one may be supporting the other, or one may not be providing the other with an opportunity to show how well s/he can 'negotiate or take turns'?
- How much should the interlocutor intervene?
- What is the effect on discourse and scores if 'significant' intervention by the interlocutor is required, or if one test taker gets more talking time than the other?

Research in these areas is growing and also addresses the group oral testing format where three test takers are required to participate in a speaking task and similar issues are at stake. Fulcher (1996a) reported from questionnaire data on a range of test tasks that students generally thought that the group discussion task generated the most natural discourse, created the least pre-test anxiety, and that over half the test-takers preferred the discussion task to other task types. However, while studies of learner perception are important, the Fulcher study did not take into account the key variables of personality or language proficiency.

Berry (cited in Fulcher 2003: 188) has undertaken extensive research into the interaction of introvert and extrovert students, and found that discourse varies according to the pairing. Both introverts and extroverts performed better when placed in homogeneous pairs, whereas in mixed pairs, introverts did not perform as well as extroverts. However, both introverts and extroverts performed better in a paired test than they did in a one-to-one interview.

In a study to determine the effects of learner acquaintanceship on test performance, O'Sullivan (2002) found strong evidence to support the hypothesis that candidates would achieve higher scores when working with a friend due to the lowering of anxiety levels. However, analysis of the discourse in the same tests revealed that there was no effect on the complexity of the language produced. Parallel to this, Iwashita (cited in Fulcher, 2003: 189) investigated the impact of the level of ability of one learner on another in the paired test format. He found that lower ability test takers talked more when paired with a higher ability test taker, but that the amount of talk was not related to the test score.

Research is also currently being conducted into the negotiation of meaning that takes place in paired or group interaction. It is claimed that this negotiation fuels L2 acquisition (Swain and Lapkin, 2001) and the format is recommended for speaking tests on the grounds that 'dialogues construct cognitive and strategic processes which in turn construct student performances, information which may be invaluable in validating inferences drawn from test scores' (Swain, 2001: 275). In other words, the pair or group format may generate language performances that allow us to test much more complex constructs than in a traditional one-to-one interview. The problem now lies in defining constructs that recognize the co-construction of discourse and meaning

where the speech sample to be assessed is seen to be produced through interaction, collaboration, and support (McNamara, 1997). Thus, social interactional investigations need to consider focused hypotheses of the complex interaction of linguistic and non-linguistic knowledge, that is cognitive, affective and conative attributes engaged in particular situations which make context critical for test development as well as for test score validation.

In relation to our own test design for the current research project, the paired and group format opens up the possibility of enriching our construct definition to include interaction and context and hence the meaning of test scores. The validity argument is plausible given our current state of knowledge; as further research is conducted it may get stronger or weaker.

II.4 TEST SPECIFICATIONS AND VALIDITY

Specifications for language tests should contain a statement of the test construct, a description of the tasks that will make up the test, a description of the test format, a statement about what kind of responses we expect test takers to make, and an explanation of how the performances are going to be scored. Thus, they bring together several of the most important features of tests and together comprise an argument or rationale for the *validity* of the test; by means of detailing why test design decisions are made and creating the test specifications, we contribute to a validity argument that relates test scores to constructs. Test specifications are dynamic, evolving documents that should be subjected to the process of test design, piloting and revision.

The concern with validity means that we need to focus on construct definition at many levels of the testing process. Whereas validity was previously seen in terms of

whether a test 'measures what it is intended to measure' (Hughes, 1989: 22), a quality that is either present or absent, the current research position now sees providing empirical evidence and theoretical rationale as a validity argument that should "present and integrate evidence and rationales from which a validity conclusion can be drawn pertaining to particular score-based inferences and uses of a test" (Chapelle, 1999: 263)

This argument encompasses all kinds of evidence that impact on our understanding of what the score might mean. It includes documentation of how a test is developed, the decisions made during the design process, and the reasons for those decisions. Fulcher (2003: 117) suggests that the kinds of activities that need to be documented are:

- identification of the members of the design team
- identification of the test takers
- definition of test purpose
- definition of the construct
- design of prototype tasks
- piloting of prototype tasks
- working on initial ideas for rating scales and band descriptors
- writing and revising task administration instructions
- carrying out research to support design decisions
- making explicit any constraints in test design

This documentation forms part of the validity argument linking constructs to tasks and rating scales through a record of design decisions, and forms the basis of the research methodology of the current project (see Chapter III). In particular, it shows how a design team attempts to avoid construct under-representation and construct-irrelevant variance by recording the development of the test specifications as they evolve. Through the process of collecting evidence during the design phase, the kinds of claims that the test designers wish to make about the meaning of scores can be supported.

Test specifications are different from a test syllabus in some important ways: a syllabus is usually oriented towards the needs of teachers and learners, and includes information such as the level of the test, its format, what is being tested, what task types are used, who the interlocutors are in a face-to-face test, and what the criteria for successful performances are. These criteria are generally extracted from the rating scales, but are presented in a more user-friendly format. Test specifications are used by test designers and refer to the overall format of the test and to the individual tasks that are included in the framework. They gradually become the basis for test and task writing and for the ongoing investigation of validity issues. These include:

- writing many tasks that ‘appear’ the same and that can be placed in a task bank for creating parallel forms of a speaking test.
- investigating whether the speaking test elicits the processes or language that was predicted by the task writers.

- varying the form of items in future versions so that the test evolves in line with future validity studies and new discoveries in language acquisition and applied linguistics.

II.4.1 The Writing of Test Specifications

Test specifications grow out of a dynamic process of discussion, piloting, and information collection based on research. Their drawing-up is often therefore not clear-cut, especially in the early phases, and recording the process can be a complex task. The resulting document is a record of test design and development decisions, forming part of a validity argument. According to Chapelle (1999: 263), “A validity argument should present and integrate evidence and rationales from which a validity conclusion can be drawn pertaining to particular score-based inferences and uses of a test.”

In contrast to an approach which centres on the task, and then attempts to assess the language sample that arises from it in terms of how well the task has been carried out, the emphasis in Chapelle’s argument is on making the rationale for the test explicit from the outset. The design originates from the concept of the construct, subsequently focuses on how to collect the evidence and finally addresses the type of tasks that will elicit that evidence. This approach also allows us to focus on the rationale behind the task, helping to generate new tasks of the same type and level for a single, or subsequent, testing sessions.

Fulcher (2003:130) suggests a test specification format derived from work carried out by Davidson and Lynch (2002). The essential features of the format are summarized below in Table 2:

- **General description** – a brief general statement of the behaviour to be tested. It also provides a summary of the construct that underlies the test.
- **Prompt attributes and responses** – a complete and detailed description of what the test taker will encounter. These provide evidence for why a task or set of tasks was selected and also allow the generation of multiple forms of the same task. They should thus provide the basis for investigation and the strengthening of validity arguments.
- **Response attributes** – these describe the type of response the student will perform, including the criteria for evaluating that response. In effect, they are a statement of how a task will be scored before it is administered.
- **Sample item** – an example of a task that will be generated by this specification.
- **Specification supplement** – a detailed explanation of any additional information needed to construct items for a given specification.

Table 2

This protocol facilitates the production of new test tasks or items and also makes it more likely that different forms of the test will elicit similar language samples, that task difficulty will be similar, and that scores will therefore be reliable. It also becomes easier to produce many tasks for the task bank when writing to a specification which is a much more efficient use of resources than writing many different stand-alone test forms.

It should be noted that the specification introduces freedoms as well constraints, highlighting elements of the tasks that the writer may vary. This degree of freedom should be explicitly discussed. If it allows the generation of task types that were not envisaged at the outset, unwanted variability that may not be construct-related may be introduced into test scores. Thus freedom should be built in to allow the generation of different, but similar tasks. However, Davidson and Lynch (2002: 65) warn that the power of a well-established test specification to guide practice can sometimes be dangerous; although it may help with test validity, it can also lead us to a false sense of security. We need to remain sensitive to advances made in research which may change our understanding of second language acquisition, discourse analysis, or interaction theory. Specifications need to be frequently revised, especially where they act as an explicit statement for the focus of learning. They can have an impact on how teachers see their role and act as a central point of discussion for syllabus design, as well as embodying a statement of what is currently being valued in language learning and teaching.

IL5 TASKS FOR SPEAKING TESTS

As we have outlined previously, the purpose of a speaking test is to collect evidence in a systematic way (by means of elicitation techniques or tasks) that will support an inference about the speaker's ability with regard to the construct as we define it from the summary of this evidence (the score). Evaluators will also usually be interested in the candidate's ability to perform in a range of situations which is much wider than those that could be sampled during the test. From a sample

performance, we need to be able to make inferences about the likely success or failure of the learner's future performance in non-test contexts.

This is one of the key challenges in testing speaking skills: designing tasks that elicit spoken language of the type and quantity that will allow meaningful inferences to be drawn from scores about the learner's ability within the construct the test is designed to measure. In order to do this, the test must avoid two threats to its construct validity:

- *Construct under-representation*, or the extent to which a test fails to capture important aspects of the construct it is intended to measure.
- *Construct-irrelevant variance*, or the extent to which test scores are influenced by factors that are irrelevant to the construct it is intended to measure.

If task definitions are not included in the construct (which is often the case), any variance attributable to the task type will constitute construct irrelevant variance.

In previous decades, there have been attempts by authors to provide extensive lists of task types with a discussion of their advantages and disadvantages for assessing spoken language (e.g. Masden, 1983; Underhill, 1987; Weir, 1990). More recent literature, however, has tended to avoid the discussion of task types and to concentrate on the role of the task in eliciting spoken language and how this can be linked to the construct definition in order to be able to vary task content and yet maintain the level of difficulty, the type of response required, and the opportunity for a similar size and quality of language sample to be produced. The development of models of communicative competence (Canale and Swain, 1980; Bachman, 1990; Bachman and Palmer, 1996) has made it possible to see tasks in other ways and not only in terms of

their general usefulness in eliciting a language sample, because they recognise that speaking takes place in specific social settings, and with particular communicative goals. This perspective has provided a common point of departure for those involved in language testing, language pedagogy and second language acquisition research.

Candlin (1987:10) defines a task as:

One of a set of differentiated, sequenceable, problem-posing activities involving learners and teachers in some joint selection from a range of varied cognitive and communicative procedures applied to existing and new knowledge in the collective exploration and pursuance of foreseen or emergent goals within a social milieu.

This is later reiterated by Bachman and Palmer (1996: 44), who state that *test* tasks are associated with a specific social situation, that task participants (test-takers) are goal-oriented and that the tasks involve the active participation of the candidates; they define a test task as “an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation.”

Candlin (1987) defines tasks in terms of seven characteristics that make up the basis of their use in the classroom:

- input, or material used in the task
- roles of the participants
- settings, or classroom arrangements for pair/group work
- actions, or what is to happen in the task
- monitoring, or who is to select input, choose role or setting, alter actions
- outcomes as the goal of the task



- feedback given as evaluation to participants

Nunan (1989) provides an almost identical list:

- goals
- input
- activity
- teacher role
- learner role
- settings

These characteristics can be used to describe tasks and thus to select and design tasks for speaking tests that allow score inferences to generalise to the domain of target language use beyond the limits of the context of the test. By providing an interface between construct definition and task description, it is possible to identify the *process* which the learner/candidate is required to engage in while undertaking the task, as well as the characteristics of the task itself, allowing the task to be viewed within the terms of the construct. We will attempt to apply these criteria in our own task design for the current project as described in Chapter III, Section III.5.

II.5.1 Test Task Characteristics

Models of communicative language ability and use provide a framework for both interpreting the components of a construct and the dimensions of different tasks by defining their characteristics. Bachman & Palmer (1996: 49) argue that, through identification of the characteristics of language tasks, it can be shown how performance on these tasks may be related to speaking 'in the real world'. They provide a list of test task characteristics which includes the setting, the test rubrics

(including instructions, structure, time allotment, and scoring method), and the input. However, to date, little research has been conducted into which task features need to be recorded in test specifications for speaking tests, so that task writers can produce sets of comparable tasks. Bachman and Palmer (1996: 44) refer to 'target language use tasks' that exist in the 'real world', to which inferences from test scores need to generalise. The analysis of the target language tasks provides the characteristics for the tasks to be used on the speaking test, and is, in part, what they mean by 'authenticity'. 'Authenticity' here forms part of a perceived relationship to validity, but Lewkowicz (2000) questions whether it is, in fact, possible to match test tasks to real-world tasks using all of the items in any given checklist. She claims that authenticity is really a concept rather than a construct, differentially interpreted by test takers, and is a matter of perception rather than of external reality.

Probably, it is more important to describe tasks in ways that contribute to the development of test specifications than to attempt a description of authentic tasks. If speech is to be scored, then we need to be aware of what kind of language we expect from the tasks chosen for the test. The Bachman and Palmer (1996) model is designed to be generic and applicable to all language tests in general. Weir's (1993) performance conditions are more specifically related to testing speaking. These contain the features of status and familiarity, missing from the Bachman and Palmer model and which prove to be more important predictors of task difficulty than many other criteria that have been investigated up to now. Weir's list also includes the items of speaking rights and responsibility for continuance of the interaction, which opens the way to describing tasks in terms of the extent to which they might encourage the

co-construction of discourse, an important feature which we have highlighted previously.

In describing test tasks, it is possible to use any of the categories from the above models as long as they are relevant for the test design process and the subsequent analysis of how successful the tasks are in eliciting samples of language that can be rated. Test designers can select categories for task description that are appropriate to their own teaching and testing contexts. Fulcher (2003: 57) proposes a shorter, workable framework for the description of tasks for speaking tests shown here in Table 3, based on the models included above (Weir, 1993; Bachman and Palmer, 1996):

- | |
|--|
| <ol style="list-style-type: none">1. Task orientation<ul style="list-style-type: none">• Open: outcomes dependent upon speakers• Guided: outcomes are guided by the rubrics, but there is a degree of flexibility in how the test taker reacts to the input• Closed: outcomes dictated by input or rubrics2. Interactional relationship<ul style="list-style-type: none">• Non-interactional• Interactional<ul style="list-style-type: none"><input type="checkbox"/> One-way<input type="checkbox"/> Two-way<input type="checkbox"/> Multi-way3. Goal orientation<ul style="list-style-type: none">• None⁵• Convergent• Divergent |
|--|

⁵ A task with no goal orientation would simply require the test-taker to carry out an instruction, such as 'Read out point six' from a list numbered 1 – 10. In a convergent task, a test-taker might be asked to speak about a topic indicated by the examiner, e.g. a recent holiday, while in a divergent goal-oriented task, they might be presented with a controversial issue and asked to discuss it with another test-taker, expressing their opinion or arguing their point of view.

4. Interlocutor status and familiarity

- No interlocutor
- Higher status
- Lower status
- Same status

5. Topics

6. Situations

Table 3

These categories are appropriate for producing comparable sets of test materials and also in subsequently evaluating them with relation to their outcomes once they have been implemented.

IL5.2 Task Difficulty in Pedagogy and Second Language Acquisition

Brown and Yule (1983: 37-53) were among the first researchers to discuss the difficulty of speaking tasks. They suggest a number of factors that might make tasks more or less difficult for different task types; these are summarised below (Table 4):

- Narrative tasks
 - increasing the cognitive load
 - using images that have different cultural implications
 - proliferation of 'same-type' participants in the story (e.g. same gender characters, so candidates cannot use he/she to distinguish between them
Brown and Yule refer to this as 'communicative stress')
- Tasks requiring descriptions and instructions
 - the more pieces there are, the more complex the task
 - the cognitive load is increased by using culturally unfamiliar material
- Extended discourse tasks
 - linguistic requirements for task completion
 - cognitive requirements for task completion
 - discourse requirements for task completion

Table 4

These concerns focus exclusively on features of the tasks themselves, assuming that this is the only variable that may affect performance. However, exploration within the second language acquisition literature of features that might make tasks more or less difficult has a wider focus, going beyond the materials themselves to consider the impact of other aspects of the context. The research of Tarone (1988) shows that variability in conditions such as physical setting, topic, and participants can vary the difficulty of tasks and she argues that the construct of a 'stable competence' is therefore untenable and that performance data can only support the weaker construct of 'variable capability'.

This variationist approach poses a problem for language *testing*, as argued by Fulcher (1995) since, if this were the case, each test would only be a test of performance in the specific context defined in the task facets of the test situation, and it would thus be impossible to generalise the meaning of test scores from any test task to any other task, or any non-test situation, unless there were a precise match between every facet of the test and the criteria. On summarising the most recent research in the field, Fulcher (2003: 62) has concluded that:

The assumption underlying present SLA-influenced approaches to studying speaking tasks is that there is variation in test-taker performance by task characteristics or conditions, and that this variation leads to different scores (or estimates of speaking ability) under different test task conditions. This encourages the language test researcher to consider task features or conditions in relation to task difficulty and how this may impact upon what inferences may be drawn

from scores on speaking tests in relation to the tasks students are asked to perform.

Chapelle (1998, 1999) had anticipated a number of possible positions in relation to this argument, which are summarised below:

- **the ‘new behaviourism’ position:** inferences drawn from test scores may only be generalised to identical tasks in other tests or the real world.
- **the trait theory position:** test scores are not task specific and tasks are, for the most part, interchangeable. Scores represent underlying constructs that enable speech and from which we can generalise to other speaking tasks in other tests or the real world. (This is the current position of the vast majority of high-stakes speaking tests currently in use).
- **the interactionist position:** some features of a task may have an impact on generalisability and these need to be investigated. While the general degree of the impact of some task features is usually small, it is these features that could be manipulated to make a test task specific to a particular situation in tests of speaking for specific purposes. (This is a potential way forward for researchers to better comprehend how much context is part of the speaking construct).

Part of the issue at stake here is the extent to which the test task should be included in the construct definition. It is only under the assumptions of the trait theory position, which has been held until very recently, that task types are seen to be irrelevant to the construct, with scores assumed to reflect underlying competence that can be transferred to innumerable situations outside the test context. The interactionist position provides a new ‘middle ground’ which admits that there may be contexts in

which the task may be part of the construct definition. The challenge for researchers is to attempt to focus on those aspects or features of tasks that influence performance in such a way as to be able to describe them within the construct of speaking and thus gain confidence in the belief that test scores are a meaningful reflection of underlying ability and competence, and hence generalisable to other contexts.

Although second language acquisition researchers have considered some of the features of difficulty in speaking tasks in relation to classroom pedagogy and how they can promote learning (e.g. Brown and Yule, 1983), the criteria that can be used to predict task difficulty in speaking tests have been conceptualised in a slightly different way. Although it has been recognised that the structure of the interaction is important in test task design in order to ensure the elicitation of a range of discourse in speaking tests (Shohamy *et al.*, 1986), psycholinguistic categories have also been used in the empirical prediction of task difficulty. For example, Wigglesworth (1997) found that one minute of pre-task planning resulted in measurable improvements in the complexity, fluency and accuracy of candidates' speech, although this was not reflected in the scores assigned by raters.

Skehan (1998a, 1998b) has suggested that various psycholinguistic categories will affect task difficulty:

- **familiarity of information:** the more familiar the information on which a task is based, the more fluent the performance will be.
- **structured tasks:** where tasks are clearly based on a sequential structure, the performance will be both more fluent and more accurate.

- **complex and numerous operations:** the greater the number of operations and the more transformation of materials that need to be performed, the more difficult the task will be.
- **complexity of knowledge base:** the wider the knowledge base on which a task draws, the more complex the language that will be produced.
- **differentiated outcomes:** where the outcome of a task requires greater justification for its resolution, the complexity of the language produced will increase.

This research refers to tasks used in a classroom context, but Brown *et al.* (1999; cited in Fulcher, 2003: 63) showed that it was impossible to replicate the prediction of task difficulty in language tests using these categories. Iwashita *et al.* (2001) and Elder *et al.* (2002) further investigated the possibility of establishing criteria for task difficulty in terms of task performance conditions. Modifying the Skehan (1998a, 1998b) model they investigated the following criteria:

- **perspective:** telling a story from one's own perspective or from the perspective of a third person.
- **immediacy:** telling a story with and without pictures.
- **adequacy:** telling a story with a complete set of pictures, and with some pictures missing from the set.
- **planning time:** with and without three minutes to prepare a task.

These studies are unusual in that they combine both an analysis of the discourse produced from the tasks, and an empirical analysis of task difficulty. They have shown that varying task conditions had no significant impact on the discourse produced under

test conditions and that there was no large significant relationship between task conditions and task difficulty (although Iwashita *et al.*, 2001 claimed that there was a tendency for candidates to produce more accurate language under the conditions presumed to be more difficult). They propose that an explanation for this may be that the differences in testing and teaching contexts are so great that they alter the cognitive focus of the tasks. For example, in an interactive task carried out with a classmate, the focus will be on the completion of the task, while in a test situation, delivery may be halting independently of whether the task is easy or difficult, because the candidates are focusing primarily on correctness. The lack of complexity in production may also be due to anxiety about how their speech is being evaluated, making them reluctant to take risks even when task conditions allow for this. This finding may be an argument in favour of replicating classroom conditions in oral tests to the extent that this is possible, that is, by carrying out group oral tests where candidates take responsibility for both directing and resolving tasks among themselves rather than simply following examiner directions.

In the Elder *et al.* study (2002), learners were also asked to complete questionnaires regarding their perceptions of task difficulty, but feedback from test takers provided no support for the Skehan model with reference to the impact of conditions on task difficulty, suggesting that it should not be relied upon as a basis for test design or for test validation arguments. Fulcher's (1996) study also found from questionnaire results that certain tasks were perceived by students to encourage more natural conversation than others. Almost half of a group of 47 test takers responded that engaging in group discussion gave them more confidence to speak than having to respond to an examiner.

The lack of score sensitivity to variation in task has also been discussed in language testing research. Clapham (2000) claims that specificity as a task condition has failed to generate enough score variance for it even to be worth maintaining subject-specific modules in tests such as the IELTS. Language for Specific Purposes (LSP) testing has still to show what it is about an LSP test that makes it specific, i.e. easier for those with the specific knowledge and more difficult for those without it (Douglas, 2000).

However, this does not question the view that changes in task or task conditions result in changes in the discourse produced. It is evident that a change in task topic or the number of participants will change the discourse produced by test takers. What is crucial here is that changes in discourse *necessarily translate into changes in test score*, and hence the estimate of task difficulty. The research cited above consistently shows that it requires gross changes in task type to generate significant differences in difficulty from one task to another, and even then the task accounts for little score variance. In nearly all studies, learner ability accounts for most score variance, and task difference for only a small part of score variance. Fulcher (1996) reports significant but very small differences in task difficulty that account for test score variance between a picture description task, an interview based on a text and a group discussion, while the only language testing studies to find large significant differences between how learners perform on tasks are those where the tasks are maximally different (Bachman *et al.*, 1995; Fulcher, 1996b, cited in Fulcher, 2003) and employ multiple rating scales (different tasks are rated on different scales in the same test).

It seems likely then that specific rating scales generate large task-specific variance. Fulcher (1996b, cited in Fulcher, 2003: 66) has shown that rating scales that do not refer to specific task types, task conditions, or tasks generate scores with most variance accounted for by test-taker ability. This finding upholds the hypotheses regarding scale specificity and independence originally presented by Bachman and Savignon (1986) which suggest that the inclusion of task-specific references in the ACTFL rating scales led to difficulties in investigating construct validity, because test method facets were built into the scoring process. Present research findings therefore suggest that it is the rating scale which invests the specificity in the task rather than any feature inherent to the test task itself. Parallel to this, Fulcher and Marquez Reiter (2003) have shown that a discussion of task difficulty only makes sense in relation to specific speakers. Hypothesising that pragmatic task conditions will have an impact on task performance, discourse and score, and that these conditions are culturally related to the candidate's L1, they found that there was a three-way interaction between social distance, degree of imposition and L1.

Thus, despite the research undertaken to date, it remains unclear as to what precisely it is that makes a task easier or more difficult for any particular group of test takers. It would seem that 'difficulty' does not reside in the task itself, but rather it is a combination of tasks, conditions and test takers. For this reason it should be possible for teachers who design tests for their own students to develop tasks within a framework that can isolate key features that affect performance. The challenge for international tests of speaking is greater, since it is more difficult to remove construct unrelated difficulty from tasks across multiple cultural contexts.

IL.5.3 Types of Tasks in Speaking Tests

As we have seen in the above discussion in relation to task design, the principle considerations that need to be taken into account in selecting task types for a test of speaking skills are whether the task will elicit a performance that can be scored, and whether it will be possible to make inferences from the score to the construct we intend to measure. Task classifications can help the test designer to select the most appropriate set of tasks for a specific purpose, given that tests need to be short enough to be practical and economical, and long enough to be reliable and to provide evidence to support valid inferences. The task characteristics and task conditions that we use to describe the tasks should also reflect the type of variables that we might expect to affect task difficulty for the intended test takers.

As we have seen above, there are different approaches in describing types of tasks. They can be seen in terms of the activity involved in carrying them out, for example, 'picture story task', 'answer questions by interpreting information given visually on a map or graph' or 'describe a picture'; or in terms of what they may be used for, such as 'tasks suitable for tape-mediated speaking tests', or 'tasks suitable for extended speaking'. However, it is probably more useful to describe them in terms of the features they possess and the type of speech sample they may elicit, thus allowing meaningful inferences to be made from scores to constructs. A summary of the categories suggested by Fulcher (2003) for classifying tasks is given in Table 5 below:

- **Task orientation:** this may be closed (completely directed by the input), guided (test takers may develop their own ideas in relation to the material

provided), or open (speakers may develop a topic in any way they wish)

- **Interactional relationship:** non-interactional (in the case of a tape-mediated test), one-way (narrating a story from picture prompts), two-way (information gap, interview format, paired test format), or multi-way (group oral format)
- **Goal orientation:** convergent or divergent
- **Interlocutor status and familiarity:** higher or same status; familiar or unfamiliar
- **Topics:** variable or limited according to the purpose of the test
- **Situations:** eg. job-related; discussion

Table 5

This method of classification is a useful one, since it enables a comparable description of varied and differing tasks used in all kinds of speaking tests that are currently in practice throughout the world. It does not, however, provide us with orientation as to the type of discourse that a task might produce, the size of the sample we might obtain, or with any information about how scoring on a particular task may relate to the construct. The way in which personality factors contribute to the co-construction of speaking and studies of how interlocutor/raters accommodate their speech to that of the test taker, are also aspects whose inclusion in construct definition prove both controversial and unclear.

It has frequently been claimed that the discussion task situation is the one that will allow assessment of the widest range of competences and knowledge. Underhill (1987) claims it is most the natural task type since it is designed to put test takers in a situation where the interaction may not resemble 'test-like' discourse. However, it

remains controversial due to the assumed willingness of test takers to enter into a dialogue in the presence of an examiner. For assessment purposes, there is also the problem of how a score can be given to an individual when the interaction is co-constructed, even in a guided task. However, Fulcher (1996a) advises that this should not deter testers from using the discussion task, since there is evidence that it is a model preferred by test takers over other task types. The examining body, Cambridge English for Speakers of Other Languages (ESOL) implies that this task type can provide evidence from which we may draw inferences not only about accuracy, but also about the complex constructs of discourse management and interactional competence, since these categories are included in the rating scales for several of their tests of spoken English.

II.6 RATING SCALES

We shall now turn our attention to test scoring procedures and rating scales. The ultimate aim of rating scale design is to link constructs, band descriptors and design processes to the types of inferences that we may make from scores on speaking tests. A rating scale, often referred to as a *scoring rubric* or a *proficiency scale* is defined by Davies *et al.* (1999: 153-4) as:

[A] scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency. Typically, such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are

commonly characterised in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion) [...] Scales are descriptions of groups of typically occurring behaviours: they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the population and test purpose. Raters or judges are normally trained in the use proficiency scales so as to ensure the measure's reliability.

Thus, the scale provides an operational definition of a linguistic construct. This is the position taken by most of those who use rating scales, which takes for granted that the scale will be used to score speech samples and to guide developers in the selection of tasks for tests. However, Alderson (1991b) has suggested other uses for rating scales:

- User-oriented scales – used to report information about typical or likely behaviours of test takers at a given level.
- Assessor-oriented scales – designed to guide the rating process, focusing on the quality of the performance expected.
- Constructor-oriented scales – produced to help test constructors select tasks for their inclusion in a test.

The level of detail in each of these types of scale may be quite different: for example, in 'user-oriented scales', band (or level) descriptors may be expressed in terms of 'can

do' statements,⁶ while assessor-oriented scales necessarily make reference to construct definition. Constructor-oriented scales will be more detailed and contain references to the types of task most likely to elicit the language sample required for scores to be meaningful. Our principal concern here is with the rating scale as a guide to the rating process, that is, with assessor-oriented scales.

Traditionally, a distinction has been made between holistic and analytic rating scales: in the former, the assessor gives an overall score to the performance and is not required to count or take into consideration specific incidents of any particular feature of the construct. The latter, by definition, is the opposite, namely the counting or tallying of incidents that occur in different areas of the sample of language under assessment in order to interpret the outcome as a score.

Hamp-Lyons (1991, cited in Fulcher, 2003: 89-90) distinguishes between *holistic scoring*, *primary-trait scoring* and *multiple-trait scoring* in tests of speaking. These can be defined in the following way:

- **Holistic scoring:** a single score is given to each speech sample, either impressionistically or guided by a rating scale. This can be problematic because it does not take into account the constructs that make up speaking. Furthermore, as we have seen, a single score may not do justice to the complexity of the speaking skill.
- **Primary-trait scoring:** assumes that a speech sample can only be judged in its context, so rating criteria should be developed for each individual task.

⁶ The European Common Framework of Reference makes use of such statements to define levels of language proficiency (e.g. Level C1: "Can understand a wide range of demanding, longer texts and recognise implicit meaning").

- **Multiple-trait scoring:** provides multiple scores for each speech sample, with each score representative of some feature of the performance, or of the construct underlying the performance. The scores are related to constructs rather than tasks, so they may be generalised across a range of task types. The main disadvantage of multiple-trait scoring is that, frequently, raters cannot make the distinctions required to assign three or four separate grades for one speech sample leading to a tendency to give the same grade across categories. This ‘halo effect’ has come to be known as *cross-contamination* (Alderson, 1981).

Fulcher (2003: 91) proposes a framework for describing rating scales based on the three categories described above:

Orientation:

- User
- Assessor
- Constructor

Scoring

- Analytic approach
- Holistic approach
 - Holistic scoring
 - Primary-trait scoring
 - Multiple-trait scoring

Focus

- real world
- construct

The type of scale selected for a particular test will depend on the purpose of the test. Test developers need to be aware of the rationale behind choosing a certain type of scale for a particular testing purpose as well as to make their decisions and reasoning explicit.

II.6.1 Approaches to Rating Scale Design

There are two basic approaches to rating scale design: one is to use ‘intuitive’ methods, and the other is to base the design on some kind of empirical research or data. Intuitive methods include *expert judgement*, where an experienced teacher or language tester writes a rating scale in relation to already-existing scales; *a committee*, similar to the above, but where a group of experts are involved in a discussion of the wording of descriptors and the levels of the scale; and *experiential*, where, having begun with one of the above methods, the scale is gradually modified by those who use it to further suit their purposes. In this way, over a period of time, users develop an ‘intuitive’ understanding of the scale in relation to performance. This is the most common intuitive method of scale development.

Several methods of empirical scale development have been defined. One of these is *data-based*, or *data-driven*, *scales* requiring the analysis of performance on tasks, and the description of key features of performance that can be observed in order to make inferences about the construct. We also have *empirically derived*, *binary choice*, *boundary definition scales*, where expert judges are asked to divide samples of performance into better or poorer categories, and then to record the reasons for the categorisation. These will then be used to write a sequence of yes/no questions that lead the rater to the score. A final empirically-based method for developing rating scales is known as *scaling descriptors*, in which many band descriptors are collected in isolation from a scale and subsequently ranked by experts in order of difficulty. This sequence is then used to create the scale. These varying approaches will be discussed briefly below.

II.6.2 Intuitive Rating Scales

The historical precursor of intuitive and experiential scale development was the FSI rating scale (see Appendix 1). It became the model for the design of many other rating scales still in use today. On the Inter-agency Language Roundtable web site⁷, we can find the following account of how the ILR and ACTFL rating scales came to be developed:

...The resulting scale [of the FSI interview] became part of the United States Government Personnel Manual. The original challenge to inventory Government employees' language ability could finally be met.

New developments continued. In 1976 NATO adopted a language proficiency scale related to the 1968 document.⁸ By 1985 the U. S. document had been revised under the auspices of the Interagency Language Roundtable (ILR) to include full descriptions of the "plus" levels that had gradually been incorporated into the scoring system. (Since then, the official Government Language Skill Level Descriptions have been known as the "ILR Scale" or the "ILR Definitions"). Although specific testing tasks and procedures now differ somewhat from one agency to another for operational reasons, all U.S. Government agencies adhere to the ILR Definitions as the standard measuring stick of language proficiency.

⁷ http://www.govtilr.org/ILRscale_hist.htm

⁸ In 1968 several US government agencies cooperatively wrote formal descriptions of the base levels in four skills - speaking, listening, reading, and writing. The resulting scale became part of the United States Government Personnel Manual.

Also in the 1980s, the American Council on the Teaching of Foreign Languages (ACTFL) developed and published for academic use Proficiency Guidelines based on the ILR definitions. Like the ILR scale, the ACTFL guidelines have undergone refinement. ACTFL also developed an OPI similar to the Government test and began training educators to test according to their scale. ACTFL and the Government have worked together closely for almost twenty years to ensure that the two proficiency testing systems are complementary.

Martha Herzog (ILR)

These scales are both assessor and user-oriented and are based on a semantic differential with the intermediate levels representing a relative amount of a quality between bipolar terms. Each of the five levels in the scale ('Elementary Proficiency', 'Limited Working Proficiency', 'Minimum Professional Proficiency', 'Full Professional Proficiency' and 'Native or Bilingual Proficiency') is defined in relation to the other levels and the only key reference point or ultimate standard (ultimate criterion reference) is the proficiency of the educated native speaker.

Intuitive rating scale development has therefore depended on the concept of the native speaker for the definition of the top band, although the exact nature of this ideal is not clear. The use of the concept of 'educated native speaker' for scale development has increasingly come under attack. The most significant problem with its use is that native speakers "show considerable variation in ability" (Bachman and Savignon, 1986: 383). In reality, the variation in native speaker competence is so great that no researchers have sufficiently defined the term to make it useful in a testing context.

The challenge faced by the test designer is to develop a testing procedure that generates sufficient evidence to be scored, and a rating scale that describes the constructs to be measured. In the development of intuitive rating scales, the correspondence between the speech samples generated and the descriptors in the rating scale has not been investigated because reliable ratings depend on 'experience'. Alderson (1991b) reports that the IELTS (International English Language Testing Service) band descriptors contain descriptions of performance which are not elicited by the tasks that make up the test. Jones (1981, cited in Fulcher, 2003:95) also found that many oral testing procedures did not relate the elicitation technique to the scoring systems in any specific way.

II.6.3 Rating Scale Terminology

In rating scale design, the notion of 'development' must be theoretically coherent and empirically verifiable. The problem with the 'zero to native speaker' rating scale is that its development relies on non-validated theories of second language acquisition that correspond to the intuition and experience of the scale users and test designers and to the definition of 'competent native speaker'. With regard to the number of levels described within a rating scale, it would seem that unless it is possible to show that what is described in the bands of the rating scale refers to the way in which students acquire language and how they really speak, the rating scale and the model of acquisition of language it claims to describe will always be open to criticism. For this reason, construct definition is a vital prerequisite to scale design.

The intuitive approach to scale development results in a certain amount of vagueness and generality in the descriptors used to define the bands. Typical band descriptors in these types of scales are not normally explanatory in an independent way (they tend to rely on references to other features or bands described within the same scale and can only therefore be defined within terms of themselves), and are often not based on actual language production but on an idealised or predicted idea of a speech sample (e.g. the FSI descriptors for *Fluency*: S4 – “Speech on all professional matters as apparently effortless as in English; always easy to listen to” and S5 – “Speech at least as fluent and effortless as in English on all occasions”).

A further problem that can be perceived here is that of providing descriptions that distinguish between one level and the next, where descriptors use adverbs such as ‘often’ and ‘sometimes’ to distinguish between one band and the next. Alderson (1991b) details some of the problems arising from the revision of the assessment criteria for the ELTS (English Language Testing Service) test. One of these was the difficulty in designing descriptors which measure levels of language at regular intervals:

It was felt that a difference between, say, Bands 6 and 7 should not be noticeably larger than that between, say, Bands 4 and 5. Getting the progression right all the way through the levels has proved to be very difficult. (In Charles Alderson & Brian North (eds), 1995: 81)

It can be seen here how employing a number to describe ability is an extremely complex issue. ‘1’ is an abstract but regular concept, but describing abilities or behaviour that correspond to one unit of difference is nowhere near as simple as the superficial appearance of numerical scaling implies.

Problems were also found in the descriptors with the exact meanings of quantifiers such as *several*, *some*, *few*, *many* and *considerable*. If our aim is not that of creating an equal interval scale, the important thing is for assessors to be able to use the scales and agree on their understanding of the descriptions that define the levels. Fulcher (1994) suggests that research be carried out on the optimum length of rating scales for practical use. Rating scales which include too much detail become inoperational, while those which have too little run the risk of being misinterpreted or interpreted differently by different raters. Alderson (1995:81) states that in the process of revising the ELTS scales “[...] much of the wording was changed since it was felt that it was too complex and metalinguistic.”

It can therefore be seen that, along with attempting to define the speaking construct, a second function of descriptive assessment scales is to provide guidance for assessors who are rating performance. The language performance elicited by the task is compared with the scale descriptions which act as a common standard for the different raters and also for the same raters in different sessions of the same test. Alderson (1991b: 73) terms this the ‘assessor-oriented purpose’. He then describes a third purpose of the scales which is that of providing guidelines for test constructors, so that the texts, tasks and items in a test are of an appropriate level to that determined for the test.

The concept of ‘experience’ in creating and applying holistic and intuitive scales is observed as an important feature in perceiving them to be meaningful and to provide reliable results with reference to the terminology employed. However, the question remains as to what extent the scales are meaningful once they are separated from the training which raters must receive in order to become certified raters.

Training and socialisation may mask problems with the wording of bands in the scale by creating the illusion of psychological reality through high rater reliability.

II.6.4 Data-Based Scale Development

A different approach to rating scale development is one in which the band descriptors are designed using validated research data which is based on observed learner behaviour as opposed to postulated or normative notions of how learners will behave. This behaviour must be quantifiable in order to make the relationship between linguistic behaviour, task and scoring procedure transparent. This is known as the data-driven approach to scale development. It has a strong theoretical and empirical underpinning and the scales derived from it are usually assessor-oriented, may require holistic or multiple trait scoring, and are construct-focused.

Fulcher (2003: 98-104) reports on his design of a rating scale for fluency based on a large database of spoken learner discourse, the analysis of which provides the foundation for more meaningful band descriptors in a multi-trait scoring system. The scale differs to those based on the FSI approach in that its content rests on descriptions and explanations of discourse features based on actual test performance. There is greater descriptive detail and there are few 'more than – less than' relationships between the bands. The scales developed through his approach were proved to be reliable across tests, test tasks and raters.

II.6.5 Empirically-Derived, Binary Choice, Boundary Definition Scales (EBBs)

This type of scale, named and developed by Upshur and Turner (1995, 1999), uses primary-trait scoring and is assessor-oriented. The procedure (detailed in Fulcher,

2003: 105-6) is to rank-order speech samples, score them and then identify features that were decisive in allocating the samples to particular bands or score ranges. Therefore, they are developed using expert judgement, rather than a direct, objective analysis of performance as in data-based scales. The resulting scales make no assumption about a linear, regular or theoretical process of second language acquisition. Instead, they rely on how sample performances are sequenced and how these can be scored by asking raters to make a series of binary (yes/no) judgements that define the boundaries between score levels. They are task-specific and therefore not transferable to tasks other than the one they were originally designed for.

Upshur and Turner (1995: 10) claim that the principal difference between traditional scales and the EBBs is that instead of trying to define a mid-point for each band, the questions on an EBB describe the boundaries between the categories. This, in turn, leads to a greater simplicity in their use, as they rely on a single judgement (yes/no) in answer to a criterial question and possibly greater reliability in scoring, especially when used by teachers who have been involved in the development of a particular scale. This strength is, paradoxically, the weakest point of the EBB concept: the very specificity of the scoring procedure means that the score relates only to the test task for which it was designed and cannot be generalised to any other test task, or to any real-world task unless that task is identical in every way to the specific task that was used in the test. In summary, the approach really only lends itself to situations where a small group of individuals are working in a particular context.

II.6.6 Scaling Descriptors

Developing rating scales through scaling descriptors is associated primarily with the work of North (1995) and North and Schneider (1998) in the context of creating the Common European Framework of Reference⁹ (2004) for assessing competency in any of the European languages and with a view to creating a 'language passport' common to the entire EU. The design method is empirical and the resulting scales are both user- and assessor-oriented, using holistic scoring with a real-world focus where the final descriptors are phrased as a series of 'can do' statements. Scale developers collect large numbers of descriptors from as many existing scales as possible and re-sequence descriptors that can be calibrated onto new scales. The measurement model underlying the process is 'multi-faceted Rasch analysis'¹⁰ because, as the descriptors are not actually written by the test designer but are chosen according to the basis of their fit to the measurement model, it allows researchers to calibrate items as well as test-takers and raters on a linear scale (for a detailed summary of North's work, see Fulcher, 2003: 108-113). The resulting scale is currently in use in the Council of Europe's 'Common European Framework for Language Teaching and Learning' and is used for assessing second or foreign language proficiency across a wide geographical area with learners of different first languages who have had experience of varied educational systems.

⁹ A Council of Europe document which aims to provide a practical tool for setting clear standards to be attained at successive stages of language learning and for evaluating outcomes in an internationally comparable manner.

¹⁰ In the Rasch model, the probability of a specified response is modelled as a function of person and item parameters. For example, in educational tests, item parameters pertain to the difficulty of items while person parameters pertain to the ability or attainment level of people who are assessed: the higher a person's ability is relative to the difficulty of an item, the higher the probability of a correct response on that item will be.

The weakness of the approach, acknowledged by the authors (North and Schneider, 1998: 242), is that it is not based on any theoretical or empirically validated description of the language learning process. This is addressed by Taylor (2000): she discusses revision procedures for the Cambridge ESOL 'First Certificate in English' rating scale, taking into account advances in research in Applied Linguistics pedagogy, testing, and measurement theory. Multi-faceted Rasch theory is used in order to be able to simultaneously revise different aspects of the test that will also be affected by any changes made to the rating scales. During the revision of the existing scales, conversational and discourse analytic techniques were used to investigate samples of speaking test interviews at different proficiency levels to confirm the criterial features of test-taker performance and to further identify the features of language that distinguish different levels of performance.

Since the revision of rating scales also has an influence on test task design, it is necessary to consider whether the test tasks are capable of eliciting a broad enough sample of candidate output to be measured against the new criteria and scales. Furthermore, the revision process takes into account the role of the raters by using verbal protocol analysis¹¹ with examiners as they actually apply the revised criteria to provide insights into the theoretical and practical problems they encounter. This can address questions such as what it is that raters actually pay attention to in their rating, how they reach a final decision with regard to scoring, and whether they find certain criteria more difficult to identify and scale than others. These verbal protocol analysis

¹¹ In protocol analysis, subjects are trained to think aloud as they solve a problem, and their verbal report forms the basic data to be analyzed.

procedures can help to improve the qualities of validity, reliability, impact and practicality in the test design.

II.7 RATERS AND RATER TRAINING

We shall now turn our attention to the figure of the rater in speaking tests. Although rater training may seem to be a practical matter, it plays an important role in the theoretical considerations of interlocutor behaviour, how this may affect interaction during the test, and how the rating scales are used and interpreted in both one-to-one tests (where it is the interlocutor who also awards the score) and those tests where an interlocutor manages the interaction and an independent rater, who does not take part in the interaction, scores the candidate(s). It is quite possible that the score might be in some way affected by the interlocutor and the nature of the development of the interaction, and we may need to question whether this is inevitable in any act of communication in a speaking test, or whether the score should be treated as independent of interlocutor interaction. The traditional view of this question is that any effect on the speech sample caused by the interlocutor constitutes construct-irrelevant variance, that is, that the test score is somehow contaminated by some aspect of the testing context that is not relevant to the construct. However, from a constructionist point of view, we may regard it as inevitable and merely attempt to control some aspects of interlocutor behaviour, and also the test environment. Our speaking competence changes, depending on who we are speaking to, the physical setting and the topic, and also the socio-affective conditions. With changes in these variables, scores may also change, so the conditions under which speaking tests are conducted need to be standardized to as large an extent as possible.

II.7.1 Rater Reliability

Historically, most of the research effort into the role of the rater in the testing of speaking skills has been directed at rater reliability and is concerned with the extent to which two or more raters are capable of agreeing with each other on the score they award to the same candidate in the same test (e.g. Lunz *et al.*, 1990; Lumley and McNamara, 1995). The principle underlying the notion of inter-rater reliability is that it should not matter to the test taker which rater they have in a test; they should be awarded the same score irrespective of who is rating their performance.

From studies of inter-rater reliability (e.g. Mullen, 1980) it has been argued that two raters are required for any speaking test, as individual raters tend to have different patterns of rating. Most other investigations of reliability in oral tests also recommend the use of at least two raters in order to avoid the possible impact that a single rater may have on a test score. The published evidence on inter-rater reliability suggests that high correlation coefficients are achieved when multiple trained raters are used to score performances (Shohamy 1983a; Shohamy 1983b; Shohamy *et al.* 1986). The correlation coefficients decrease dramatically when untrained raters are used (Barnwell, 1989). However, although there is often substantial agreement in the sequencing of test takers' performances by raters, and hence high correlation coefficients, the severity of raters differs widely (Bejar, 1985). Awarding the mean score of multiple raters is therefore suggested as a correction device for varying rater severity.

Rater reliability is also concerned with intra-rater reliability, or the extent to which the same rater awards the same score to the same performance over time. Upshur and Turner (1999) found that while teacher-raters were not equally severe,

they were individually consistent in their scoring and interpretation of the rating scales. These authors also addressed the question of whether teacher-raters were biased with respect to test-tasks, finding that although some were, it was not a significant number. They also found that teachers who had been scale constructors rated differently from other raters in that they were stricter markers. It seems that raters are more lenient when they do not fully understand the scale or have not internalised it.

II.7.2 Rater Training

Studies of rater training have shown that training reduces random error in rating, but that it is not possible to completely eliminate the differences in severity between raters (Weigle, 1994, cited in Fulcher, 2003: 142). A second reason for rater training is to try to standardise raters' interpretations of the rating scale. The meaning of a score on a speaking test is contained in the descriptor that defines the level or band. Although training reduces rater-related score variance, the fact that raters differ both in their perceptions of what it is they are rating and how, has been long recognized. McNamara (1996: 218-22) shows that in the 'Occupational English Test'¹² raters are paying attention to grammatical accuracy rather than the communicative criteria embodied in the scale descriptors. Orr (2002) uses rater 'think-aloud' protocols to attempt to identify the frequency with which raters for Cambridge ESOL 'First Certificate in English' make judgements on the basis of criteria that are not contained in the rating scale. He concludes that raters frequently make judgements on

¹² The OET is an Australian language proficiency test for overseas qualified medical and health professionals whose first language is not English. It assesses English language competency as used in medical and health professions in Australia and is adaptable as a language competency testing model for other professions.

the basis of their own personal constructs rather than the criteria contained in the rating scale.

Recently the question has arisen as to whether these perceptions should somehow be included in the scale descriptors. This approach is taken by Upshur and Turner (1995), where the rating scales were constructed *after* the test had taken place, firstly by rank ordering the speech samples and then using raters' perceptions of them to construct verbal descriptors for the scale. The authors claim their findings show important qualitative differences in the salient characteristics of the discourse produced in different tasks and that systematic effects of tasks on performance test scores cannot be ignored. For this reason, they conclude that rating scales should be task-specific.

However, it is possible that this approach may only compound the problem of variable perception; in an attempt to reflect as many variables as possible in the descriptors, the method could simply produce a scale with no meaning for anyone. However, if variable perception is used to represent types of variation that are likely to occur, it may be possible to reduce the impact of the perceptions in the interpretation of a rating scale more effectively, and these may, therefore, be useful as information to support rater training programmes.

The process of rater training is designed to 'socialise' raters into a common understanding of the scale descriptors and train them to apply these consistently in operational speaking tests (Alderson *et al.*, 1995: 108). Variation in severity between raters and the change in severity of individual raters over time have both been identified as a cause for concern: McNamara (1996: 237-8) found that changes may occur over a period of 18 months. Re-certification for high-stakes speaking tests usually takes place at one or two-year intervals, but further research is required before

it is possible to recommend fixed times for re-training and re-certification on a sound empirical basis.

The problem of the relationship between rater training and the development of a validity argument for a speaking test also needs to be addressed. Fulcher (2003: 146) points out that:

Rater training *presumes* that the rating scale designed for the test is valid and that training raters to use it in operational testing situations is legitimated. 'Valid' here means that the rating scale has been constructed in some principled way and that the level descriptors are meaningful for the purpose of the test, so that the inferences that score users make from the scores can be justified theoretically and empirically. Part of the validity argument will require using individuals to make judgements about sets of performances on sample tasks, and the consistency with which they are able to make those judgements contributes to validity evidence. If raters are trained, or 'socialised', before the validity argument is constructed, the training itself becomes a facet of the test that cannot be separated from the construct. This fusion contaminates any validity evidence that uses scores from these raters.

Fulcher's point is that raters should not be trained until *after* a validity argument has been constructed by the test designers themselves, which was the procedure used in the design of the 'Common European Framework' (North, 1995). The same designer/raters that have been involved in the construction of the test cannot be a reliable source for evaluating or claiming its validity, or the quality of the descriptors; this can only be proved by the training and subsequent consistent application of the

rating scale by raters from outside this process, thus showing that the scales and descriptors are in fact a practical, understandable and reliable guide for those involved in the evaluation of speaking performance in test situations.

IL7.3 Interlocutor Training

With our understanding of the interactive nature of discourse and the observation that some types of interaction or tasks in speaking tests are not ‘natural’ conversation, interest has grown in the role of the interlocutor (the examiner who manages the test and directly engages in interaction with test-takers, unlike the rater who does not participate). Interlocutors may differ along a range of parameters that introduce variation into the speaking test, and which may affect scores (Brown, 2003). Such variation in a speaking test is a confounding variable that reduces our confidence in the inferences we draw from the scores. Cambridge ESOL has addressed this problem by introducing interlocutor frames for all its speaking tests in order to reduce the difference between the spoken contributions of different interlocutors. The frames provide everything the interlocutor should say during the test, including greetings, instructions to candidates, back-up questions to be used in the case of a breakdown in communication, and ending the test. Interlocutors are required not to deviate from the frame at any point during the test.

Most of the work on the impact of the interlocutor has involved discourse analysis of interview type tests, primarily looking at how the interlocutor accommodates to the level of the test taker (Ross, 1992, 1998). This research on accommodation has shown that interlocutors vary their speech corresponding to a range of strategies that support the test taker. Ross recommends that interlocutor

training should include the study of various forms of accommodation and support, so that unnecessary support that may affect scores is not provided for the test taker.

The variability in support provided by an interlocutor means that the test-taking experience is frequently very different for each test taker. Lazaraton (1996: 19) identified a number of distinct types of interlocutor support in the CASE test (Cambridge Assessment of Spoken English): (i) priming new topics before they are raised; (ii) supplying vocabulary or helping to complete a test taker's turn; (iii) giving evaluative responses (e.g. the 'good' typical of teacher-talk); (iv) echoing or correcting test-taker talk (modelling language for the test taker); (v) repeating questions more slowly with over-articulation to improve test-taker understanding; (vi) reformulating questions as statements that only require the test taker to confirm or disconfirm the statement; (vii) drawing conclusions for the test-taker based on an answer they have provided to a question; and (viii) rephrasing questions to help test-taker understanding. She concludes that rater training should focus on the variable aspects of interlocutor support to make the interactions more similar, since failure to take this type of variability into account is a direct threat to score interpretation and fairness to all test takers.

The most important aspect of the discussion of interlocutor behaviour which needs to be considered for training, as pointed out by Lazaraton (1996), is that there is evidence for variability within performance that may have a variable impact on test-taker performance and scores. If this is defined as part of the construct, the variability needs to be understood so that it does not become a built-in feature of unfairness to test takers. If it is not defined as part of the speaking construct, it needs to be controlled

through interlocutor frames that limit the possibilities for construct-irrelevant variability to affect test scores.

II.8 TEST ADMINISTRATION

Test-taker performance may also be affected by the way in which the test is administered and the environment in which it is carried out. These factors are considered in the following sections.

II.8.1 Administration

Test administration is a largely unresearched area, with anecdotal evidence and intuition playing an important role in what is felt to be the correct and efficient running of tests. Generally, the test will run more smoothly and with an air of professionalism if everything that is required for it to take place is prepared well in advanced, with clear indications for candidates of times and rooms to avoid confusion.

In speaking tests, the presence of an usher who can avoid contact between candidates who have just taken the test and those who are about to take it is often essential, especially where sets of materials are limited and the number of candidates in any one testing session large. Ushers need to be coached to behave in a friendly but professional manner in order not to negatively influence test takers prior to taking the test.

II.8.2 Environment

For all types of speaking test it is recommended that the physical conditions should include good lighting, the absence of strong smells, appropriate heating or

cooling, and protection from distracting noise. Noise may come from an outside source, such as a street or corridor, or even be generated by other test takers in the case of computer generated tests or other semi-direct speaking tests being taken simultaneously by candidates in the same room, all speaking into microphones.

The arrangement of the furniture is also important in terms of providing as relaxed and non-threatening an atmosphere as possible. In cases where there are both an interlocutor and a rater and the latter is not involved in the interaction, he or she should be conveniently placed within the field of vision of the candidates, but not so as to have an intrusive influence on them. If the test is being taken in pairs or a group, with candidates being required to speak to one another, then the chairs should be placed so that they can see each other and the interlocutor throughout the interaction.

II.8.3 Test Accommodations

Test accommodation is defined by the American Education Research Association (AERA, 1999, cited in Fulcher, 2002) as the modifications or adaptations of the testing situation or the materials necessary 'to minimize the impact of test-taker attributes that are not relevant to the construct that is the primary focus of the assessment'. This means that accommodations should be made for candidates with disabilities in order to avoid the disability affecting the test score. The provision of accommodations is therefore driven by a concern with validity, and would therefore ideally be based on an empirically-founded link between the disability and the accommodation. Often, legislation is in place requiring accommodations to be applied to candidates with disabilities, but it is not usually specific as to exactly how these should be carried out. While there are no definitive solutions to these issues, it is clear

that anyone involved in applying tests, producing scores and subsequently using those scores for decision-making cannot avoid the ethical and fairness issues that surround some of the most practical aspects of test administration.

IL9 CONCLUSIONS

In this chapter we have given a brief overview of the history as well as the current state of research concerning the design of speaking tests, their construct definition and validity, and their corresponding rating scales, as well as considering how these may be implemented and interpreted. In the next chapter we will describe the research methodology used in the current investigation project to compare the performance and scores of the same learners on two different kinds of oral test: the traditional individual proficiency interview and the group oral test. In doing so, we will also attempt to provide a rationale based on the discussion of research in the current chapter for the design of our group oral test which we intend to implement as an effective and practical method for testing speaking skills in our own university teaching and learning context.

III. RESEARCH METHODOLOGY

The following chapter will provide a comprehensive account of the design procedure and methodology for the research project that constitutes the basis of this dissertation. The motivation for the current investigation was produced by a number of shortcomings perceived in the test procedure being used to test oral competence in the subject *Lengua BII*, a second year English as a foreign language component of the degree in 'Translation and Interpreting' (Faculty of Translation and Interpreting) at the University of Las Palmas de Gran Canaria. These preoccupations referred both to the concepts of validity and reliability, as we have seen in the previous chapter, two major interdependent areas of importance in testing, and also to concerns relating to the compatibility of the test with the syllabus being taught over the year.

Several questions needed to be addressed with regard to the validity and reliability of the oral interview as an adequate means for assessing our students' oral competence. The most obvious one was that the speaking construct was formally defined nowhere, nor was there a clear description of the level of competence required to pass the examination. Without these guidelines, we cannot be certain that all candidates are being assessed in the same way, since we are relying solely on internalised examiner beliefs or impressions of what speaking entails and a subjective idea of how well interview candidates perform. These may easily have more to do with socio-affective factors, such as empathy with a candidate's viewpoint or familiarity with them as a student, than with their use of language or level of competence.

A further concern with the interview procedure was that there was no formal or standardized structure to the test: each candidate was asked arbitrary questions, using

non-standardised wording and the examination proceeded according to the responses given. It could be argued that, in some ways, this replicates authentic conversation (although there are many types of ‘conversation’, defined by characteristics such as setting, knowledge of topic, or social and power structures), but a conversation, by definition, cannot be a test. In order for a test to qualify as such, it must have certain features that mean that it can be scored in a standardised way (a definition of the construct to be measured and a scoring rubric), it can be replicated with other candidates taking the same test (there are test tasks which may vary in content, but that can remain constant in procedure), and also that its results can be generalized from the particular instance of the test itself to the whole area of competence involved (a description of the way in which the construct definition refers to authentic language use and underlying ability). As we have seen, without these features, a test cannot be valid. This is not to say, however, that a test may not attempt to reproduce some of the features of authentic language use, but its primary function will always be as a measurement tool and, as such, there will be certain constraints placed on its authenticity as an instance of language use. Whether or not a test is an authentic *test* (it is coherently related to the syllabus in such a way as to efficiently measure progress and learning through it) is, as we have seen above, a different issue.

Another major area of concern was the scoring procedure used to assess student performance. The University of Las Palmas de Gran Canaria¹, like practically all educational institutions in Spain, uses a universal 0 – 10 marking scale as the only possible way of grading students within the official administrative system, so that all students, whether they study Modern Languages, Medicine, Law or Marine Science,

¹ From here, referred to as ‘ULPGC’.

will receive a score according to the loosely defined scale: 0 – 4.9 = *Fail*; 5 – 6.9 = *Pass*; 7 – 8.9 = *Very Good*; 9 – 9.9 = *Excellent*; 10 = *Honours* (meaning free registration for the follow-on subject). It is possible to see here that the range of marks is unevenly balanced across the scale, with the first part (a possible 50 scores) indicating only that the required level has not been achieved. *Pass* and *Very Good* have a mere 20 scores each in comparison, with *Excellent* only 10 possible marks and *Honours* just one score. These are the only definitions of the scores provided by the University itself and those charged with the responsibility of awarding the scores are therefore left to devise their own means of implementation which are necessarily based on internalised criteria and personal interpretation of their meaning.

The interpretation of the score 10 is interesting; whilst a score of 10 can be achieved in some kind of straightforward (and totally objective) Maths test, it can be argued that perfection to such a degree is simply an idealized and unattainable goal within university education, which is concerned with progress, discovery, the critical analysis and questioning of new ideas (where teachers/examiners may or may not agree with student opinion), and the exploration of what is already known with reference to how this may have a bearing on future study. With so many factors still unknown, how can it ever be possible to achieve an absolute 10? In a language-based subject this may be perceived to be even more difficult. Others, however, may see the accurate reproduction of what they have delivered as input to their students during the programme of study as a quantifiable entity, and therefore have no problem in awarding numerical scores that indicate the percentage of correct information that has been relayed back to them in assignments, tests or examinations. In this case, a 10 is a perfectly acceptable score. Still others may take the view that examination candidates

need to be judged according to what can reasonably be expected of them in a given situation and at a certain, previously defined level. In this case too, a 10 is a possible, although unusual and outstanding score. The question these examiners ask themselves is whether anything more can be expected or asked of the student at the level and in the circumstances under which they are being examined.

Of further interest would be a study of just how individual teachers throughout the ULPGC use the marks. In *Lengua BII* we only use the whole scores and .5s, while in the subjects *Lengua A*, *Lengua BI* and *Lengua BIII*, the whole range of scores is employed. The lowest mark ever given in *Lengua BII* is 3.5 (and this is extremely unusual), with 4 being a more frequent score where a candidate has failed to reach the required standard, since it is felt that this is a sufficient indication of non-achievement which has the same essential effect as the lower scores. Teachers of other subjects in the 'Translation and Interpreting' degree, however, award official scores such as 1.8 or 2.3 since they feel them to be a true indicator of a candidate's attainment and do not wish to imply that with only a little more effort and slight improvement, a student will be able to reach the required standard. For these teachers/examiners, very low scores signal that extensive further study and understanding are necessary before students will be able to pass a subject.

From these general patterns it is easy to see that members of the ULPGC teaching staff even within the same faculty (Translation and Interpreting) interpret and apply the scale according to their own criteria and that, in fact, universally the scores cannot therefore hold an inherent, objective meaning. A further irony is that students will be given an average mark for their whole degree, which may contain between 30 and 40 subjects in total, arrived at from the total scores awarded to them by 30 or 40

different members of teaching staff, each having created their own personal interpretation of the marking scale. The validity of an average based on scores which are only ostensibly elements of a single system is questionable, but how exactly to go about a unification of the system to give it greater authenticity and reliability requires extensive research far beyond the scope of this dissertation. The essential point here is our need to doubt, question, and reflect upon the actions we take in awarding scores that affect the lives of others in what may sometimes be major ways. This should, at least, give us a heightened sense of responsibility for these actions and dispel our false sense of security in a system which is widely accepted by all its users and is traditional and time-worn, but which, for all these qualities, may not necessarily be the most valid and reliable.

III.1 RESEARCH QUESTIONS

Our study therefore aims to compare the traditional method of oral assessment, the one-to-one 'Individual Oral Proficiency Interview', which has been carried out to date in the Faculty of Translation and Interpreting at the University of Las Palmas de Gran Canaria, with a new type of oral test where the students are examined in a group with other candidates taking the same test at the same level, that is, the 'Group Speaking Test'. We will focus principally on (i) *test procedure*, in particular on the different skills and amount of attention required by the interviewer to manage the one-to-one interview and the group test and, from the students' perspective, on the socio-affective aspects involved in taking the two different tests; (ii) the use (or absence of) descriptive *rating scales* and the subsequent objectivity and reliability of the scores, as well as the usefulness to the students themselves of the marks obtained. In relation to

both test procedure and rating, we will also consider the value of introducing a rater to the test along with an interlocutor, whose role is exclusively to assess candidates' performance, and who takes no part in managing the interaction.

In order to bring the project into line with current trends in language testing and assessment, a third perspective, that of (iii) *self-assessment* was introduced. Self-assessment may play an important and fundamental role in learning, driving students' motivation, helping them to become aware of their strengths and weaknesses, and raising confidence and self-esteem. The Council of Europe is in the initial stages of implementing the 'Common European Framework of Reference for Languages' and the 'European Language Portfolio' (www.coe.int) which involves individuals in assessing their own capabilities in a foreign language. In line with these developments in Europe, our experiment included a self-assessment procedure in order to appraise just how far students' conceptions of their own abilities coincide with external, 'objective' assessment. If students' perception of how well they perform on speaking tasks is similar to, or not significantly different from, those of the external observers, then there may be a case for taking their self-assessment into consideration when awarding an overall grade for the subject *Lengua BII*. This may help to compensate for students who, for whatever reason, perform to a lower level than expected on the day of the exam, or who are simply not good test-takers, and should also provide motivation and confidence which are essential factors in successful speaking performance.

In order to become familiar with the criteria for assessment, the subjects of the study were requested to give a self-assessment of their speaking ability in English in general terms before taking either of the two tests included in this project, and then to

assess their own performance on each of the tests they took immediately after the test was carried out. For the purposes of our research, these self-perceptions will be contrasted with the students' external scores on the different tests in our analysis of the data in Chapter 4. Depending on the degree of correlation between the scores, the students' general self-assessment and the one carried out following the 'Group Speaking Test' may also have a bearing on the student's final mark in the subject *Lengua BII*. (The mark received for the 'Individual Oral Proficiency Interview' will not be taken into account here).

Our research questions thus cover three broad areas which we will attempt to look at from the two different perspectives of the participants involved in the speaking test: the candidates and the examiners. These are *test format* (how does the test format affect the test taking experience for both examiners and candidates?); *scoring procedures and rating scales* (how do interviewers/raters award scores and how are these interpreted by students?); and *self-assessment* (how similar are students' perceptions of their own ability and performance to those of the examiners and can it play a role in the learning process?).

III.1.1 Test Format

Our first area of interest concerns the characteristics of the test format from both the perspectives of the candidates and the examiners. From the perspective of the students, we wish to explore the socio-affective aspects of the test-taking experience. We are particularly concerned with anxiety and whether the different test formats have an effect on increasing or lowering student anxiety, either due to the power structures

generated in the test situation, or to the similarity (or lack of it) between test tasks and classroom tasks. The principal research questions here are:

1. Does taking a speaking test in a group reduce the anxiety inherent to speaking tests in general and, if so, is anxiety lower than in a one-to-one interview situation?
2. Does familiarity with the task and/or test type have a bearing on performance?
3. Do students feel that the test format for both types of tests allows them to demonstrate their speaking ability?

From the perspective of the examiner, we wish to consider some of the features of test management, particularly the difficulties involved in simultaneous interview management and rating procedures, and whether raters feel they can give more objective and accurate scores when they are not directly involved in the interaction. Principally, here we focus on the following questions:

1. Do examiners feel that they can manage test materials and interaction, as well as give accurate and objective scores for candidates' speaking performance at the same time?
2. Is managing a test with three students easier than managing and simultaneously scoring an individual test?
3. Does the test format influence the size of the speech sample produced by candidates, either facilitating or hindering assessment?

III.1.2 Scoring Procedures and Rating Scales

Our second area of interest lies in the scoring of speaking tests. We will address questions of how raters interpret, and subsequently use, a traditional 0 – 10 marking

scale which does not include definitions of level or descriptors that assign a meaning to the mark awarded, and contrast this with the introduction of a rating scale that describes the features that it attempts to measure and defines a reduced number of scores (0 – 5). We will try to determine some of the features that examiners focus on when scoring candidates for speaking tests in an attempt to discover whether they use norm-referenced or criterion-referenced procedures. Our main questions here are:

1. Do examiners feel more confident in awarding scores when using a descriptive scoring scale than when using a traditional 0 – 10 scale?
2. How do they interpret the meaning of these two types of rating scale?
3. Do they focus on a wider range of features of speaking when using a descriptive scoring scale?

From the students' perspective, we will investigate the extent to which they understand their test scores, how they interpret them, and whether these scores have any pedagogical value beyond indicating to them how they measure up against the other students who took the course and the test at the same time as them. We will attempt to answer questions such as:

1. Is an analytic score, which relates to a set of descriptors, more meaningful than a mark received on the traditional 0 – 10 scoring system?
2. Do analytic scores indicate areas of strength and weakness to students, and hence have a pedagogical value?

III.1.3 Self-Assessment

Finally, we will examine the role of self-assessment. From the students' point of view, we are interested in how accurately they can perceive the success of their own

performances and also in the potential effects of self-assessment on learning. We will investigate student opinion on the following questions:

1. How useful is self-assessment in learning and making progress?
2. Should self-assessment be taken into account as part of the final mark for the subject *Lengua BII*?
3. Can self-assessment give an accurate reflection of speaking ability?

From the perspective of the teacher/examiner we are more concerned with attitudes towards including self-assessment in learning programmes and with how this may be valued within the formal testing structure. Teacher/examiners were asked to express an opinion on:

1. Should self-assessment be incorporated into our teaching programmes and testing procedures?
2. How accurate can students be in their self-assessment?
3. Can self-assessment be useful in helping students to improve their language skills?

After eliciting the attitudes of our participants towards self-assessment by means of questionnaires, we will attempt to discover whether there is any empirical evidence which indicates that self-assessment should be included in our teaching, learning, and testing programmes. A possible indicator of this will be a comparison of the scores assigned on the tests by the interviewer/raters and by the students. If there is a significant difference between them, this would indicate that students are not really aware of their own level of speaking performance, while a similarity in the scores awarded would mean that students have a reasonable or good perception of their ability in speaking. We will therefore try to answer the question:

1. Is there any empirical evidence to support an argument for introducing self-assessment into our study programme for the subject *Lengua BII*?

The exploration of the theories set out in the previous chapter has shown that there is no single best test, or best test method, and the design and development of a test will therefore take into account the specific context of the test, who it will be used by, and for what purpose. A possible starting point for the design process of a new test therefore, is a description of the target test population and the situational context of their language learning and use.

III.2 PARTICIPATING SUBJECTS

The subjects of this study were our own students at the Faculty of Translation and Interpreting at the University of Las Palmas de Gran Canaria (ULPGC) registered in the subject *Lengua BII (inglés)* in the academic year 2003/04. A total of 152 students were registered during this period, but it was not expected that all of them would participate in the study, since the total number of students registered never complete the course (the second test was part of the final exam), and the first test was optional (presented as part of our study but with the incentive of the opportunity for a practice test before the final exam).

III.2.1 Learning Context

Below, we will examine the educational context and profile of our students and the major factors which influence teaching and learning in our context. Based on the results of our study, we will subsequently propose and implement a change in oral

testing which we consider to be educationally coherent and administratively feasible to operate in this learning/teaching/testing situation.

III.2.2 The Global Context

The students who make up the sample for the current study are in the second year of a four-year degree in ‘Translating and Interpreting’, whose core content is dictated by the Spanish Ministry of Education. That is to say, the Ministry provides a title and one-line description of the core subjects (*asignaturas troncales*) which are taught in all Spanish faculties offering this degree, and the individual teachers responsible for these subjects draw up their syllabuses with whatever they feel is appropriate content. The Faculty of Translation and Interpreting at the ULPGC then decides which subjects are compulsory for the degree in their own institution (*asignaturas obligatorias*); *Lengua BII (inglés)*, the subject which this study addresses, falls into this category. Students must also complete a required number of credits offered as optional courses (*asignaturas optativas*) where they can ostensibly choose from a range of subjects related to their chosen degree, although in practice financial constraints dictate that there is little or no choice, since there are not sufficient resources to run courses for a small number of students (an average class size in our faculty is 45 students, which is considered very small by Spanish university standards), so students may be forced to take a particular elective subject in order to complete the necessary number of credits. The final category of subjects known as ‘free credits’ (*libre configuración*) have, as a requirement, nothing in common with the principal content of the degree itself. These are often taken from courses run by other faculties and departments for their own students, but may also be short courses run by outside

institutions such as art galleries or cultural organisations. The idea behind this is to give students a broader general knowledge and an insight into other areas which might be of interest to them, but in practice it just contributes to further overloading them with class hours.

The degree has a total of 300 credits, with approximately 76 corresponding to the second year. Currently, one credit is equal to ten hours of study, but up until now there seems to have been a misinterpretation of what these ten hours correspond to. We have been operating a system where students are expected to attend ten hours of classes per credit, with additional time being spent on private study and individual course work, rather than reducing contact hours.

Some specific faculties in Spanish universities are currently involved in the pilot scheme for adaptation of their study programmes to the European Area of Higher Education (the Faculty of Translation and Interpreting is one of these). We are in the process of revising our curricula in order to adapt them to a common European purpose, and a change of direction is taking place, with the emphasis shifting from teaching to learning, from information to formation and with a more practical, work-oriented stance being introduced. In line with university education reforms across Europe, we are likely to see the disappearance of annual subjects and the introduction of semester-long modules which will facilitate student mobility across its institutions and the incorporation of study time into the credit system, so that credits do not correspond directly to the number of taught hours in any one module.

III.2.3 The Micro-Context

Currently, our students may be in class for 4 or 5 hours every weekday, and also be expected to produce extensive work at home for assessment in all subjects (which may total around nine), as well as sit final examinations at the end of the course. The degree in 'Translation and Interpreting' is based on a minimum of three languages, Language A (Spanish) and Languages B and C (English, French or German in any combination, with Russian as an option for C only). Students must have passed Language B at the school-leaving examination stage and also have passed an entrance test to the faculty in that language. Language C may be started from scratch.

The students that make up the sample for the present study are taking English as their first foreign language, *Lengua B* in the second year. In the first year they will have had 5 contact hours per week. In the second year there are 4 hours per week of English Language, which is then reduced to four hours for only one semester in the third year, and none at all in the final year. Parallel to this, is intensive input of the second foreign language, with 6 hours in the first year, 6 hours in the second year, and 4 in the third year.

The vast majority of students who enter the faculty do so with English as their first foreign language, since it still seems to be unusual for Spanish state schools to offer French or German as a first foreign language. A typical First Year intake for the FTI would be 10-12 students in German (from the Deutsche Schule), 4-5 in French and 120 in English. This is a just reflection of the status which the English language enjoys in Spain, despite the fact that the outlets for German speakers are equal to, if not greater than, those for English speakers in the Canary Islands due to the tourist industry.

III.2.4 The Learners

In Spanish state schools, children are starting instruction in English at an increasingly early age during the first years of primary education. However, very few qualified primary school teachers have specialised in English, and consequently the progress made between the ages of 5 and 12 is not very significant. By the time students reach university, they have been studying English for at least ten years but have failed to consolidate basic structures and have a very limited vocabulary. In the final three years of school, they have used successively lower-intermediate, intermediate and upper-intermediate text books (such as Oxford University Press's *Headway*) which are mostly higher than the level the majority of pupils have achieved in previous courses. Although most schools use up-to-date text books, they often adapt them to their more traditional methods of teaching, translating the dialogues, making little or no use of the accompanying CDs or cassettes, or giving de-contextualised vocabulary tests on reading texts.

The fact that upper-intermediate text books have been used in the final year of school means that in the first year at university only an advanced level course book can give the impression that university level is higher than that expected at school. These texts are generally beyond the level of the majority of students, and tend to contribute further to the failure to incorporate the new language into their language use, so they continue to employ the few structures they have already automatised, using many incorrect fossilised forms (e.g. '*I am agree' '*What means ... ?').

There is also a major change in the teaching methods employed at university, so that emphasis shifts from accuracy to fluency and students are asked to carry out communicative tasks in groups to develop their speaking skills, which mostly they

have only had limited controlled practice in until now. In our own faculty, First Year teachers are currently in the process of writing their own course manual which will hopefully begin to remedy this situation by gradually introducing communicative tasks and requiring students to focus on grammar through exploring different text types which, at the same time, will serve as an introduction to, and orientation towards, translation skills.

It should, however, be noted that changes are beginning to take place in the Spanish state education system where several recent studies have investigated the speaking skills of primary and secondary school students with a view to introducing a compulsory foreign language speaking test to the current school-leaving/university entrance examinations (PAU) in 2008. At present, these studies have been limited to assessing whether students answer correctly, partially correctly or incorrectly a series of questions divided into three broad categories: (i) giving personal information, (ii) responding to a visual prompt, and (iii) expressing a personal opinion (ICEC, 1998; INCE, 2002; INECSE, 2004). A corpus of speech produced on such a test in the Canary Islands has recently been published by Wood *et.al.* (2007).

Parallel to the change our students currently experience in the approach to teaching in their English classes in *Lengua BI*, they begin an intensive second foreign language course from scratch in which progress seems to them to be very rapid, contrasting markedly with what they see as stagnation in English. The absence of prescriptive grammar classes in the first year adds to their perception of failing to advance in English, and although the majority of students do undoubtedly make progress, by the time they reach the second year many are far more motivated in their

second foreign language than in English, principally because progress there is so much more easily quantifiable.

In the second year, we try to maintain the fluency which has been gained over the previous year both in speaking and in writing in the subject *Lengua BII*. Learners continue to work on communicative tasks in groups and also start writing a dialogue journal with their teacher which, for those who carry it on through the whole year, has proved to be a motivating experience. It is a point of individual contact in large classes and an excellent way to get to know the students better. It also humanises the teacher and creates a good working atmosphere in the classroom. At the same time, we reintroduce a focus on accuracy in language production with one class each week being devoted to grammar revision and extension. Whether or not this contributes to the language learning process is uncertain, but it does give the learners the sensation of receiving new input in the form of ‘factual information’, which has been an essential part of their learning experience for thirteen years (other university subjects continue in this mode) and which they perceive to be an addition to their knowledge of the language.

III.2.5 The Influence of Translation

In the second year, the students that are the focus of our study are also presented for the first time with a subject in which they translate texts from Spanish into English. This is taught for 3 hours a week over the whole year. We believe that full advantage should be made of this time to reinforce the language learning that takes place in *Lengua BII*, by placing emphasis on flexibility of expression and grammatical accuracy, as well as practising translation skills. In the recent past there has been a

general feeling amongst ESL theorists and practitioners that translation should be avoided because it leads to more errors caused by transference of L1 structures into the second language. However, this will depend on the focus that is given in the classroom and we have found that translation can, in fact, be used positively in the language learning process.

The 'Grammar Translation' approach to language teaching was not based on any kind of theory, either of language or of learning, but on the traditional approach to the teaching of classical languages (Richards and Rodgers, 1986: 5). It aimed at formal accuracy and the type of mental gymnastics required to match the grammatical structure of one sentence to another in the target language, with little or no attention paid to meaning other than the 'symbolic meaning' (Widdowson, 1990: 82), and leading to meaningless sentences of 'the nose of my aunt' type so extensively quoted in the literature. However, this cannot really be classed as 'translation' in the modern sense of the word, since the learner here is engaged only in the mechanics of reproducing structures and lexical items in the target language in decontextualised sentences which lack a communicative or interactional situation. As soon as a context is provided, meaning varies and hence also the rendering of an utterance in another language. Form ceases to be the sole or principal aim and meaning becomes paramount.

Translation requires learners to be able to focus on meaning from two points of view, that of the writer of the source text and of the reader of the target text, and to act as a mediator in the expression of that meaning. As such, it involves interaction and negotiation of meaning in several directions and is thus an eminently communicative activity. It also lends itself easily to the demands of a monolingual classroom, since it

does not need suspension of disbelief (such as other learning activities like role-play) to be seen as a valid activity. If students can be encouraged to see translation not as a system of one-to-one correspondence between two languages, but as a demanding communicative skill depending on contexts and cultures, they may become more aware of the differences in structure between the L1 and the L2.

The idea of translation as an ability which necessitates linguistic precision means that the classroom process must focus on illocutionary force as it is expressed through form. Students are required to 'notice' structures as they focus on the text and this may then become a language learning strategy for them, facilitating the process of restructuring, proceduralisation, and the eventual automatising of correct language (Batstone, 1994: 45).

The incorporation of parallel texts (texts in the source or target language that are comparable to the text to be translated in terms of subject matter or text type) in this course helps to provide the learners with cultural input, not only in terms of physical location but also of its expression through language by drawing attention to style, register, structure and lexis. It is in this light that we would like to see the influence of the translation subject on our learners in the subject *Lengua BII*, not only as a skills training process, but also as a positive and key element in the language learning which takes place in the Faculty of Translating and Interpreting.

III.2.6 Teacher and Learner Expectations

Both classroom interaction and the type of learning students undertake are influenced by the expectations of the teacher and the learners themselves. In the present context, the teachers expect learners to participate actively in classroom tasks,

to communicate with each other in English in the classroom even though they all have Spanish as their L1, and to take risks in order to extend their productive use of the language. This is in contrast to the learner expectations, built up over more than ten years of learning experience in all kinds of subjects where appropriate classroom behaviour is to sit passively and be given input which can then be memorised and reproduced in tests and exams. The students also expect to be grammatically accurate when they speak, which leads to a very low incidence of risk-taking and a maintenance of automatised, often fossilised, language learnt several years previously at school. These learners would also expect to be evaluated on written grammatical accuracy, but not on cohesion, coherence, or discourse structure and they are mostly unaware of the importance or relevance of these.

The teacher's task then, in the second year subject Lengua BII, is to set up classroom tasks and activities in such a way that learners are motivated to speak in English and are provided with opportunities for 'noticing' and focusing on new structures, in the hope that they will restructure and eventually proceduralise the new (and revised) forms so that they become available for production and automatisation. At the beginning of the course, we concentrate on strategy training and raising awareness so that learners may become more independent and responsible for their own learning. This attempt to shift the focus from the teacher to the learner as the central element of the learning process is an essential part of the attempt to change the learners' perception of their classroom role and hence their expectations of the course. It can lead to greater participation in classroom activities and motivation to speak in the foreign language, and hence a greater improvement in fluency and often in accuracy as well.

Parallel to more traditional composition writing, students are also encouraged to improve their writing skills by working on the dialogue journal (mentioned above) in which communication is exchanged with the teacher on a weekly basis and which can help to reduce anxiety in using English orally in the classroom. Here, learners write about anything they like and the teacher responds by commenting on perhaps one or two of the language mistakes which have been made in the writing and then to the message itself, trying to work into the reply new language which is relevant to the ideas the learner wishes to express. This can act as comprehensible input which the learners 'notice' and begin to use immediately, since it has direct bearing on what they wish to communicate. In turn, the dialogue journal may have a positive influence on the learners' speaking skills, since it builds confidence by showing that what they have to say is important and can be responded to as a communicative act and not only as piece of text for teacher correction.

It has generally been observed that learners do become much more motivated to practice their oral skills during the year, and that the consequent gain in confidence leads to a more relaxed and positive classroom atmosphere. Although further research is required to establish whether or not this actually improves language acquisition, one would expect that, at the very least, it makes for a more enjoyable learning experience for both the teacher and the learners.

The introduction of self-assessment as both a tool for enhancing learning and for awarding grades is a new concept in our teaching situation, and one with which students and teachers are unfamiliar. Further training is likely to be required for students to be able to view it as a valuable exercise, and as teachers we need to develop a more constant and wide-reaching approach to interpreting students' evaluation of

themselves as well as better ways to record and make use of their perceptions and judgements. The journals mentioned above may also prove be an aid in developing self-monitoring and metacognitive self-assessment strategies of this kind.

III.3 TESTING PROCEDURE

Having provided an account of the subjects and context relevant to our study, we shall now proceed to describe the methodology for our research project. In order to contrast the validity and reliability of the two types of oral test, the 'Individual Oral Proficiency Interview' and the 'Group Speaking Test', and also their socio-affective implications for students, we initiated a study based on assessing the students' speaking performance in the two kinds of test format within a fairly short time interval (approximately six weeks) in order for there to be relatively little or even no change in the external manifestation of their speaking ability. After each test, the candidates were subsequently asked to fill in a self-assessment sheet (see Appendix 11) about their own perceptions of how well they had performed on the test and, after receiving their results, a questionnaire (see Appendix 5) about the test itself and their experience of it. Prior to this, students had filled in the same self-assessment sheet reflecting their perception of their speaking ability in English in general outside test conditions. This was carried out immediately following a speaking activity in the classroom a week before the individual oral interview took place in order to familiarise students with the criteria they would use to assess their test performances.

III.3.1 Student and Examiner Test Preparation

The common approach to the individual interview test format is to assume that students already take for granted that the interviewer will ask questions and they will be expected to respond. For this reason, students were prepared in advance for the interview only inasmuch as they were informed that the test would be based on a text which they would read before entering the interview room and that this would be used as a basis for the topic of the discussion. It was made clear that reading comprehension was not the objective of the test. The students did not, however, have any practice interviews or demonstrations of what was likely to happen during the test. Since students took the test on a voluntary basis, they were encouraged to see it as a mock speaking test prior to the final *Lengua BII* examination, and an opportunity to demonstrate their speaking ability after having developed it in class sessions over the preceding months.

In contrast to the 'Individual Oral Proficiency Interview', prior to the examination period, students spent a classroom session preparing for the 'Group Speaking Test' in the subject *Lengua BII*. All students were given a copy of the same text with accompanying questions, and the class subsequently watched a demonstration of the group test, performed with three volunteers from the group and their teacher as the interlocutor (no rater was present and no marks were given). There was a class discussion with the teacher about the test they had just observed, and students were able to ask any questions they felt to be necessary to clarify their doubts. The teacher then distributed a selection of materials packs and all students practised at least two different speaking tests in groups of three.

One of the intentions of our research was to try to discover whether familiarity with the test format, combined with peer support, can reduce anxiety and consequently enhance test performance. By giving students the opportunity to see exactly what would happen in the test and allowing them to practise one or more tests themselves, it was hoped they would become more confident in their approach to the tasks, and especially in understanding how we expected them to interact with each other. The fact that they were already familiar with the rating criteria from the two self-assessments they had already carried out (the general self-assessment of speaking ability done in the classroom and the one following the one-to-one interview) should have further consolidated knowledge of what would be judged and how the examiners would be looking at their performance. This procedure was in accordance with the other final written papers for the subject *Lengua BII* where practice tests are given at some time during the second semester to familiarize students with the exam format and to encourage them to use appropriate strategies in each part of the examination.

Since our examiners alternatively carried out the roles of both interlocutor and rater it was necessary to give standardisation training to all examiners in both capacities before the examining sessions took place. This was done by meeting with the examiners and discussing the written categories and rating scales, the two test procedures as laid out in the examiner's instructions and the level at which the test was to take place (defined as advanced course-book level or European Common Framework of Reference Level C1). Examiners were then presented with two or three sets of dummy candidates and were asked to carry out the tests and score them. These scores were then discussed and analysed by the group of examiners in order to set a

common standard for the actual examining sessions. A critical review of how the test was managed also took place among the examiners and the test designer.

III.3.2 Data Collection

The interviewers and raters worked in paired teams and carried out both types of test, the 'Individual Oral Proficiency Interview' and the 'Group Speaking Test', alternating their roles in order to experience managing and rating the two tests. In the one-to-one interview situation, the interviewer was required to carry out the dual role of interlocutor and rater, managing the interaction and simultaneously rating the candidates' performance on a scale of 0 – 10 according to the prescribed university marking scale. S/he was also then required to award the candidate a score using a 0 – 5 detailed rating scale containing descriptors of performance within different categories of spoken language production (see Appendix 3)². The design of this rating scale is discussed fully in Section III.5.

In the group speaking test, the interviewer was only responsible for managing the test and giving a global impression mark at the end, according to rating scale descriptors on a 0 – 5 scale designed for global rather than detailed assessment, while the rater focused exclusively on the task of assigning an analytic score to the candidates, using parallel but more detailed marking criteria and without becoming involved in the interaction at any point (see Appendix 3). After the administration of the tests, the interviewers and raters completed a questionnaire about the test itself and

² An independent rater was also present in the 'Individual Oral Proficiency Interview' test format for experimental control purposes.

their experience of managing the interaction and rating the candidates in the different situations.

In both the case of the candidates and the interviewers, the questionnaires contain a number of items which focus on the experience of the two interviews and on the application and understanding of the marking criteria in order to ascertain the possible impact of the different procedures on performance and the accuracy of the statements made about the candidates through the marks they achieved on the tests. The data-collection process is summarized below:

- 1) Students carried out a self-assessment of their general speaking ability, directly after a speaking task in the classroom a week before the first test. They were provided with the same detailed criteria on a 0 – 5 scale (Appendix 3) as the ones that would be used for the testing procedures and these were explained by the teacher. The criteria were slightly modified in terminology to make them more student-friendly, but had the same explicit meaning as those used by the examiners. In this way, the students were made familiar with the criteria before they were required to use them to assess their own performance on the tests, and their meaning was clarified before they received their results for either of the speaking tests. Students then filled in a score sheet (Appendix 11) with the marks they felt to be appropriate to describe their level of spoken English.
- 2) The following week, students took part in an '**Individual Oral Proficiency Interview**', lasting 5-6 minutes in the one-to-one format, but with a rater also present for the purposes of objectivity. The interview was recorded on cassette. Candidates were asked to read a text before the interview which formed the basis for the topic of the interview, but they were made aware that in-depth

textual comprehension would not be tested. After a short introductory phase consisting of questions to elicit personal information, the interviewer asked the prescribed questions from the relevant materials pack (see Appendix 4). The rater took no part in the interaction, but recorded the candidate's score according to the analytic scale on the mark sheet. The interviewer then gave the rater his/her impressionistic mark out of ten, followed by the analytic score on the same scale as that used by the rater. The interviewer and rater were instructed **not to discuss** these marks. On leaving the interview room, students were asked to complete a self-assessment sheet using the analytic rating scale and descriptors used by the rater, but with modified 'student-friendly' terminology, to evaluate how they thought they had performed on the interview.

- 3) A week later, students received the marks awarded to them by the interviewer (not those given by the rater) on both the impressionistic 0 – 10 scale and the analytic 0 – 5 scale. They were then asked to fill in a questionnaire referring to these marks and to their experience of the one-to-one test situation (see Appendix 5, Questionnaire 1 – Student).
- 4) On completing all the tests, the interviewers were asked to fill in a questionnaire (see Appendix 5, Questionnaire 2 – Interviewer) which focused on the experience of the one-to-one test situation and on using the two different marking scales.
- 5) The '**Group Speaking Test**' took place as part of the final examination for the subject *Lengua BII* in June, approximately six weeks after the individual oral interviews. It lasted 15-18 minutes and was recorded on video. Students were

examined in groups of three, choosing the day, time, and fellow-students with whom they wish to do the test. The procedure for this test was the same as that for the oral interview: each of the candidates was given a copy of the same text before the exam and they were allowed to read and discuss it together. At the start of the test, and as a settling-in phase, each candidate had an individual turn in which the interviewer asked them one or two questions each from among those provided in the materials pack (Appendix 4) to elicit personal information. The interviewer then gave each of the candidates a copy of three questions based on the topic of the text which they had just read and invited them to discuss the questions among themselves. S/he then withdrew from the interaction, avoiding eye-contact and therefore signalling extra-linguistically to the candidates that s/he would not participate in the conversation. The chairs were arranged in such a way as to encourage the candidates to talk to each other, rather than solely to the interlocutor, before they entered the room. The materials pack contained copies of both the text and the questions for interviewer reference, and provided an extra question which the interviewer could ask if the test was too short, or if one candidate in particular produced a significantly smaller speech sample than the rest. (See Appendix 6 for full instructions to examiners).

During the test, the rater gave each candidate a score on the analytic scale of 0 – 5, using the descriptors provided in each category on the rating scale (see Appendix 3). The interlocutor also gave a mark on a 0 – 5 scale, which was modified from the Rater Scale to provide criteria for a global impression mark

(see Appendix 3). S/he gave this mark to the rater to record on the mark sheet, but the marks were not discussed.

On leaving the interview room, students were asked to complete a self-assessment sheet (see Appendix 11) using the analytic rating scale and descriptors used by the rater, but with modified, 'student-friendly' terminology, to evaluate how they thought they had performed in the 'Group Speaking Test'.

- 6) One week later, the students were given their marks for the speaking test and were requested to fill in a questionnaire with particular emphasis on the experience of doing the test in a group (see Appendix 5, Questionnaire 3 – Student).
- 7) On completing the full session of group speaking tests, the interlocutors/raters were asked to fill in a questionnaire referring particularly to the experience of managing the 'Group Speaking Test' and awarding a global impression mark using an analytic scoring procedure (see Appendix 5, Questionnaire 4 – Interviewer).

A week before each oral test was due to take place, examiners were provided with a folder containing a brief description of the four categories of speaking competence to be assessed, the marking criteria and score sheets, the instructions for the procedure of the test (Appendix 6), a choice of questions for the introductory phase of the test, and a set of materials packs to be used for the individual tests (Appendix 4). The examiners were required to familiarize themselves with the instructions and the materials packs before commencing each examining session.

III.3.3 Introductory Phase

The initial stage of both the interview and the group test is a short settling-in phase which aims to put the candidates at ease by eliciting some general, personal information (see Appendix 4). The easiest thing for most people to talk about is themselves and most students are confident in expressing their plans for the immediate future (summer vacation; ‘Erasmus’ exchange visit), or in talking about a recent past experience, such as a film they have seen, or a place they have visited. This short phase aims to break the ice, put the candidates at ease and build confidence for the rest of the test. It also helps the rater to tune into the candidates’ pronunciation, while leaving aside judgement in the other categories for the main part of the test. For purposes of standardization, the interviewer adheres to the questions in the script, repeating them more slowly if asked to do so by the candidate or if s/he appears not to understand. A question may be paraphrased if it is still not understood after slower than normal repetition.

III.3.4 Materials Packs and Test Tasks

The materials packs consist of sufficient copies of the text for the candidates who will take the test (one in the case of the interview, and three in the case of the group test) and an extra copy for examiner reference, together with copies of questions referring to the topic of the text. The interviewer’s copy of the question sheet contains an extra back-up question in case the test comes to a premature end through the failure of the candidates to continue the interaction, or to be addressed to an individual student who has produced a significantly smaller speech sample than other candidates. This question may be used at the discretion of the interviewer.

The materials packs are designed to encourage candidates to express their opinion by providing a topic for discussion and by focusing their attention by means of questions addressing the differing perspectives from which it may be viewed. The topic is presented in the form of a short authentic newspaper article from a recent publication and it is hoped that the majority of students will be familiar to some extent with all of the topics chosen, either through general cultural knowledge, or through a personal identification with the issue which is similar to a situation in their own sphere of experience. In both test formats, candidates are informed before the test that the text will only be used as a springboard for discussion, and that an in-depth understanding of the text itself is not necessary. After being handed the text and prior to entering the interview room, they may use a dictionary to find the meaning of any words they do not understand. Reading comprehension is tested on a separate paper in the final exam for *Lengua BII* and it would therefore be inappropriate to test it here, or to confuse it any way with the speaking construct.

It is important to note here that the articles are not all of the same length and that some experts may feel that the texts differ somewhat in difficulty. It is our belief, however, that it is extremely complicated to accurately grade texts according to the difficulty they may present to the reader and also virtually impossible to establish criteria by which all the texts chosen for a particular purpose present the same level of difficulty. We therefore prefer to grade the tasks, which is a feature that is much easier to control. The questions for each of the texts follow a pattern of “more specific → more general”, referring in the first instance to the situation described in the text, and then becoming wider-reaching in order to allow the discussion to develop in a reasonably natural, but connected, way.

The materials packs are numbered and the pack used for each test is recorded on the candidate's mark sheet so that it is possible to trace any markedly easier or more difficult materials packs with reference to the scores obtained using those packs. The candidates and interviewers may also provide comments on any texts or topics they find to be especially difficult to handle or which have particular appeal.

III.3.5 The Test Environment

For all types of speaking test it is recommended that the physical conditions should include good lighting, absence of strong smells, appropriate heating or cooling, and protection from distracting noise. However, one of the major constraints from which the Faculty of Translation and Interpreting suffers is a lack of suitable rooms for language teaching in general and the availability of acoustically adapted rooms is very limited. Air conditioning is unavailable throughout the site. The tests took place in the most adequate rooms available, which always included a false ceiling to facilitate the comprehension of spoken language and to create a more agreeable atmosphere.

The furniture was arranged in a non-threatening manner in both the individual interview and the group test formats. In the 'Individual Oral Proficiency Interview', the interviewer sat opposite the candidate, to one side of the table, so that the interview did not take place across the desk. In the 'Group Speaking Test', the interlocutor faced the three candidates who were sitting in a semi-circle so that they all had a clear view of both the interlocutor and the other candidates in the group. In both test formats, the rater sat within the field of vision of the candidates, but at a distance which marked their non-participatory role in the interaction.

Care was taken to avoid distractions caused by a candidate sitting facing a window with strong sunlight shining through, or facing the door where a passer-by might suddenly appear and stare into the room or gesticulate. A notice was placed on the door indicating that an oral exam was in process and that no interruption should be made and the usher, who was administering the self-assessment sheets, dealt with any queries, late arrivals, or changes to the programmed sessions.

A second room, adjacent to or opposite the interview room was used for the candidates to read the texts prior to taking the test, either individually or in their groups, but away from other students who may have been arriving for or finishing their interviews. On concluding the test, candidates returned to this room to fill in the self-assessment sheets.

III.4 RATING SCALES AND DESCRIPTORS

III.4.1 Defining Features of Speaking for Assessment

In deciding how to score the candidates' speaking competence, three existing rating scales for testing at a similar level were consulted (ARELS, Trinity, and Cambridge ESOL) in order to study different informed approaches to assessing the speaking construct before designing our own rating scale for the *Lengua BII* speaking tests. These are discussed in some detail below.

III.4.2 ARELS Marking Key for the Higher Certificate Examination in Spoken English and Comprehension (see Appendix 7)

Firstly, it is important to note here that this exam does not take place in a face-to-face situation and is not marked in real time, but is recorded in a language

laboratory with the answers being recorded on tape for later assessment by at least two independent raters. As a model, therefore, it is highly impractical for our own situation, since for the final examinations in *Lengua BII* we have 120 candidates and a language laboratory with 19 places, and also a very limited number of personnel for administration and correction. However, it is still of interest to examine the rating scales used in order to identify points of coincidence which may be useful for our own situation or to extract from them the theoretical basis for the assessment procedure.

The first thing of interest to note is that in the general rubrics at the beginning of the marking key, the final grades awarded (*Fail, Pass, Credit and Distinction*) are arrived at through a percentage which is calculated from the candidates' score sheets. The rubric notes that: "[These] criteria can sometimes work in an over-arbitrary way, so markers are asked to give an impression mark of grade, independent of percentage total." They are asked to do this before totalling the points awarded in each section of the test. This statement seems to place doubt on the accuracy of the assessment criteria, or to imply that there may be some underlying failing in the marking key. The fact that the rater can have an overall impression of a candidate's performance that differs markedly from the final grade achieved according to the scores awarded on the test also suggests that either they have applied the criteria incorrectly or the criteria themselves are not clearly interpretable.

Either of these circumstances is possible, but a close examination of the criteria may tend to indicate the latter. Each section of the test has a different method of scoring, focuses on different aspects of the construct and uses a different scale. In the instructions to the rater for assessing Section 1 (a short talk on a prepared topic) we find the following scales:

Holding the listener's attention (by the interest and relevance of what the candidate has to say and the skill with which he says it) 0 – 12

Fluency 0 – 12

Accuracy of all aspects of the candidate's English 0 – 6

Here, no definition is given of each of the scores, and it is up to the examiner to choose a number somewhere on the scale.

Section 2 uses a scale of 0 – 4, each of the scores having a descriptor and focusing on appropriateness, comprehensibility, lack of ambiguity and “faultiness” of the response (e.g. 2 = “A response that is comprehensible and reasonably appropriate, although with quite serious faults”). In Section 3 the test looks at different phonetic and phonological aspects of pronunciation, each item being scored for a different feature, and some items being scored on a scale of 0 – 1, while others are scored on a scale of 0 – 2. Finally in this section the rater is required to give an overall impression mark on a scale of 0 – 8, a descriptor being provided for every other score, i.e. five descriptors for the scores 0, 2, 4, 6 and 8. Section 4 tests listening comprehension and is therefore not relevant to our discussion. Section 5 is a test of “fluency and accuracy in extended speech”. The rater is asked to consider three groups of features of the speaking construct: (i) pronunciation, stress, rhythm and intonation; (ii) appropriate and varied use of vocabulary and dialogue; and (iii) appropriate and varied use of structure. Each of these is scored on a scale of 1 – 10 with a descriptor for every other score, i.e. six descriptors for the scores 0, 2, 4, 6, 8 and 10. Finally, in Section 6, the focus of attention is “accuracy and control of an area of syntax or vocabulary”. Some of the items are scored on a scale of 0 – 1, and others on a scale of 0 – 2. There are no descriptors.

As can be seen from the above analysis, the scoring system is a complex one which raters will need to refer to constantly as they listen to and correct the candidates' tapes. The fact that there are descriptors for some of the scores and not for others may lead to some confusion, especially when applying the 0 – 12 scale in the first section, since this is an unusual number of points for any scale to contain, and it is likely that most raters would convert it to a percentage in order to be able to use it with any sense of accuracy. Even after standardisation training, this would still mean that they were using an internalised and personalised version of a scale to which each individual assigns their own meaning. In this sense, it would be similar to the traditional university 0 – 10 scale as used at the ULPGC, where no external meaning other than pass or fail is assigned to the scale prior to its use.

In other aspects, the scoring procedure is admirable in its wide-ranging attempt to isolate and assess so many aspects of the speaking construct in such detail, but the complexity of its structure and the need to make constant, detailed reference to the instructions would make it very time-consuming to use with an acceptable degree of accuracy, and impractical for our own circumstances where the examiner/candidate ratio is very unfavourable on the administrative side.

III.4.3 Trinity Grade Examinations in Spoken English for Speakers of Other Languages (see Appendix 8)

These tests of spoken English are independent, and do not form part of a combined written and aural skills syllabus. They are taken in a one-to-one format and form a series of twelve progressively graded tests, divided into four broad stages (*Initial, Elementary, Intermediate* and *Advanced*), moving from a low level of

proficiency (Grade 1) to an advanced level of proficiency approaching first-language ability (Grade 12).

Just how the level of first language ability is defined is not quite clear from the syllabus, although the assessment criteria seem to indicate that once more, the concept of the 'educated native speaker' underlies the descriptors of ability: statements from the Grade 12 assessment criteria such as "responding appropriately with confidence and ease at all times"; "entirely appropriate content of all contributions to the conversation"; "evidence of strategies to initiate and control the conversation"; and "competent organization of content of contributions to the conversation" indicate that it is a particularly adept class of native speaker who is being conjured up here.

Each grade is assessed in four areas of the speaking construct that are defined by the descriptors themselves: 'Readiness', 'Pronunciation', 'Usage' and 'Focus'. The twelve levels have different descriptions of what candidates are expected to be able to do in these four categories to pass at each grade. 'Readiness' includes understanding and responding appropriately, with maintaining the flow of conversation and taking initiative also being included at higher stages. 'Pronunciation' considers the production of individual sounds, as well as stress and intonation patterns. 'Usage' includes grammatical accuracy and lexis, and 'Focus' takes into account the appropriateness and organization of the content of candidates' speech. These assessment criteria, therefore, seem to cover all the areas that are addressed in the rating scales for other tests and examinations, although they use slightly different terminology in their category titles. Unfortunately, details of the breakdown of marks were not available for consultation for this study and it is only known that candidates receive an evaluation report and a mark out of 100, with 85+ being equivalent to a Pass with Distinction, 75-

84 a Pass with Merit and 65-74 a Pass. Here it can be seen that the traditional 50% pass mark has been abandoned, with candidates needing to be able to achieve at least 65% of the required objectives in order to obtain the certificate. This may be partly due to the fact that the stages are based on the Council of Europe's 'Common European Framework of Reference' which uses 'can do' statements to define the different levels. Obviously, if an individual can only do half the things required in the definition of a level, then s/he cannot really be considered to have attained that level.

The concept of twelve levels of speaking competence, each progressing towards the next and adding new material at the same time including what has gone before, is complex and if we turn to the definitions of the assessment criteria we will find that sometimes it is, indeed, difficult to establish marked differences between one grade and the next. For example, the difference between Grade 11 and Grade 12 'Readiness' is defined by nuances such as "understanding changes in register" (Grade 11) and "understanding changes in register and emphasis" (Grade 12) made by the examiner. Also in the 'Readiness' category, we find "controlling and maintaining the flow of conversation with ease" (Grade 11) and "controlling and maintaining the flow of conversation in a natural way" (Grade 12). In "Pronunciation" we find *occasional sounds* replaced with *rare sounds* that "deviate from an internationally intelligible model", and in 'Focus' there is a change from *adequate* to *competent* organization of content in contributions to the conversation. Apart from these, there are no other differences in the assessment criteria for these two levels. For such minor, subtle differences, it is questionable whether an examining board is entirely justified in encouraging candidates to take examinations progressively through their levels, requiring them to pay each time they take a test. However, what we do learn from this

close study is that writing assessment criteria that differentiate clearly between levels and attainment grades is an extremely complex and arduous task, and one which requires a comprehensive study of existing scales and reflective consideration of the task at hand. In designing our own rating scales, we need to be aware that, in order to be meaningful, rating scales and criteria need to be carefully thought out and extensively trialled and piloted before they become fully operational in an educational setting.

It is also questionable whether it is possible for an interviewer/rater to notice the candidate's use of "the full range of conditionals" (Grade 11) as opposed to second and third conditionals, conditionals with *unless* and *could have* plus participle (introduced from Grade 7 onwards) at the higher stages of proficiency where the candidate is expected to perform to near first language ability. It is extremely difficult to process the content of what is being said at the same time as listening for the range of grammatical structures someone uses as they speak, and it is certainly not a common thing for native or near-native speakers of a language to do as it would not make for natural conversation to attempt to use a range of grammatical structures in order to prove that you *can*. The only way to engineer this would be for the examiner to ask questions which elicited the use of these different tenses, but it appears that at the higher grades the candidate is encouraged to take control of the conversation, in which case the interviewer/rater would be required to focus even more attention on the content and direction of the discussion since this will be fairly unpredictable. These difficulties seem to strengthen the case for the presence of an independent rater in some form, either through tape recording of the test (which requires each test to be "performed" at least twice) or the physical presence of another examiner in the room

whose attention is not taken up with interview management or involvement in responding to the candidate (which leads to an even more unbalanced power structure). These are questions which were addressed in the design of our own test procedure.

The idea of an evaluation report in the ‘Trinity’ test is a particularly interesting and innovative one which should help to provide candidates with the necessary guidelines for identifying their own strengths and weaknesses and allow them to pinpoint areas in which they can improve their speaking skills. We have tried to incorporate this notion into our own test through the combination of the self-assessment process and familiarization with the rating scales, which will give the students a reasonable idea of what they do well and where they need to improve. Providing students with the same criteria as those employed by the examiners means that teachers are not involved in labour-intensive individual report-writing for over 100 students, while at the same time the latter receive information about how they have been judged which goes beyond a number on an abstract Pass/Fail scale.

III.4.4 University of Cambridge ESOL Examinations (see Appendix 9)

The speaking test for the ‘Cambridge ESOL’ suite of exams is a component of a larger written examination which also includes a listening comprehension test. The bands and scores for the speaking test throughout the suite follow the same basic design, but are adjusted in wording according to the level of the examination in question. There are various levels at which the Cambridge examinations can be taken, which correspond to the levels of the Common European Framework of Reference: Learning, Teaching, Assessment. These are defined by the Council of Europe as *Basic*

User. A1 and A2; *Independent User*. B1 and B2; and *Proficient User*. C1 and C2³ and correspond to the Cambridge levels of ‘Key English Test’ (A2), ‘Preliminary English Test’ (B1), ‘First Certificate in English’ (B2), ‘Certificate in Advanced English’ (C1) and ‘Certificate of Proficiency in English’ (C2). The level we are concerned with in our study and which corresponds to the stage of learning of our second year university students is that of C1 and consequently the Cambridge scale we have taken as a starting point is the one for the Certificate in Advanced English.

The CAE oral test takes place in real time in a candidate paired format with two examiners present, and lasts 15 minutes. It aims to test “interaction in conversational English in a range of contexts” with tasks focused on “exchanging personal and factual information, expressing and finding out about attitudes and opinions” (*CAE Handbook*, <http://www.cambridgeesol.org/exams/cae.htm>). It is divided into four parts: (i) an interview section, (ii) an individual long turn, (iii) a collaborative task, and (iv) a three-way discussion (two candidates and the interlocutor).

The speaking test is assessed in four areas: (i) grammar and vocabulary, (ii) discourse management; (iii) pronunciation; and (iv) interactive communication. While the candidates have access to the features of these categories that will be judged in any of the specific course-books that prepare them for the test, the rating scales and the scale descriptors are not made available to the public domain and therefore cannot be reproduced here.

³ Full definitions of these can be found at:
http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/levels.html

III.4.5 Rating Scales for *Lengua BII* Speaking Tests (see Appendix 3)

Due to personal involvement and previous training across the whole suite of speaking tests, and the similarity of the overall examination process where speaking is a component of a wider examination in all skills, it was decided that the rating scale for the two *Lengua BII* speaking tests carried out in this project would be broadly based on that used by 'Cambridge ESOL'. The names of the categories and the rating scale itself have been modified to suit our own circumstances and needs, and the definitions of the features to be judged in each category are original. The descriptions of the categories as they appear in the *Lengua BII* examiners' instructions are reproduced below:

Grammar and vocabulary

In this category the aim is to evaluate the grammatical accuracy of utterances. Occasional or minor inaccuracies are not important, especially in the "settling in" phase, but frequent and repeated inaccuracies should be taken into account, especially if they impede understanding.

The range and appropriacy of the vocabulary used by the candidate is also judged here. Students will be expected to have a good range of vocabulary to talk about themselves, and an adequate range for dealing with the other topics of conversation. Credit may be given for paraphrasing of more complex concepts, but not where a word or phrase should be part of the expected working vocabulary at this level.

Pronunciation

Here, both the pronunciation of individual words and general patterns of rhythm and stress should be taken into account. Candidates should not be penalised

for having an L1 accent, unless this is so marked as to prevent understanding. However, they should be awarded positive marks for approximation to more native-like pronunciation and for attempting to use pronunciation features like weak forms and stress.

When considering candidates' patterns of rhythm and stress, assessors should put themselves in the place of a tolerant speaker of English in order to decide how much strain the pattern of the candidate's speech produces and how much it may impede understanding.

Discourse structure

This deals with internal coherence, i.e. the student's ability to organize speech coherently by making appropriate use of cohesive devices and the tense system so that at this level s/he is able (for example) to present an argument (not necessarily an extensive one) or statement and support it in a relevant way, without leaving utterances unfinished or pausing for too long to search for language or order ideas.

Interaction

Here the aim is to judge the candidates' ability to interact with others in the conversation, both on a sociolinguistic level by being sensitive to turn-taking, using politeness strategies for disagreeing, encouraging others to participate, etc., and at the level of external coherence by responding logically to leads, questions, and the general direction of the conversation.

III.4.6 Designing the Rating Scale

The rating scale for our own testing procedure (see Appendix 3) was arrived at firstly through a consideration of the categories that could be considered to be

component parts of the speaking construct, as discussed above in II.2, followed by the design of descriptors that would summarise a range of ability across the level to be measured. Given that several raters and interviewers would use the scale, it needed to be visually clear, concise, and easy to manipulate, and use terminology that would make it possible to differentiate without difficulty between the features that distinguished each score. For this reason, it is set out in the form of a table where the categories and the scores are immediately apparent, using a single page format so that it is not necessary to turn over sheets or refer to several pages in order to assess the candidate. This is especially important in assessment in a real-time setting, since the sight of examiners leafing through multiple papers in their presence can be an added cause of anxiety to the candidates during the test.

As we have seen above, attempting to clearly define and differentiate ten scores can prove to be an almost impossible task, with descriptors simply substituting vocabulary items such as *most* for *nearly all*, or resorting to intensifiers like *very* in order to change *frequent* to *very frequent* to try to justify discrete scores. We have also seen that often the scale descriptors even repeat themselves as happens with the Trinity Level 11 and 12 assessment criteria, where many of the items described in Level 11 are simply repeated for Level 12 (see Appendix 8). For this reason, and in order to simplify the process of creating the rating scale, here we have reduced the scale to five points as in the 'Cambridge ESOL' model, with the middle band representing the attainment necessary to be considered to have achieved a satisfactory score to pass at this level. By doing this, it is hoped that the scale will be much easier to use and that there will be a clear differentiation between the scores.

The decision was made to define only three scores of the possible five: the lowest score (1), the required achievement to pass (3), and the highest score at the level being examined for *Lengua BII* (5). This again was felt to make the scale easier to use in real-time, since raters would not have to cope with so many descriptors and the “adequate” score would be a clear starting point for assessment from which they could move up or down according to candidate performance. It would also seem to be beneficial to candidates taking the test to assume at the outset that they will have an adequate level, rather than to start at the bottom of a scale and see how far they can go, which is not generally conducive to the awarding of high marks. Many non-standardised oral tests use only negative marking strategies, focusing on the number of mistakes made by the candidate during the test without taking into account the positive features of performance. We would argue that this is an unrealistic way to judge oral production, since all speakers, including very proficient ones, make mistakes in their first language but clarify and correct themselves, and it would therefore be absurd to make an assessment of a foreign language speaking test based on isolated mistakes that were made during the examination, especially if interaction and effective communication were achieved. We propose that while perfection in speaking is probably not possible, it is feasible for candidates to attain the highest score at a particular level, when that level and score have been defined within achievable aims which refer to a maximum cut-off point for the level.

III.5 QUESTIONNAIRES

After each of the tests, a structured questionnaire was administered to both the interviewers and the students in order to collect information about their opinions on the

procedure, format and marking of each test. The interviewers completed their questionnaires (Questionnaires 2 and 4) after all the tests in each examining session, while the students were asked to respond when they received their marks a few days after the tests in order to be able to express an opinion on their understanding of the mark they received (Questionnaires 1 and 3).

III.5.1 'Individual Oral Proficiency Interview': Student Perspective

The final version of the questionnaire was arrived at through the following matrix:

QUESTIONNAIRE 1 – STUDENT

Experience of one to one interview	Socio-affective aspects	1. I felt nervous throughout the whole test. 2. I think I did well in the test. (Give yourself a mark from 1-10)
	Performance	3. I performed to the best of my ability in the test. 4. I think I spoke enough for the tester to judge my ability.
		Procedure
	Test and task features	Task familiarity
Level of difficulty		7. I could answer the questions without difficulty.
Topic		8. I could find enough to say about the topic.
Global mark vs. analytic mark	Fairness	9. The global mark I received was a fair mark. 10. The analytic mark I received was a fair mark.
		Understanding mark

		13. The global mark I received was easier to understand than the analytic mark.
	Improving	14. The global mark helped me to understand what steps I need to take in order to improve my speaking.
		15. The analytic mark helped me to understand what steps I need to take in order to improve my speaking.

+ Please add any other comments you would like to make about the test itself and/or your experience of the test.

The aim of this questionnaire is to focus mainly on the experience of being interviewed in a one-to-one situation, with a second person (the rater) also in the room for purposes of objective assessment. From the outset, the candidate is in a position of inferior status; there are two people in the room who will make a judgement on what just one candidate says. In the accepted social structure, the student is inferior to the university lecturer, and in this case there is a ratio of two lecturers to one student, compounding and strengthening this inferiority. It is therefore expected that the candidate will feel anxiety produced not only by the test situation itself, but also by the socio-affective aspects inherent to the interview test-type and the unbalanced social situation, and this is addressed in the first two questions.

In the power structure resulting from the interview situation, the interviewer has absolute control over the interaction, with the 'speaking rights' to initiate the exchange, continue or change the topic, and to bring the interaction to a close as s/he chooses, which may translate into a dominant or controlling force over the candidate's performance. Questions 3 and 4 attempt to elicit from the candidates whether they feel that the test situation in the interview allows them to effectively demonstrate their

speaking ability, and whether they think that this procedure is an adequate way of testing their oral ability (Question 5).

A further consideration for this investigation is the extent to which task familiarity and format affect performance on the test. The interview format is an extremely common test-type used for assessing oral ability, yet it is almost never, if ever, employed as a classroom activity and therefore students have little or no practice or experience of this type of task. We have assumed that this negatively affects candidate performance and that this may be reflected in both Questions 5 (Procedure) and 6 (Task familiarity). The level of difficulty of the tasks and the general interest of the topics should remain constant in both tests, since the design of the materials packs is based on the same criteria (see Appendix 4). However, it is possible that candidates find it more demanding to answer questions which they only hear, which occurs in the interview procedure; Questions 7 and 8 deal with these features.

Finally, the questionnaire addresses the issue of impressionistic assessment on a scale of 0 – 10 following the traditional university grading system and how this is generally understood by students, in comparison with how they might interpret the new marking criteria on a reduced scale of 0 – 5 which have clear descriptors assigned to them. It will be of interest to see whether these are more meaningful and easier for students to interpret, and whether they are more useful in indicating which areas of their speaking ability they need to improve. The questionnaire attempts to gauge student perceptions of these issues with Questions 9-15.

We also invite candidates to comment on any other issues they consider to be relevant and that have not been included elsewhere in the Questionnaire, with a view

to widening our perception of how students are affected by the oral test situation and how valid they feel it to be a reflection of their oral competence.

III.5.2 'Individual Oral Proficiency Interview': Interviewer Perspective

The final version of the questionnaire was arrived at through the following matrix:

QUESTIONNAIRE 2 – INTERVIEWER

Managing the test	Simultaneous rating and interviewing	1. I was able to manage the interview and give the student a global mark on a scale of 1-10.
		2. I was able to manage the interview and give the student a detailed score at the end of the interview.
		3. I was more focused on managing the interview than on the rating criteria.
		4. I felt comfortable in the dual role of interviewer and rater.
		5. I felt happy about the test procedure.
	Size of speech sample	6. The student produced a large enough speech sample for assessment.
	Interaction	7. It was easy to assess how well the candidate was interacting.
Global marking vs. analytic rating	Understanding	8. I understood what I was assessing in giving the global mark.
		9. I understood what I was assessing in giving the analytic score.
	Focus of assessment	10. The most important part of my assessment in giving the global mark was grammatical accuracy.
		11. The most important part of my assessment in giving the detailed score was grammatical accuracy.
	Fairness	12. I think I awarded the student a fair mark in giving the global mark. (Reason:)
13. I think I awarded the student a fair mark in giving the analytic score. (Reason:)		

	Converging/diverging opinion	14. It was easier to mark a student who expressed an opinion similar to mine in giving the global mark.
		15. It was easier to mark a student who expressed an opinion similar to mine in giving the analytic score.

+ Please add any other comments you would like to make about the test itself and/or your experience of managing and rating the interview.

This questionnaire focuses on two main aspects of the interview procedure from the point of view of the interviewer/rater: (i) test management and (ii) global or impressionistic marking contrasted with using an analytic rating scale with detailed descriptors of language ability to assess the candidate. The first five questions focus on the issue of managing the interview at the same time as making a judgement on the candidate's ability and to what extent this is possible or even desirable. It is our belief that, in fact, it is extremely difficult to simultaneously conduct the interview and objectively assess a candidate's performance, and that, while interviewers may believe they are capable of doing both these things at the same time, actually they may be influenced by factors other than the candidate's oral ability in awarding the mark. Especially in the case of the global impression mark on a scale of 0 – 10, examiners often rate students by comparing them to one another rather than by judging them in an objective way, and it is possible that they have been influenced by external features such as empathy with the candidate, coincidence or divergence of opinion on a particular topic, or extra-linguistic communication skills. Question 1 therefore refers to the confidence with which the interviewer feels s/he can competently award the mark from 0 – 10, but without making reference to what the interviewer actually understands these marks to mean. Question 2 refers to the interviewer's ability to use

an objective analytic rating scale in retrospect, since it was felt to be too complex and off-putting for the candidate for the interviewer to use it during the test, while interview management and test procedure are the object of the following three questions.

Question 6 contrasts the interviewer's opinion of the amount of candidate speech produced with that of the candidates themselves (recorded on 'Questionnaire 1 – Student', Question 4). In our analysis of the data, we will be interested to discover whether students produce a large enough sample for assessment in this test format since, in our experience, there is often a temptation for the interviewer to speak more than necessary, thus reducing the amount of time available for candidate speech. If this does in fact occur, then a candidate may be awarded a high mark for an interview where they have empathized with the interviewer, but where their speech sample for assessment is very small.

The last question in this section aims to focus the interviewer on the task of judging how well the candidate is able to interact, assuming in the first instance that global impressionistic marking is influenced by the ease with which the interview has proceeded, and also possibly by the convergence of the test-taker's views with those of the interviewer. On examination of the data, we will also attempt to find out what type of interaction the interviewer and rater are assessing, assuming that it will almost exclusively be responses to content questions, and that asking questions, taking turns, switching topic, inviting an opinion or initiating and closing interaction, amongst other important sociolinguistic skills will not be present in the conversation and therefore the candidate's skills in these areas will not have been evaluated.

The second part of the matrix contrasts assessment using a global impression mark common to the university system (scale of 0 – 10) with scoring on the analytic scale using descriptors in different categories of speaking competence. The first two questions focus on the examiner's understanding of what the marks they award actually mean, and how far it is possible to assign objective or descriptive meaning to the 0 – 10 scale. It is our belief that in using this scale, examiners are in fact measuring candidates not according to an objective statement of their ability, but according to how they compare to one another which means that the 0 – 10 scale is, in fact, a norm-referenced (not a criterion-referenced) measurement scale. We also predict that, for the most part, examiners will be unaware of this, and believe that they are giving objective and descriptive marks which clearly indicate a degree of competence to the students themselves and to other users of academic marks such as parents, university administrators, local and national governmental education managers, and employers in general.

The following two questions address the focus of the examiner's assessment. Even when rating scales are specifically designed to give credit to candidates' communicative competence rather than their grammatical accuracy, examiners may persist in their tendency to make this aspect the main basis for their assessment. Here, we will try to discover whether the analytic rating scale can help examiners to focus on different aspects of speaking competence by contrasting replies to Questions 10 and 11 in the questionnaire.

Questions 12 and 13 attempt to elicit whether impressionistic marking is perceived by the interviewers to be more or less accurate or reliable. If, as postulated above, they measure candidates against one another when using the 0 – 10 scale, their

scoring will depend on the memory they have of the performance of students whom they have previously interviewed. Since it is probably not possible to retain a detailed memory of performance for more than one or two previous interviews, interviewers may reach a point of uncertainty or confusion about how they are applying the scale, even with reference to their own internalised version of it. Since the object of the analytic marking criteria is precisely to avoid this, it will be of interest to see whether interviewers feel more confident about awarding scores in this more objective way.

Finally, with the last two questions we aim to look at how far interviewers are influenced in their assessment by a convergence or divergence of opinion with the candidate. Generally, as socialized human beings, we are more positively disposed towards others who show opinions similar to our own on certain topics, and it is therefore likely that examiners are also influenced in this way and tend to give higher marks to candidates whose ideas or ideologies coincide with theirs. The interview situation, by its very nature, compounds this tendency by drawing the interviewer into a much more personalised dialogue with the candidate. Because the 0 – 10 scale does not use written, objective, or stable criteria, it is logical to assume that it will be more susceptible to influence by such circumstances. We will try to determine whether this effect can be inhibited or reduced by using the analytic scale.

We also provide the space and opportunity for the interviewers to comment on any other aspect of the test which they feel to be of interest or use in the study and which has not been covered by any of the items in the Questionnaire. It is the hope here that we will gain some further insight into what is actually happening in the mind of the interviewer/raters as they are conducting and simultaneously assessing the test.

III.5.3 'Group Speaking Test': Student Perspective

The final version of the questionnaire was arrived at through the following matrix:

QUESTIONNAIRE 3 – STUDENT

Experience of group testing	Socio-affective factors	1. I felt nervous throughout the test.
		2. I think I did well in the test.
	Performance	3. I think I performed to the best of my ability in the test.
		4. I think I spoke enough for the examiner to judge my ability.
	Procedure	5. I felt comfortable with the procedure of the test.
		6. I knew exactly what I had to do.
Test and task features	Task familiarity	7. The test was similar to the kind of task practised in class.
	Level of difficulty	8. I could answer the questions without difficulty.
	Topic	9. I had enough to say about the topic.
Marking criteria	Understanding mark	10. I understand what my mark means.
	Improving	11. I know what I need to do in order to improve my speaking.
Self-assessment	Accuracy	12. I think that my general self-assessment was a true reflection of my speaking ability in English.
		13. I think that my self-assessment in the Group Speaking Test was a true reflection of my speaking ability in English.
	Usefulness	14. I think self assessment can play a useful role in learning generally.
		15. I think my self-assessment should be taken into consideration in my overall grade for the subject Lengua BII.
		16. We should be given the opportunity to use self-assessment more frequently in this subject.
	Training	17. We should be trained in how to assess our language skills in this subject.

+Please add any comments you would like to make about the test itself or that you feel would be helpful in improving the test for the future.

This questionnaire aims to highlight the differences that students perceive between the two different test types and to discover whether these may be reflected in test performance and consequently in candidates' scores. As a point of departure, we presume that taking the test with other students (in most cases class-mates and friends), together with familiarity with the test procedure, will reduce anxiety and that this, in turn, may lead to enhanced performance and an improved score. Therefore, in order to be able to contrast how students perceive the 'Group Speaking Test' in comparison with their experience of the one-to-one interview, the first five questions from Questionnaire 1 are repeated, with an additional question (6) in the Procedure section '*I knew exactly what I had to do*', which is intended to address the fact that students were prepared for the test in advance and, if they had attended classes, should be familiar with the test format and procedure. This is in contrast to the interview situation, where students were given no information about the test prior to its taking place, reinforcing the positions of power and control held by the interviewer and rater, who knew exactly how they would proceed and what they expected of the candidates.

A further expectation is that the similarity of the test procedure to the kind of speaking activities carried out in class during the year will mean that candidates are less anxious about the test procedure and should also have more expertise in carrying out the required tasks. This point is addressed in Question 7, although we only expect that those students who have attended classes regularly will respond positively. However, we anticipate that there will be a contrast in the response to the same question in Questionnaire 1 where we presume that the majority of students will indicate that the test bore no similarity to classroom practice.

Questions 8 and 9 again address the topic of materials design and we expect that there will be little variation from the responses to the first questionnaire, since the materials were designed according to the same principles and using the same criteria. The following two questions are also repeated from the first questionnaire and elicit the candidates' opinion on the analytic marking criteria, and whether or not they can help students to identify their areas of relative strength and weakness in speaking in English. It is possible that in the 'Group Speaking Test', candidates are more aware of their interactive ability or performance since they will see it in relation to that of their fellow students, and we may see more agreement with this statement than in the first questionnaire.

Questions 12 to 17 attempt to gauge student opinion on the usefulness and validity of self-assessment, both as a tool for learning and improving language skills and as a part of the overall evaluation process. Firstly, we look at the aspect of accuracy, eliciting in Question 12 how far students believe their own evaluation of their speaking ability should be considered an integral part of what they are able to do in English. The following question (13) requires students to contrast what they believe their general ability is in comparison with how well they were able to demonstrate this on the test (underlying ability vs. performance). Question 14 elicits how students feel about the usefulness of self-assessment as an instrument for learning. Their answers will be of interest to identify the extent to which they are aware of their own role in the learning process, and how far they see themselves, rather than their teachers, as part of a process of strategy implementation which can help them to develop their own language skills. It is our belief that many students will still be reluctant to see themselves in a role of major responsibility in the learning process and that much more

training in learning strategies is required than that which we implement at present in this course.

Question 15 takes the concept a step further and asks students how far they see self-assessment as a valid measure of their own ability. Whilst many teachers would argue that self-assessment is clearly open to abuse, it is our experience that the majority of students seem to be very accurate in evaluating their own ability and are also honest in their assessment of how much effort they have invested in their work (e.g. Cranfield and Clouet, 2006). This belief is reflected in Question 16 which invites an opinion on student interest in developing this area of pedagogy which will undoubtedly be a new concept to them. Question 17 recognizes that it is possible that students will require more practice and even training in this area in order to feel confident in its implementation, and we also realise that it will take some time for teachers to partly relinquish control over an area in which they have traditionally held absolute power. However, in doing so, we should become aware that in fact our marks gain in reliability and validity precisely because they are not wholly dependent on marking schemes and scoring systems that are, at best, internalised versions of an objective description of ability, and possibly often no more than a subjective accumulation of ideas and experience with no stable reference point.

Finally, students are invited to make any other comments about the test that they feel to be important and which have not been covered anywhere else in the questionnaire, but always with a view to them making constructive observations and not simply complaining about a personal experience. We feel it is important for students to recognise that they can play a role in educational change and reform. All

too often, those most affected by educational innovations are never consulted about them to the detriment of all concerned.

III.5.4 'Group Speaking Test': Interviewer Perspective

The final version of the questionnaire was arrived at through the following matrix:

QUESTIONNAIRE 4 – INTERVIEWER

Managing the test	Simultaneous rating and interviewing	1. I was able to manage the interview and give each student a score at the end of the test using the rating scale provided.
		2. I was more focused on managing the interview than on the rating criteria.
		3. I felt comfortable with the test procedure.
	Size of speech sample	4. The students produced a large enough speech sample for assessment.
	Group format	5. It was difficult to manage the test with three students participating.
		6. I felt comfortable in the dual role of interviewer and global rater.
Rating	Using the rating scale	7. I knew what features to focus on while assessing the candidates.
		8. It was easy to assess how well the candidates were interacting.
		9. It was useful to have a rating scale to refer to when giving the global score.
	Converging/diverging opinion	10. It was easier to assess students who expressed an opinion similar to mine on the topic.
	Understanding	11. It was easier to use a scale from 0-5 than one from 1-10
		12. It was easier to assign meaning to a scale of 0-5 than to one of 1-10.
	Fairness	13. I think that I awarded the students a fair score. (Reason:)
Accuracy	14. I think that students can give a true reflection of their general speaking ability using the criteria provided.	

Self-assessment		15. I think that students can give a true reflection of their performance in the Group Speaking Test using the criteria provided.
	Usefulness	16. Self-assessment is a useful tool for helping students to know how improve their speaking ability in English. 17. Self-assessment can play a useful role in learning generally.
	Purpose	18. Self-assessment should be taken into consideration in the students' overall mark for English Language subjects at the ULPGC.

+ Please add any other comments you feel might be useful and constructive in developing this test for the future.

The second interviewer questionnaire principally focuses on contrasting the experience of carrying out the one-to-one interview and rating simultaneously with that of managing the interaction in the 'Group Speaking Test' with three students present, and on using an analytic rating scale to award a global mark to each candidate at the end of the test. However, we also address teacher opinion on the role of self-assessment in teaching, learning and evaluating students' work as an area of interest in itself and also in order to be able to contrast it with the views of the students in the study.

The first three questions are repeated from the first interviewer questionnaire and may help to reveal whether it is in fact easier for interviewers to use a descriptive rating scale to give a global impression score to three candidates whose interaction is not directly dependent on the interlocutor, than to be engaged in one-to-one interaction and simultaneously rate the performance of a single candidate. Again, we also address the question of whether candidates produce a large enough speech sample for

assessment. We assume that interviewers speak significantly less in a group speaking test format than in a one-to-one interview, but it remains to be seen whether candidates actually produce more speech since the time allowed per candidate is ostensibly the same, and the time for interaction is shared among the students taking the test at the same time. A drawback of the procedure is that it allows the opportunity for an extrovert candidate to dominate the interaction and so we need to ensure that, should this occur, interlocutor training to prevent or compensate for it is sufficient.

Finally in this section, we invite the interviewer's opinion on the difficulty of handling interaction between three candidates and on whether inter-candidate interaction frees up interlocutor attention for more accurate assessment. By contrasting the responses to these questions with those obtained in Questionnaire 2, it should be possible to discover an interviewer preference for one type of interview as far as test management is concerned.

The second section of the matrix looks at the rating procedure and how the implementation of an analytic rating scale affects the testing procedure from the interviewer's point of view. Questions 7 to 9 implicitly allude to the descriptors as a fixed point of reference in scoring, assuming that this will provide interviewers with more confidence in awarding marks and will focus their attention on the different aspects of the speaking construct that are distinguished in the scale used by the rater so that grammatical accuracy, for instance, does not predominate over other features of the construct.

Question 8 looks again at interaction and attempts to compare the interviewer's impression of assessing candidate interaction whilst being simultaneously involved in it in the one-to-one interview (Questionnaire 2) with a much more objective, outsider

view in the 'Group Speaking Test'. Here, the interlocutor does not form part of the discussion that takes place and can therefore focus attention on different sociolinguistic aspects of the candidates' interaction, such as turn-taking, asking and responding to questions, inviting other people's opinions, supporting others in the conversation, and changing the topic. It is hoped that in the group test format it may be possible to assess candidates on a range of interactive abilities and not just on whether they "speak" or not.

Question 9 makes reference to the rating scale and to whether interviewers find it easier to rate candidates when there are fixed criteria to refer to. The intention of the rating scale is to provide a description of competence as it is reflected in performance and to thus avoid the comparison of one candidate to another since such statements are meaningless outside the context of the test. It will be of interest to see whether interlocutor/raters become aware of this, or whether the traditional way of marking has become so much a part of their habitual activity that they internalise the new scoring system and adapt it to fit their own interpretation of it. These issues are further addressed in Questions 11 and 12, where interviewers are asked to say whether they found it easier to employ a reduced scale, and whether this was more meaningful since it had descriptors assigned to it. The intention of the descriptors on the 0 – 5 scale is to explicitly state the abilities that candidates demonstrate in the test, and the scale is necessarily reduced because of the impossibility of assigning meaningful and distinctive descriptors to a wider range of numerical marks. We hope to gain an insight into interlocutor/rater opinions on what they believe they do when they employ the 0 – 10 scale, and on whether the reduced scale with definitions can improve on this. Finally, Question 13 attempts to discover whether interlocutors consider the use of the

objective rating scale to be a fairer and more consistent way of awarding marks to students. Here, we will need to compare answers with the same question on the previous questionnaire to establish whether there is any difference in views and also what the reasons are for the beliefs held.

The issue of converging or diverging opinion is re-addressed in Question 10; we postulate that if, in the individual interview situation, interviewers have found it easier to assess students whose ideologies coincide with their own through being more favourably disposed towards them, in the 'Group Speaking Test' they may be more detached from the interaction, and therefore in a position to focus more clearly on the quality of language production without it being essential to follow the candidate's train of thought and ideas in order to be able to continue the interaction.

The final part of the questionnaire turns its attention to the role of student self-assessment from the teacher's perspective, with Questions 14 and 15 looking at how accurate teaching staff consider their students to be in their self-appraisal. It will be possible to measure this accuracy by contrasting the data collected from the scores awarded to students on the tests by the raters and the scores they award themselves directly after the tests, and we should also therefore be in a position to advise teachers whether their opinions about the validity of student self-assessment are founded or not. In the following questions, we also ask them to consider how useful a role self-assessment can play in learning generally and in speaking in particular. Should this prove to be given a positive high profile, it would indicate that we need to incorporate into our teaching programmes strategies for self-assessment and also to take into account their results when assigning marks and scores to our students. If we find that Questions 16 and 17 are given agreement or high agreement scores while Question 18

receives a low agreement score, this would indicate that we need to address the issue of teacher education and modification of teaching/learning programmes in order to be coherent in our approach to both teaching and testing.

III.6 MARK SHEETS AND STUDENT SELF-ASSESSMENT SHEETS

The candidates' test scores were recorded on the mark sheets specifically designed for each test format (Appendix 10). These were then used to transfer all data to the appropriate computer programmes for analysis.

In order to speed up the testing procedure, candidates were given a mark sheet before entering the interview room where they filled in their own name and, in the case of the 'Group Speaking Test', the names of the other candidates who took the test with them. These sheets were handed to the interviewer who then passed them to the rater who was exclusively responsible for completing them. The examiner in the role of interviewer/interlocutor had been asked not to write on the mark sheets, and the examiners should have given their marks independently without discussion and should not have modified these according to the other examiner's assessment.

Students completed three self-assessment sheets at the different stages of the investigation indicated above. They received help in interpreting the criteria if they requested it, but were not influenced or assisted by other students or teachers in any other way. Throughout the study the sheets had the same format (see Appendix 11) and were filled in using the same criteria. At the end of the entire data collection process, the sheets were numbered in the same way as above, and the data entered into the appropriate computer programme for interpretation. The following chapter provides an analysis of the data obtained.

IV. RESULTS

In the following chapter, we will present the results of our investigation in two sections. Section 1 deals with the first part of the study and presents the data obtained for the 'Individual Oral Proficiency Interview' test, together with the opinions collected from the students in Questionnaire 1, and from the interviewers/raters in Questionnaire 2, about their experience of this test format. The second section shows the results obtained for the 'Group Speaking Test' and the views expressed in Questionnaires 3 (students) and 4 (interviewers/raters) about the experience of taking and rating this test.

In both sections, the test results are presented in the form of graphs which show the mean values in each of the speaking construct categories assessed. This is followed by the findings of the correlation study between the different scores awarded by the rater, interviewer and student in the first test (the individual interview), and the rater and the student in the second (the group test). The final part of each section presents the data collected via the questionnaires and which throws light on some of the affective and subjective aspects of testing speaking skills which we may need to take into account if we are to improve and defend the objectivity of the scores we obtain.

These findings, their possible causes and consequences, as well as their implications for our teaching and testing programmes are discussed in Section 3. Here we will try to give a global view of what our study reveals on a general level about issues in testing speaking skills and, on a more specific level, of how we may take account of these results in order to improve our approach to speaking tests in our own university context.

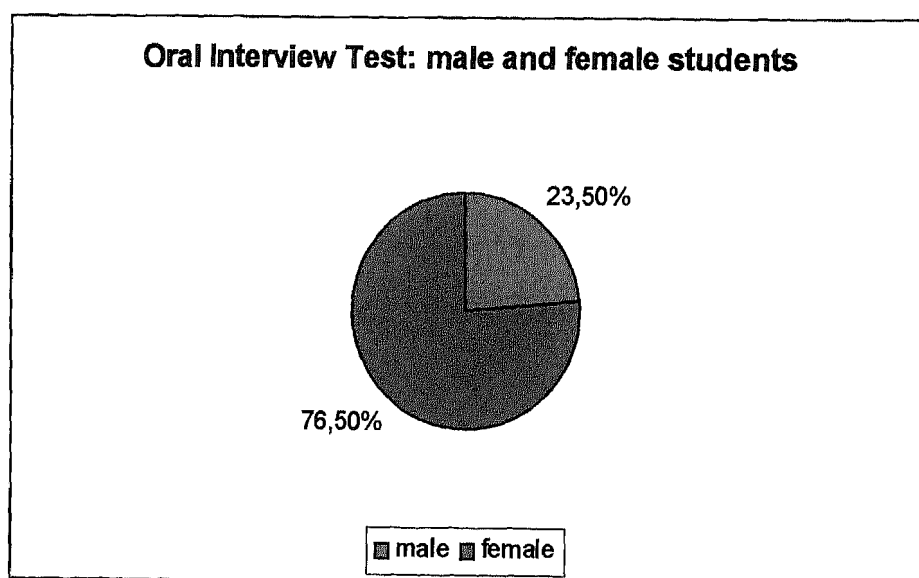
IV.1 INDIVIDUAL ORAL PROFICIENCY INTERVIEW

The total number of students who took the individual interview test (carried out with an interviewer and with a rater present only for experimental control purposes), filled in the self-evaluation sheets and completed the questionnaires was 51. Of these, 37 (72%) achieved a pass mark for the test according to the score awarded by the rater.

Male/Female

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos male	12	23,5	23,5	23,5
female	39	76,5	76,5	100,0
Total	51	100,0	100,0	

Of this total, 76.5% were female and 23.5% male.



As set out above, in Chapter 3 *Research Design*, the performance data for the first test (the ‘Individual Oral Proficiency Interview’) were obtained from the *Oral Interview Assessment Sheet* (see Appendix 10) where the rater assessed the candidates in the categories of ‘Grammar and Vocabulary’, ‘Pronunciation’, ‘Discourse Structure’ and ‘Interaction’ on a scale of 0 – 5. These same features were also evaluated by the interviewer who, in addition, gave a global impression mark, using the traditional scale of 0 – 10. The *Assessment Sheet* also recorded the test pack used for the interview.

The other instrument used for data collection was the *Self-Assessment Sheet: Speaking – Lengua BII – Interview* (see Appendix 11), where the students assessed themselves according to how well they thought they had performed on the test, using the same scale and categories, as well as similar criteria, to those used by the rater and the interviewer. The final step was to collect the students’ opinion on the interview test format using a questionnaire consisting of 15 items. All these data were recorded on an Excel spreadsheet (see Appendix 12), and were then introduced into the SPSS Statistics Package in order to commence the statistical study¹.

However, before beginning the study it was necessary to determine the types of variables which might affect the results. Of the 34 variables present (see Appendix 12), the first (student) and the second (male/female) can be considered nominal, since they only provide information about the student’s identity and gender. Similarly, variable number 19 (test pack) is also nominal, because it simply distinguishes which of the ten test packs was used in each interview.

¹ The rights to use the SPSS statistics package have been purchased by the ULPGC only in Spanish, and it is not possible to change the main labelling of the graphs and tables into English.

Sixteen variables (numbers 3 to 18) are made up by the marks awarded to the interview by the rater, interviewer and the student (through self-assessment). If we take into account that, as described above, each of these assessments was carried out in four categories ('Grammar and Vocabulary'; 'Pronunciation'; 'Discourse Structure' and 'Interaction') and that in addition, the interviewer gave a global mark out of 10, we would expect to obtain a total of 13 variables, and not the 16 indicated here. This difference can be accounted for because we have also calculated the arithmetical mean of the specific marks awarded by the rater, interviewer and student, thus giving a total of 16 scale (or quantitative interval) variables which make up the score achieved by each student.

The last 15 variables (numbers 20 to 34) include the students' answers to the 15 items on the questionnaire. Here, the students responses are reflected according to the answer to each question on a scale of 1 – 4 ('strongly disagree', 'disagree', 'agree' and 'strongly agree'), and are therefore quantitative ordinal variables².

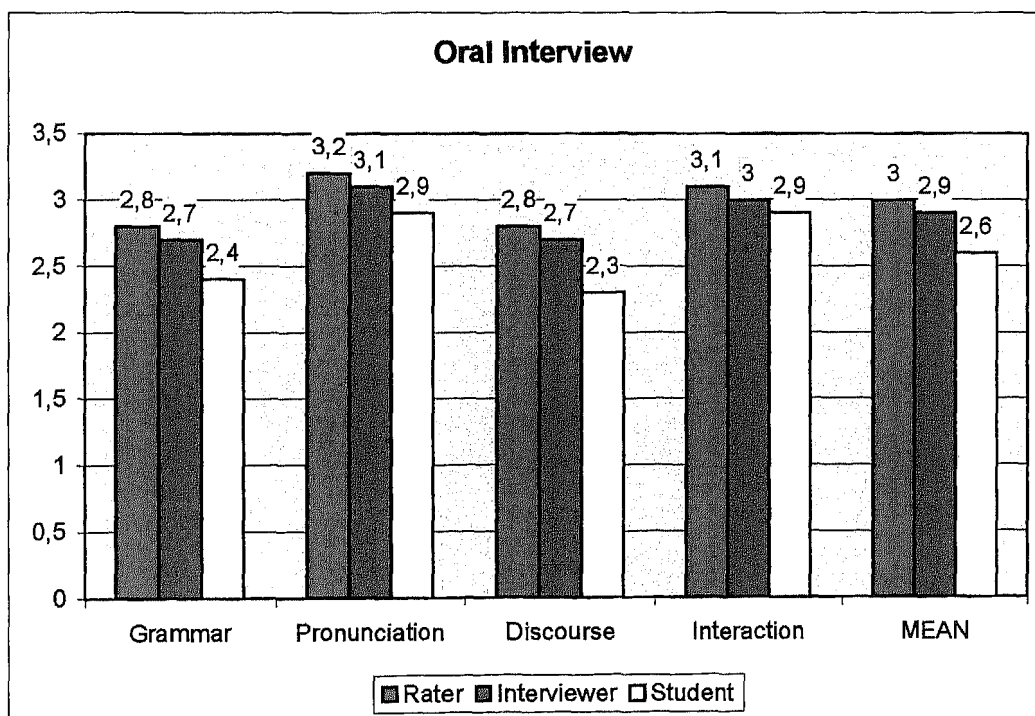
IV.1.1 Scores Obtained

First, we will compare the averages obtained in each of the variables related to the scores awarded by the rater, the interviewer and the students themselves in the 'Individual Oral Proficiency Interview' which consisted of a one-to-one interview between the interviewer and the student with a rater present for objectivity purposes. Since these are scale, or quantitative interval, variables they will be analysed using the

² Variables in which the order of data points can be determined but for which the numerical differences between adjacent attributes are not necessarily interpreted as equal, e.g. Likert scales.

Paired Samples *t*-Test³ since they compare two means referring to the same group of students.

The following bar chart represents an overall view of the results obtained for the individual oral interview test. They clearly show a general pattern that is repeated in every area scored: the rater consistently gives the highest score, the interviewer gives a slightly lower score (which in all but one case is not statistically significant) and the students always give the lowest rating for their own performance (on all but one occasion with statistical significance).



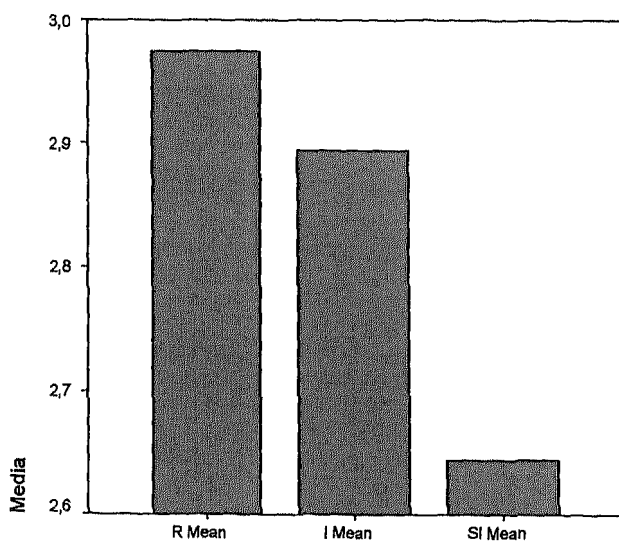
³ The Paired Samples *t*-Test compares the means of two variables. It computes the difference between the two variables for each case, and tests to see if the average difference is significantly different from zero.

a) **Mean Values**

First, we will analyse the *global means* of the specific scores awarded by the rater (R), the interviewer (I) and the student (SI) in the one-to-one test. The results obtained are presented in the following table and graph. At a glance, we can appreciate that the mean score awarded by the rater is the highest (2.97), closely followed by that of the interviewer (2.89). The student's self-awarded score is the lowest of all (2.64).

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. t.p.
R Mean	51	1,25	5,00	2,9750	,91434
I Mean	51	1,00	5,00	2,8946	,86218
SI Mean	51	1,25	4,50	2,6446	,69659
N válido (según lista)	51				



The following step is to compare the means obtained using the Paired Samples *t*-Test in order to detect any possible significant differences. We will firstly compare the mean score awarded by the rater with that awarded by the interviewer, and subsequently compare both of these mean scores with the mean self-assessment score of the student. So as not to interfere with the reporting of the results in this section, the results of the statistical analyses carried out have been included in Appendix 13.

In order to determine whether there are significant differences between the mean scores, it is necessary to focus on two principal aspects. If the result of the *t*-Test gives a probability of less than 0.05, or the confidence interval⁴ does not include the value 0, then we can assume that there is a significant difference between the two means.

Here, on comparing the global mean scores of the rater and the interviewer, it can be observed that the result of the *t*-test is 1.202, which gives a probability of 0.235 (greater than 0) and also a confidence interval of -0.0539 to 0.2147 which includes the value 0. For this reason we can conclude that there is no significant difference between these two mean scores. Likewise, there were no significant differences found on comparing the global means of the Interviewer and the Student. However, there were significant differences on comparing the global means of the rater and the student, where the *t*-test result was 2.553, giving a 0.014 probability (less than 0.05) and a confidence interval (0.0705 to 0.5903) which does not include 0.

We can therefore see that both subjects involved in the interaction, interviewer and student, have a similar perception of the action, while the rater, on the contrary,

⁴ A research study can show absolutely only the outcomes or results for the study participants themselves; the study results may not be true for others. The confidence interval (CI) is a mathematical description of how likely it is that others will have the same result as the study participants.

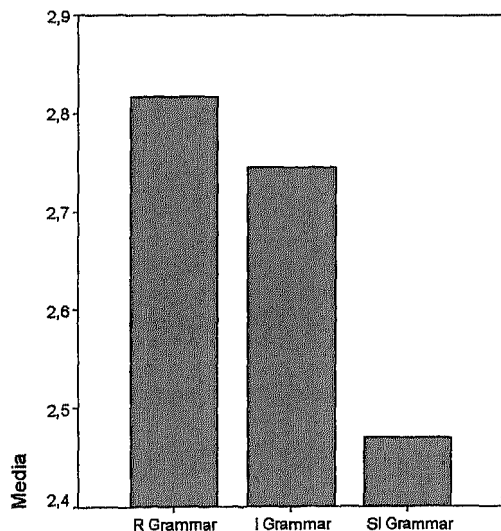
from a position of detachment and relative objectivity, has a significantly different impression of what is taking place, which in this case favours the student considerably. It is possible to postulate several reasons for this. The first is that, as is the case in many events in life, an outside observer has a much clearer overall view of a situation and is able to judge circumstances and performances with a greater degree of accuracy than those who are immediately involved in the course of the action. In this case, it may be that the rater has become aware of, and taken into consideration, the unbalanced power situation inherent to the interview format and has accounted for this in the scores s/he awarded. It is also possible that, since s/he is able to focus solely on the performance of the candidate, without having to pay attention to managing the test, or even necessarily to what the interviewer is doing or saying, the rater can pay attention to features of performance that neither the interviewer or the student notice. From the students' point of view, it may be that they are so overwhelmed by the intimidating power situation produced here that it completely obscures their capacity to evaluate their performance with any objectivity and they are aware only of the feeling of inferiority and anxiety they experience in their struggle to communicate and justify their opinions in the foreign language. Whatever the case, these overall results would indicate that there is indeed an important role to be played by an objective rater in the scoring procedure of speaking tests. If this figure is to be included, we can also infer that it will be necessary to address the situation created by the imbalance in power by introducing at least one other student into the testing procedure in order to create a 2-2 examiner/candidate ratio.

Our next step will be to analyse the data obtained in each of the categories of the oral interview that were assessed: 'Grammar and Vocabulary', 'Pronunciation', 'Discourse Structure' and 'Interaction'.

Grammar and Vocabulary

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Grammar	51	1,00	5,00	2,8176	1,00692
I Grammar	51	1,00	5,00	2,7451	,91855
SI Grammar	51	1,00	5,00	2,4706	,87984
N válido (según lista)	51				



We can see here that the scores awarded by each of the three subjects follow the general pattern of the overall mean scores. While there is no significant difference between the scores of the rater and the interviewer, nor between those of the interviewer and the student, if we compare the scores of the rater and the student, the

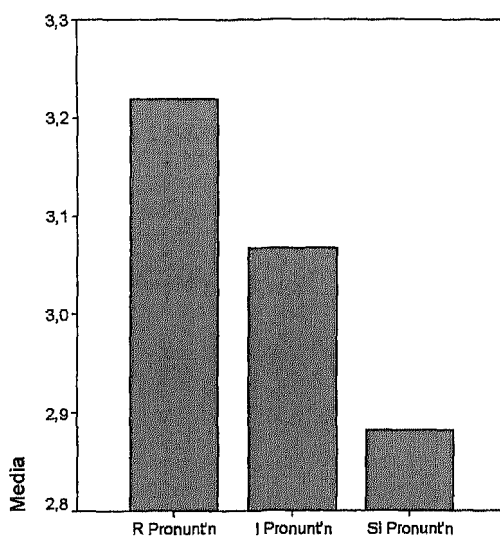
result of the *t*-Test is 2.062 which gives a probability of 0.044 (less than 0.05) and a confidence interval (0.0090 – 0.6851) which does not include 0 and therefore there is a statistically significant difference between the scores in the ‘Grammar and Vocabulary’ category.

The student self-assessment score shows that they have a significantly poorer perception than the rater of their control of grammatical structures when speaking and that they do not find it easy to access their vocabulary knowledge in the rapid real-time activity of speech. When grammar is not automatised, speaking in the foreign language generally produces a conflict between focusing on form and focusing on meaning so that when one prevails, the other is temporarily abandoned. In the one-to-one interview situation, the candidates are almost certainly paying attention to the message they wish to communicate to the interviewer, and are therefore aware of a greater struggle with grammatical structures. This may be a reason for them assessing their performance with a much lower score than the rater, whose attention is free to ‘notice’ how accurately they are using grammatical structures. The interviewer, on the other hand, will probably be more focused on the message in order to respond or contribute when necessary, and will be more likely to notice errors than correct structures which will just form part of the flow of speech. For this reason, they may give lower scores than the rater.

Pronunciation

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Pronunt'n	51	1,00	5,00	3,2196	,96727
I Pronunt'n	51	1,00	5,00	3,0686	,91662
SI Pronunt'n	51	1,00	5,00	2,8824	,86364
N válido (según lista)	51				



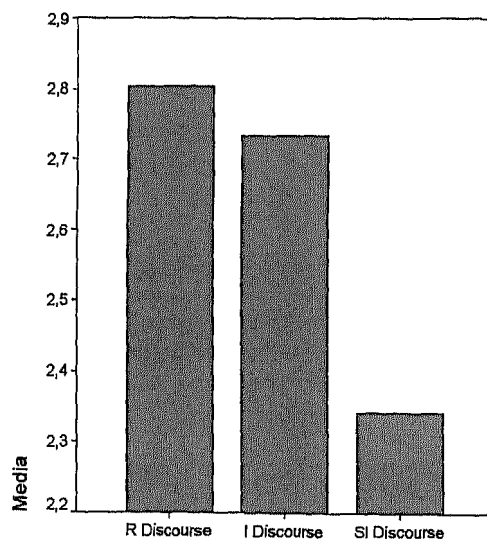
As in the previous two cases, there is a significant difference only between the mean scores of the student and the rater, with the student again giving the lowest score of the three although with much higher marks than before. It may be that it is very much more difficult to be aware of non-native-like pronunciation in one's own speech than in the speech of others; we do not even usually perceive ourselves to speak with any kind of accent in our first language, so it is unlikely that we will recognise our own features of pronunciation in a foreign language. Students therefore give

themselves a higher score in this category than in the previous one, although they continue with the general trend of lower scoring, marking themselves below the rater's judgement, probably caused by test anxiety and inferiority in the test situation.

Discourse Structure

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. t.íp.
R Discourse	51	1,00	5,00	2,8039	1,02994
I Discourse	51	1,00	5,00	2,7353	,92926
SI Discourse	51	1,00	4,00	2,3431	,80306
N válido (según lista)	51				



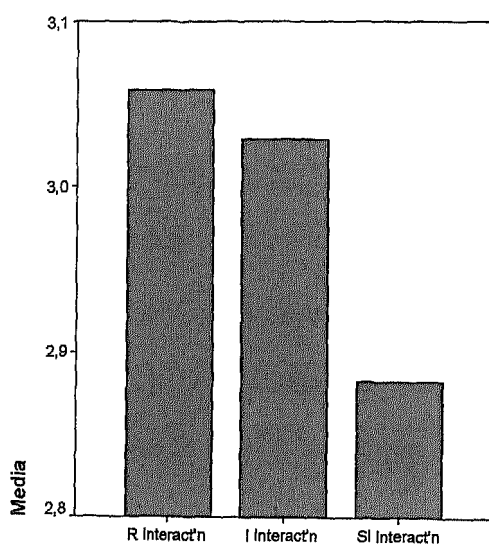
On judging performance in the category of 'Discourse Structure', we find that there are significant differences between both rater and student mean scores (probability 0.005 and confidence interval 0.1465 – 0.7750), and interviewer and student scores (probability 0.011 and confidence interval 0.0954 – 0.6889), although

there continues to be no significant difference between the mean scores of the rater and the interviewer. This is the only category where the interviewer and student scores show a significant difference, and also the category in which the students award themselves the lowest mark. This tendency to score lower here may be because the interview candidates are more aware of their struggle to organise ideas and grammatical structures coherently in comparison with the way these thoughts and ideas are ordered in their L1, and hence their perception of this organisation is fairly negative. It is also possible that listeners are, in fact, much more patient and accommodating than we tend to imagine as foreign language speakers, and that both the rater and interviewer have an internalised understanding of the speed of delivery that they would expect from candidates at this level which may be quite a lot slower than for native speech, thus accounting for the significant difference for both examining roles. A further possibility that might account for the students' feeling of inadequacy in structuring discourse in English is that we do not provide them with enough repair strategies in the foreign language which native speakers so frequently use to give themselves time and space to organise what they want to say. Extensive exposure to spoken English would be necessary for students to notice and acquire these on their own, and this may be an argument in favour of including this type of strategy teaching in our classrooms.

Interaction

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. t.p.
R Interact'n	51	1,50	5,00	3,0588	1,09839
I Interact'n	51	1,00	5,00	3,0294	1,10640
SI Interact'n	51	1,00	4,00	2,8824	,81602
N válido (según lista)	51				



Here, there are no significant differences between the means of any of the scores. The reasons why students may feel both their interaction and pronunciation skills are superior to those of grammatical accuracy and discourse structure are unclear. There may be more socio-affective factors in play in the perception of these features of the speaking construct, closely related to the subjects' perception of self-worth. Despite recognising in themselves a lack of formal language knowledge, the students rate their competence at interacting, which depends more on socio-affective strategies than on cognitive ones, much more positively.

b) Correlation Study

On completing the means comparison study, and having found that in all cases there were no significant differences between the mean scores awarded in any of the categories by the rater and the interviewer, we decided to carry out a further analysis to determine whether, in contrast, there was a high incidence of correlation⁵. Bearing in mind that the variables we wished to compare were quantitative and that we were also interested in examining the possible relationship between them, we decided to use the Pearson correlation test⁶ (Camacho Rosales, 2000: 258). This test is one of the most commonly employed in Applied Linguistics research (e.g. Cristóbal Ruano, 1992).

With the results obtained using the Pearson test (see Appendix 14), we can observe that in all cases there is a very high correlation between the scores awarded by the rater and the interviewer, at a level of 0.01, which is highly statistically significant since it is much greater than 0.05. It is possible that this was brought about in part by the examiners tendency to compare and discuss their scores as they filled in the mark sheets, despite specific instructions to the contrary. It is natural in a new or unknown situation for human beings to require corroboration that they are going about their task correctly and therefore the inclination is to try to reach some kind of consensus; very few of the marks recorded guided by the new rating scale were greatly disparate. This may have been a failing of the standardisation procedure, which should have given more practice and feedback on using the scales before the test and stricter instructions

⁵ This is the figure we obtain by applying a formula to two sets of data to test whether or not they are associated. If the variables are highly positively correlated, the figure we obtain will approach 1. If there is no relationship whatsoever, the figure will approach zero, and if there is a strong negative correlation, the figure will approach -1.

⁶ The Pearson test also gives the probability ('Sig. bilateral'). If this probability is less than 0.05, it can be deduced that the correlation is statistically significant.

to the examiners. However, if in fact the raters and interviewers did not compare the scores they awarded candidates, we may also consider these results to reflect the success of the new rating scale in providing stable and well-defined criteria for assessing speaking skills.

Of much greater interest is the correlation study carried out to compare the scores of the rater and the student. Here, a significant correlation (0.01) can be observed between the mean values of the scores recorded by the student and the rater. There are also correlations, although at the lower level of 0.05, between the rater and student scores in the categories of 'Pronunciation', 'Discourse Structure' and 'Interaction'. However, there is no correlation between the scores recorded by the rater and the student in the 'Grammar and Vocabulary' category.

In this case, it is impossible that the students and raters conferred during the scoring procedure and the correlations can therefore be said to be accurate. We can infer from this that although students may perceive their performance to be weaker than the rating of an objective observer, the general pattern of recognition of their strengths and weaknesses is, in all but one category, the same. This may indicate that there is a favourable argument for including self-assessment in both our testing and teaching programmes.

IV.1.2 Data from Questionnaire 1 (Student)

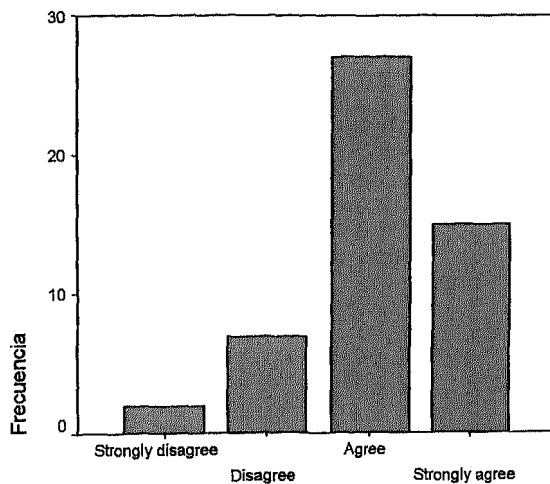
As explained above (see p.162), the last 15 variables (numbers 20 to 34) correspond to the 15 items on Questionnaire 1. In our analysis of the data, we will present details of the study of the descriptive data obtained for each item. Here, it is necessary to bear in mind that these are quantitative ordinal variables which

correspond to a Likert scale (1 – 4), and for this reason the Wilcoxon test⁷ was used to compare the pairs of items that elicited the opinion of the students on contrasting their global mark with the analytic mark obtained on the ‘Individual Oral Proficiency Interview’.

Question 1: *I felt nervous throughout the whole test.*

Quest-I 1

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	2	3,9	3,9	3,9
	Disagree	7	13,7	13,7	17,6
	Agree	27	52,9	52,9	70,6
	Strongly agree	15	29,4	29,4	100,0
	Total	51	100,0	100,0	



Quest-I 1

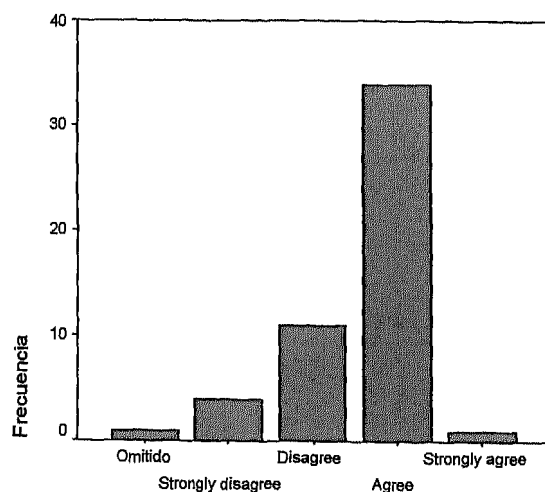
⁷ A statistical test of the equality of similar or matched groups of data to determine whether they differ significantly from one another.

Here we can see that the vast majority of students (82.3%) felt nervous throughout the test, with many expressing that they experienced a high level of stress in the one-to-one test situation. This result was in agreement with our original hypothesis that this test format creates a heavily unbalanced power situation which causes great anxiety in the candidates.

Question 2: *I think I did well in the test.*

Quest-I 2

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	4	7,8	8,0	8,0
	Disagree	11	21,6	22,0	30,0
	Agree	34	66,7	68,0	98,0
	Strongly agree	1	2,0	2,0	100,0
	Total	50	98,0	100,0	
Perdidos	Sistema	1	2,0		
Total		51	100,0		



Quest-I 2

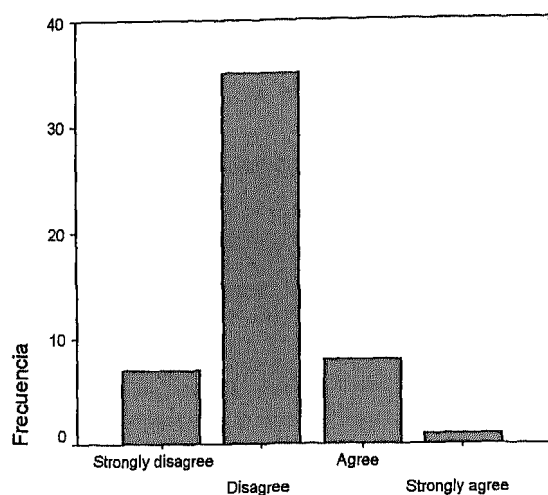
The graph shows that majority of students (70%) felt that they performed well on the test.⁸ These data are interesting, since although we would have expected a more negative perception of test performance given the level of anxiety expressed and taking into account the self-assessment scores, the majority of students still appear to leave the interview room with their self-esteem intact. We may give credit here to the interviewers who, despite the unfavourable conditions for doing so, were able to put the candidates at their ease in most cases and give them a positive impression of the way they had coped with the situation.

Question 3: *I performed to the best of my ability in the test.*

Quest-I 3

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	7	13,7	13,7	13,7
Disagree	35	68,6	68,6	82,4
Agree	8	15,7	15,7	98,0
Strongly agree	1	2,0	2,0	100,0
Total	51	100,0	100,0	

⁸ It should be noted here that in cases where the student did not respond to an item, the data is recorded as missing (*Perdido* (table); *Omitido* (graph)).



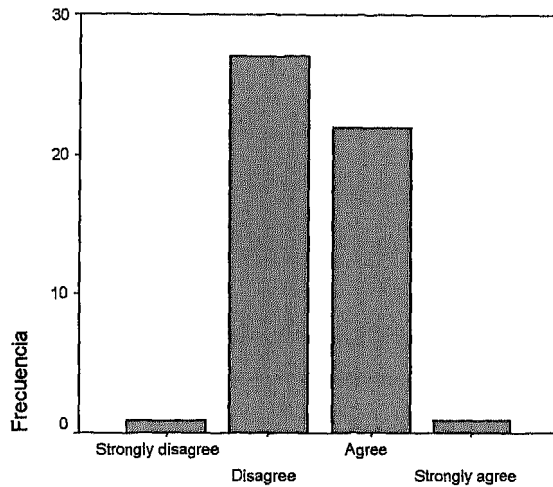
Quest-I 3

Here, in contrast to the previous graph which showed that the majority felt that they had performed well in the test, most of the students (82.3%) clearly thought that they had not performed to the best of their ability. Again, we would have expected this due to reasons of anxiety and also to the restrictive nature of the test format, where the control of the discourse is clearly held at all times by the interviewer.

Question 4: *I think I spoke enough for the tester to judge my ability.*

Quest-I 4

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	1	2,0	2,0	2,0
Disagree	27	52,9	52,9	54,9
Agree	22	43,1	43,1	98,0
Strongly agree	1	2,0	2,0	100,0
Total	51	100,0	100,0	



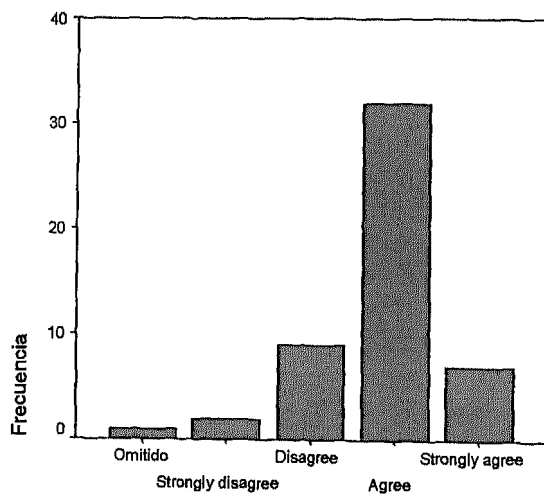
Quest-1 4

Here, we can see that while slightly more students felt that they had not produced a large enough speech sample for the tester to be able to make an accurate judgement of their speaking ability, just under half had the contrary impression of having spoken enough during the test. In contrast to our belief at the outset that there is a tendency for interviewers to speak more than, or at least as much as, candidates, this balance between agreement and disagreement shows that the test format itself probably did not have a very important impact on the amount of speech the candidates felt they were able to produce in the time allowed for each test, but that these impressions depended more on individual perceptions.

Question 5: *I was happy about the procedure of the test.*

Quest-I 5

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	2	3,9	4,0	4,0
	Disagree	9	17,6	18,0	22,0
	Agree	32	62,7	64,0	86,0
	Strongly agree	7	13,7	14,0	100,0
	Total	50	98,0	100,0	
Perdidos	Sistema	1	2,0		
Total		51	100,0		



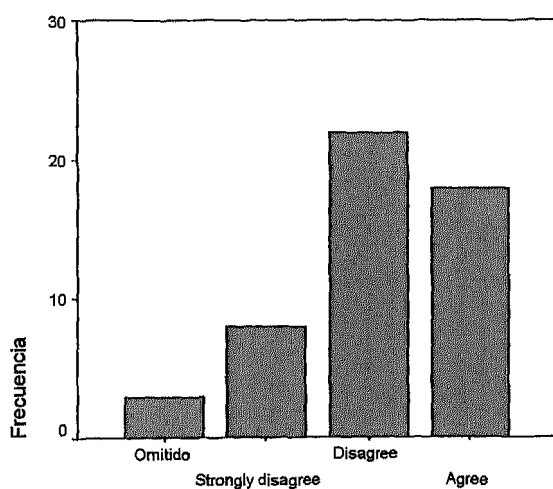
Quest-I 5

This graph shows that the great majority of students who took the individual oral test (78%) were happy with the procedure, despite the majority claiming to have found it stressful and having the impression that it did not allow them to perform to the full extent of their ability. This reflects the result found above, that the students were generally happy with their test performance despite the negative aspects they express in other responses, and once again gives credit to the professional test management of the interviewers.

Question 6: *The test was similar to the kind of task done in class.*

Quest-I 6

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	8	15,7	16,7	16,7
	Disagree	22	43,1	45,8	62,5
	Agree	18	35,3	37,5	100,0
	Total	48	94,1	100,0	
Perdidos	Sistema	3	5,9		
Total		51	100,0		



Quest-I 6

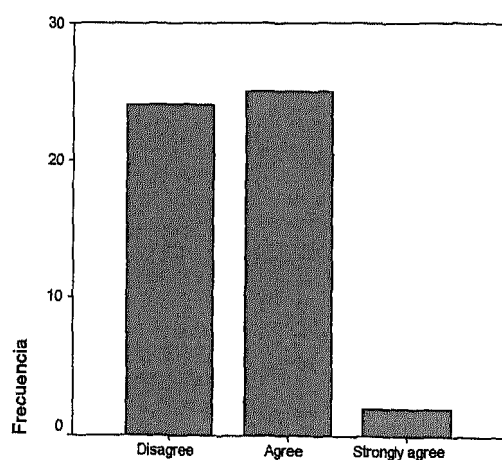
As we can see here, most of the students (62.5%) thought there was little or no similarity between speaking tasks carried out in the classroom and the one-to-one oral test. Furthermore, for this item there were no cases of 'Strongly agree'. This was an expected finding, since it is very unusual for classroom practice to include one-to-one interview-type discourse; students engage in co-operative or collaborative tasks and the teacher may join a group and talk to its members, but will rarely be involved in one-to-one interaction unless it is for dealing with problems or requests. This is one of

the reasons for attempting to change the speaking test procedure, since the oral interview does not bear any resemblance to the content of the course programme.

Question 7: *I could answer the questions without difficulty.*

Quest-I 7

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	24	47,1	47,1	47,1
	Agree	25	49,0	49,0	96,1
	Strongly agree	2	3,9	3,9	100,0
Total		51	100,0	100,0	



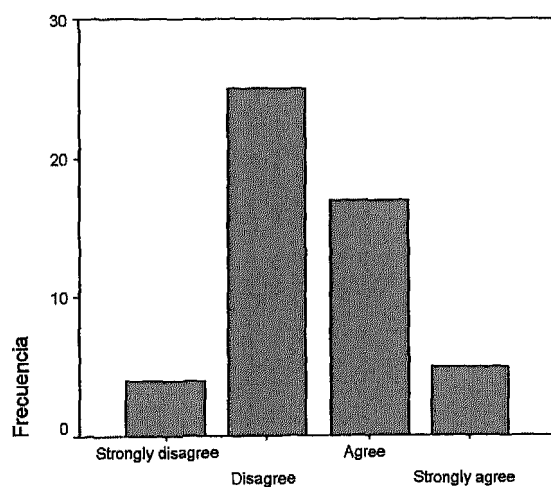
Quest-I 7

Although we can see that there were no cases of 'Strongly disagree' for this item, the students' opinions about the questions they were asked during the test are almost equally divided between the positive and negative. As in the case of Item 4 above, we can postulate that these opinions are due more to individual impressions than to the test format or materials, and hence they do not lead us to draw any conclusions about either of these in relation to the performance relevant to this test.

Question 8: *I could find enough to say about the topic.*

Quest-I 8

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	4	7,8	7,8	7,8
Disagree	25	49,0	49,0	56,9
Agree	17	33,3	33,3	90,2
Strongly agree	5	9,8	9,8	100,0
Total	51	100,0	100,0	



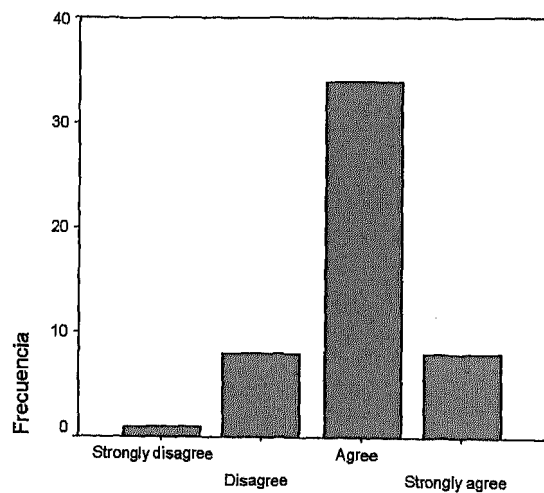
Quest-I 8

Again, we can observe that the students are almost equally divided in their views on whether they could find enough to say about the topic of their test, although a few more reported that they found that they did not have enough to say, while in the previous item, more students thought that the questions were not difficult to answer. We would, therefore, likewise assume that these responses are due to individual differences that we cannot account for in this study.

Question 9: *The global mark I received was a fair mark.*

Quest-I 9

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	1	2,0	2,0	2,0
Disagree	8	15,7	15,7	17,6
Agree	34	66,7	66,7	84,3
Strongly agree	8	15,7	15,7	100,0
Total	51	100,0	100,0	



Quest-I 9

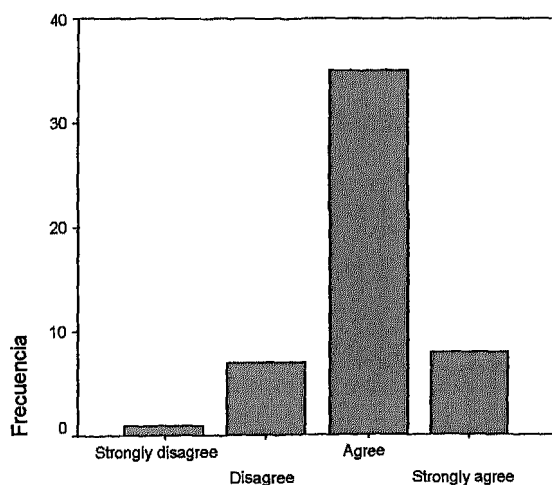
The vast majority of students (82.4%) agreed that they had received a fair global mark for their test. This could be expected given our observations in the first part of the study in the section relating to scoring. The rater always gave the students a higher score than their self-assessment and, human nature being what it is, it is fairly unlikely that many students would have claimed they thought their score was unfairly high. It is also important to remember here that the 0 – 10 marking scale is the one

students are totally familiar with and for which they have an internalised concept of meaning.

Question 10: *The analytic mark I received was a fair mark.*

Quest-I 10

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	1	2,0	2,0	2,0
	Disagree	7	13,7	13,7	15,7
	Agree	35	68,6	68,6	84,3
	Strongly agree	8	15,7	15,7	100,0
	Total	51	100,0	100,0	



Quest-I 10

Again, the majority of students (84.3%) felt that their analytic score was a fair mark which is in concordance with the response to the previous item.

Since, as we have seen above, Questions 9 and 10 contained identical wording with the exception that the first item referred to the global mark awarded while the

second was concerned with the analytic mark, we felt that it was of interest to explore any differences or similarities between the answers given by the 51 students to these two questions. The Wilcoxon test carried out on the two related samples gave the following results:

Rangos

	N	Rango promedio	Suma de rangos
Quest-I 10 - Quest-I 9 Rangos negativos	4 ^a	5,00	20,00
Rangos positivos	5 ^b	5,00	25,00
Empates	42 ^c		
Total	51		

a. Quest-I 10 < Quest-I 9

b. Quest-I 10 > Quest-I 9

c. Quest-I 9 = Quest-I 10

As we can see in the table of ranges above, 42 students considered that both their global and analytic marks were fair. Of the nine who did not do so, four gave a higher value to the global mark (1 – 10 scale), and five to the analytic mark. As shown in the table below, this difference is not significant.

Estadísticos de contraste^b

	Quest-I 10 - Quest-I 9
Z	-,333 ^a
Sig. asintót. (bilateral)	,739

a. Basado en los rangos negativos.

b. Prueba de los rangos con signo de Wilcoxon

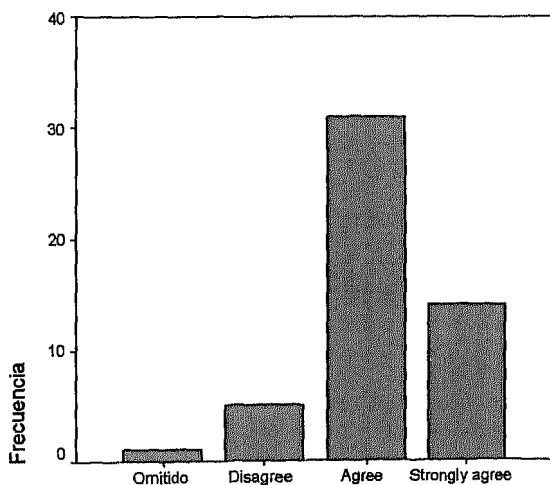
This result is slightly disappointing; we had hoped that the analytic score would meet with greater acceptance than the global mark since its meaning was actually

described on the score sheet the students had used. We must assume here that the influence of traditional scoring methods is deeply ingrained and that evidently it is not easy to change.

Question 11: *I understand what my global mark means.*

Quest-I 11

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	5	9,8	10,0	10,0
	Agree	31	60,8	62,0	72,0
	Strongly agree	14	27,5	28,0	100,0
	Total	50	98,0	100,0	
Perdidos	Sistema	1	2,0		
Total		51	100,0		



Quest-I 11

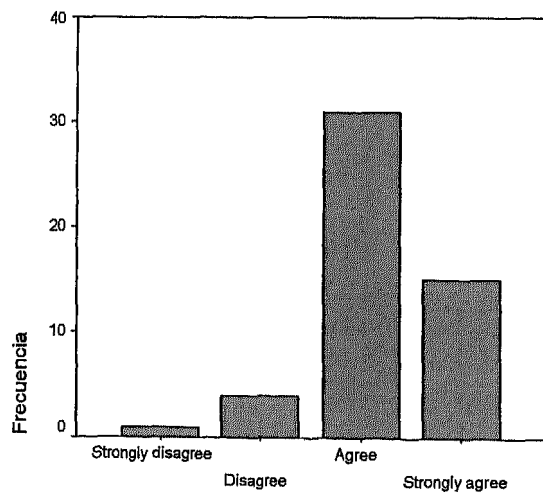
The graph shows that nearly all the students (90%) felt that they understood the meaning of the global mark they received, which follows the pattern seen in the Wilcoxon test for the previous question above. The 0 – 10 scale is so familiar that its meaning has been internalised and this seems to have become synonymous with

assigning the marks an explicative meaning which does not exist in reality beyond the concept of pass or fail and where an individual is placed on a linear scale of achievement in relation to others who took the test at the same time.

Question 12: *I understand what my analytic mark means.*

Quest-I 12

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	1	2,0	2,0	2,0
Disagree	4	7,8	7,8	9,8
Agree	31	60,8	60,8	70,6
Strongly agree	15	29,4	29,4	100,0
Total	51	100,0	100,0	



Quest-I 12

Again, nearly all the students (90.2%) agreed that they understood the meaning of their analytic mark, with only five students disagreeing. This result is more encouraging, since at least the students, while not contrasting one type of score with

another, recognise that the scale descriptors say something about their strengths and weaknesses in different aspects of the speaking construct.

Since, as we have seen above, these two questions compare the students' opinion of their level of understanding of the global and analytic marks they received, so once again, the Wilcoxon test was used to provide comparative data of the answers given in each case:

Rangos

	N	Rango promedio	Suma de rangos
Quest-I 12 - Quest-I 11 Rangos negativos	3 ^a	3,50	10,50
Rangos positivos	3 ^b	3,50	10,50
Empates	44 ^c		
Total	50		

a. Quest-I 12 < Quest-I 11

b. Quest-I 12 > Quest-I 11

c. Quest-I 11 = Quest-I 12

Here we find a balance between the results, since of the 50 students who answered both questions, 44 assigned the same value to both types of scoring procedures. The remaining six are equally divided, three valuing the global mark more highly and three giving greater value to the analytic mark, which not only means there is no statistical significance between the two, but also that absolute equality exists between the ranges.

Estadísticos de contraste^b

	Quest-I 12 - Quest-I 11
Z	,000 ^a
Sig. asintót. (bilateral)	1,000

a. La suma de rangos negativos es igual a la suma de rangos positivos.

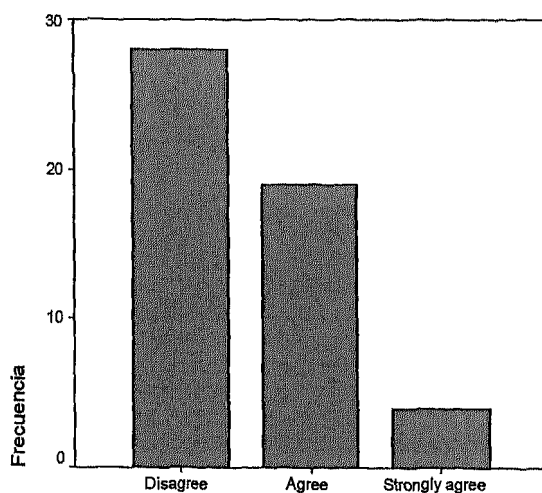
b. Prueba de los rangos con signo de Wilcoxon

It seems that in order to introduce new methods of scoring, students will require prior training in valuing and understanding them; we cannot take for granted that since they seem clear and meaningful to us they will necessarily be accepted as such by their primary end-users. However, we should emphasise that the reception reflected in these results is not negative either; the new scale is simply valued equally with the traditional system which, due to years of accumulated experience with it on the part of its users, is to be expected.

Question 13: *The global mark I received was easier to understand than the analytic mark.*

Quest-I 13

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Disagree	28	54,9	54,9	54,9
Agree	19	37,3	37,3	92,2
Strongly agree	4	7,8	7,8	100,0
Total	51	100,0	100,0	



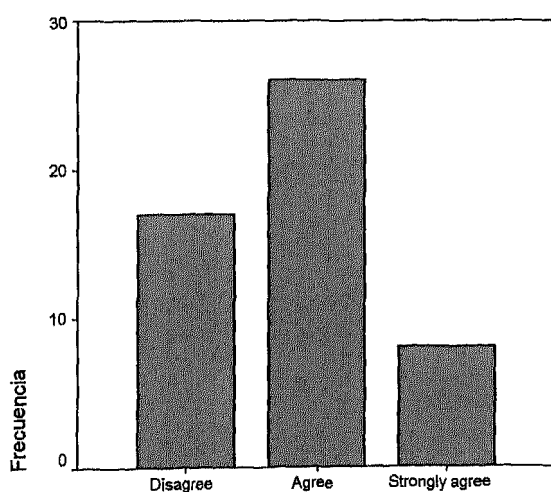
Quest-I 13

Although there are no cases of strong disagreement here, we can observe that more students disagreed that the global mark they received was easier to understand than the analytic mark. Again, we believe that having been accustomed to the traditional scale for so long accounts for this response and that a degree of open-mindedness which could be cultivated is shown.

Question 14: *The global mark helped me to understand what steps I need to take to improve my speaking.*

Quest-I 14

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	17	33,3	33,3	33,3
	Agree	26	51,0	51,0	84,3
	Strongly agree	8	15,7	15,7	100,0
	Total	51	100,0	100,0	



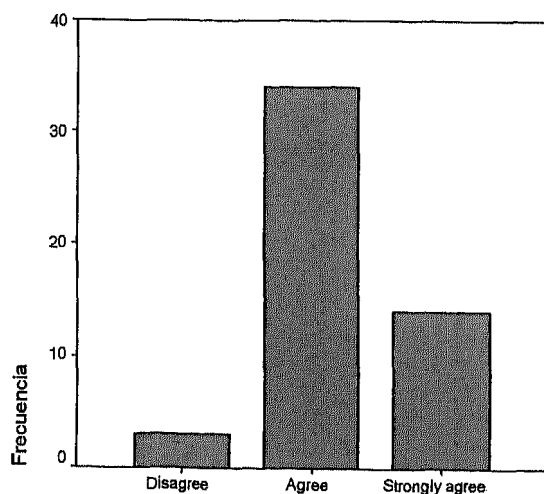
Quest-I 14

Here we can see that while most students agreed that the global mark they received helped them to understand the steps they needed to take in order to improve their speaking skills, a relevant number did not agree with this statement. Again, there were no cases of *Strongly disagree* for this item.

Question 15: *The analytic mark helped me to understand what steps I need to take to improve my speaking.*

Quest-I 15

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	3	5,9	5,9	5,9
	Agree	34	66,7	66,7	72,5
	Strongly agree	14	27,5	27,5	100,0
	Total	51	100,0	100,0	



Quest-I 15

We can observe here that many more students (94.2%) felt that an analytic score helped them to understand the steps they needed to take in order to improve their speaking skills. adding to our understanding of the usefulness to the students themselves of the test marks they receive. Again, we can see that there are no cases of ‘Strongly disagree’ and only three students disagreed with the statement. If we contrast this data with that shown for the previous item, we find that many more students appreciate the usefulness of being provided with descriptive statements about their speaking ability in order to understand which areas they need to concentrate on in order to improve. For this reason, as with the above items where we used the Wilcoxon test to compare two similar questions, we contrasted the data collected from Questions 14 and 15 which referred to the way in which the scores contributed to the students’ understanding of how to improve their speaking. The following ranges were obtained:

Rangos

	N	Rango promedio	Suma de rangos
Quest-I 15 - Quest-I 14 Rangos negativos	2 ^a	7,50	15,00
Rangos positivos	17 ^b	10,29	175,00
Empates	32 ^c		
Total	51		

a. Quest-I 15 < Quest-I 14

b. Quest-I 15 > Quest-I 14

c. Quest-I 14 = Quest-I 15

Here, it can be observed that most students (32 of 51) assign an equal value to the usefulness of the two methods of scoring as a means of indicating how to go about

improving their speaking skills. However, of the remaining 19 cases, the majority value the analytic score more highly: 17 cases against only 2 who show a contrary opinion. The following table determines the statistical significance of this data:

Estadísticos de contraste^b

	Quest-I 15 - Quest-I 14
Z	-3,386 ^a
Sig. asintót. (bilateral)	,001

- a. Basado en los rangos negativos.
- b. Prueba de los rangos con signo de Wilcoxon

The test shows that a significant difference (0.001) exists between Items 14 and 15. That is to say, a significantly greater number of students thought that the analytic mark helped them to better understand the steps they needed to take in order to improve their speaking skills. This finding provides support for our original hypothesis that students would find the analytic score more useful as an indicator for how to proceed in order to improve their speaking skills, and encouragement for the continued implementation and development of the scale in further testing sessions.

IV.1.3 Data from Questionnaire 2 (Interviewer)

Since just four interviewers took part in the testing sessions, we will present the results of the questionnaires they answered in the form of a table which repeats the format of the original questionnaire. The numbers indicate how many respondents chose each of the answers.

QUESTIONNAIRE 2 (Individual Oral Proficiency Interview)

	Strongly disagree	Disagree	Agree	Strongly agree
1. I was able to manage the interview and give the student a global mark on a scale of 1-10			3	1
2. I was able to manage the interview and give the student a detailed score at the end of the interview		1	3	
3. I was more focused on managing the interview than on the rating criteria			1	3
4. I felt comfortable in the dual role of interviewer and rater		2	1	1
5. I felt happy about the test procedure		1	1	2
6. The student produced a large enough speech sample for assessment			4	
7. It was easy to assess how well the candidate was interacting		1	3	
8. I understood what I was assessing in giving the global mark		1	1	2
9. I understood what I was assessing in giving the analytic score			1	3
10. The most important part of my assessment in giving the global mark was grammatical accuracy	2	1	1	
11. The most important part of my assessment in giving the detailed score was grammatical accuracy	2	2		
12. I think I awarded the student a fair mark in giving the global mark		1	2	1
Reason:				
13. I think I awarded the student a fair mark in giving the analytic score		1	2	1
Reason:				
14. It was easier to mark a student who expressed an opinion similar to mine in giving the global mark	1	3		
15. It was easier to mark a student who expressed an opinion similar to mine in giving the analytic score	1	3		

In order to contrast the perspectives of the raters/interviewers with those of the students in relation to the individual interview test format and to obtain an overall view of the effects of test format and scoring procedures, the results obtained from this questionnaire will be discussed with reference to the research questions they address at the end of this chapter (Section 3).

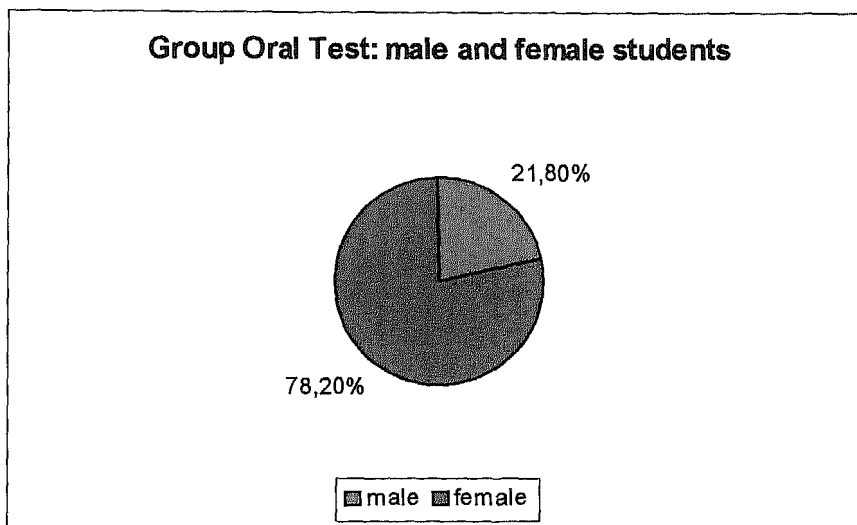
IV.2 'GROUP SPEAKING TEST'

The total number of students who took the 'Group Speaking Test', completed the self-evaluation sheets, and subsequently filled in the questionnaire was 78. Of these, 52 (67%) achieved a pass mark for the test according to the score awarded by the rater. We should point out here that more students took this test because it formed part of the final exam for the subject *Lengua BII*; for the previous one-to-one interview, participation was voluntary. This test consisted of three students, an interlocutors and a rater.

Of the total number of students, 78.2% were female and 21.8% male.

Male/Female

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Male	17	21,8	21,8	21,8
	Female	61	78,2	78,2	100,0
	Total	78	100,0	100,0	



As explained earlier in Chapter 3, in this second test data was obtained through the *Group Speaking Test Assessment Sheet* where the rater recorded scores in the categories ‘Grammar and Vocabulary’, ‘Pronunciation’, ‘Discourse Structure’ and ‘Interaction’ on a scale from 0 – 5 for each candidate taking the test. In this case, the interlocutor only gave a global, rather than an analytic, mark to each student, but also on a 0 – 5 scale with a descriptor for each score (see Appendix 3).

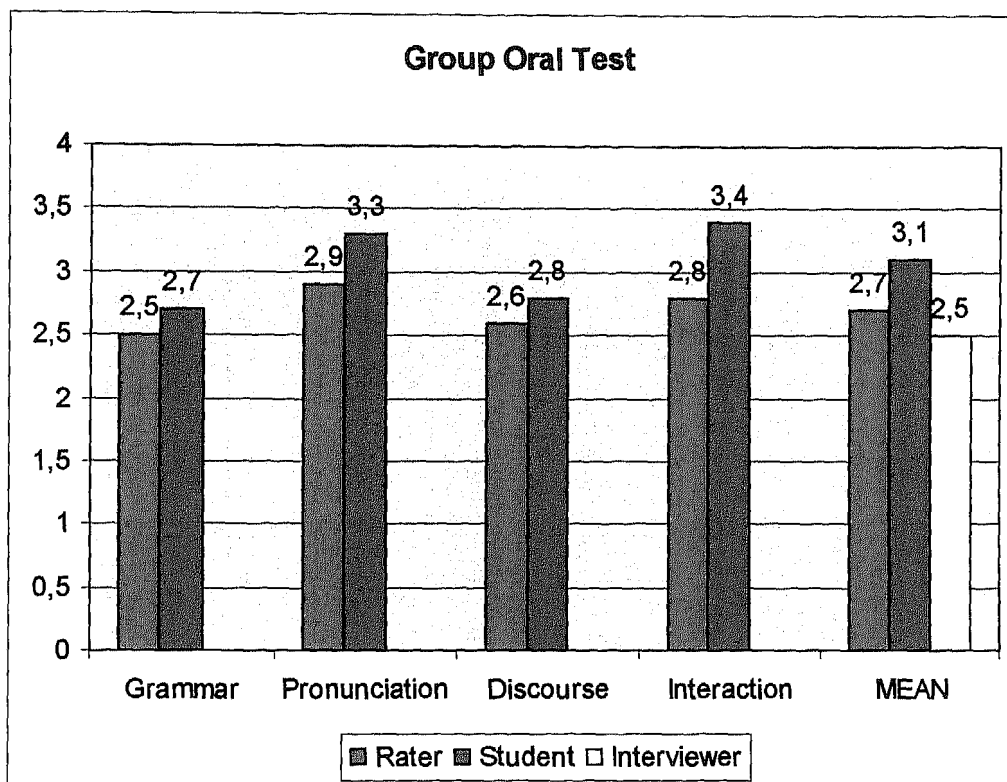
The second instrument for data collection was the *Self-assessment Sheet: Speaking (Group Speaking Test)* (Appendix 11) where the students filled in scores in the same categories as those mentioned above, using the slightly re-worded descriptors (see Appendix 3). Finally, the students answered Questionnaire 3 (see Appendix 5) from which it was possible to obtain information about their opinions on the format and scoring procedures of the second test by means of the analysis of their responses to 17 items.

After transferring these data to the SPSS statistics package (see Appendix 15) we began the statistical study by firstly determining the type of variables with which

we would work. Of the 31 variables, the first, (student) and second (male/female) are nominal since they only identify the student and their sex. Similarly, variable 14 (test pack) is also nominal because it only tells us which of the ten packs of test materials was used in each case. Eleven variables (3 – 13) represent the scores awarded to the student during the test by the rater, the interviewer and the student themselves. These make up the quantitative interval variables. The last 17 variables (15 – 31) represent the answers given by the students to the 17 questions on the questionnaire; they are all quantitative ordinal variables.

IV.2.1 Scores Obtained

The graph below gives an overall view of the results obtained for the ‘Group Speaking Test’. Again, they show a general pattern that is repeated across all the areas scored, but this time, in contrast to the ‘Individual Oral Proficiency Interview’, the students consistently award themselves the highest score. In this test, the Interviewer was not involved in assessing all the categories and only gave a global score at the end of the test, and for this reason the graph only reflects one score (the mean) for the Interviewer. In this case, the pattern is the same as for the ‘Individual Oral Proficiency Interview’: the interviewer gives the lowest score of all.

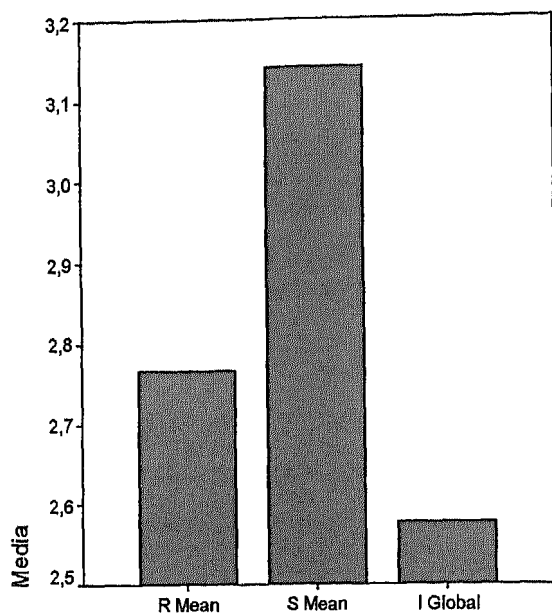


a) Global Values

We will begin by comparing the global mean score, consisting of a mean value over the four categories, of the rater (R), the interviewer (I) and the student (SI). The results obtained are presented in the table and graph below.

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Mean	78	1,3	5,0	2,765	,7325
S Mean	78	1,8	4,8	3,141	,5845
I Global	78	1,0	5,0	2,577	,8085
N válido (según lista)	78				



We can immediately see that in this case, and in complete contrast to the previous test where the students' self-awarded mark was always the lowest, the score given by the student is the highest of the three (3.14), while of the marks awarded by the examiners, the rater's mark is higher than that of the interviewer as occurred in the first test (2.76 contrasted with 2.57).

We will now compare the mean scores obtained (using the paired sample *t*-test) in order to detect any possible significant differences. Once again, we have included the statistical results of the test in Appendix 16 in order to facilitate the reading of the present section.⁹ In this case, on comparing the global mean scores of the rater and the interviewer, we can see that the result of the paired *t*-test was 2.879 which gives a probability value of 0.005 and a confidence interval of 0.058 to 0.319. This does not include 0 and therefore we can conclude that there is a significant difference between

⁹ As we saw above, in order to determine whether there are significant differences between the means obtained, it is necessary to focus on at least two aspects. If the result of the *t*-test gives a probability (Sig. bilateral) of less than 0.05 or the confidence interval does not include 0, we can consider that there is a significant difference between the two means.

the two means, that is, the rater's mean score is significantly higher than that of the interviewer. This contrasts with first test, where no statistical significance was found between the rater and interviewer scores. These results are interesting since initially we had predicted that there would be a greater coincidence between the rater and Interviewer scores in the group speaking test than in the one-to-one interview, since the interviewer here is able to take a much more objective view of the interaction. However, what we seem to find is that, on a giving a global mark the interviewer may still be focused more on the message than on the discrete features of the speaking construct and therefore tends to listen more for 'mistakes' and hence awards a lower score than the rater who is perhaps also paying attention to positive features of candidate performance.

The comparison between the mean score of the rater and the student results in a t value of -4.725 with a probability of $.000$. Similarly, on comparing the mean self-assessment score of the student with that of the interviewer ($t = 5.568$), we again obtain a probability of $.000$. Therefore, the differences between the mean scores of the student compared with those of both the rater and the interviewer have a notable statistical significance in the group test format, whereas in the individual oral interview there was only statistical significance between the scores of the rater and the student. In this case, the students assess their own performance much more positively than either of the examiners, possibly due to greater self-confidence inspired by the group test format.

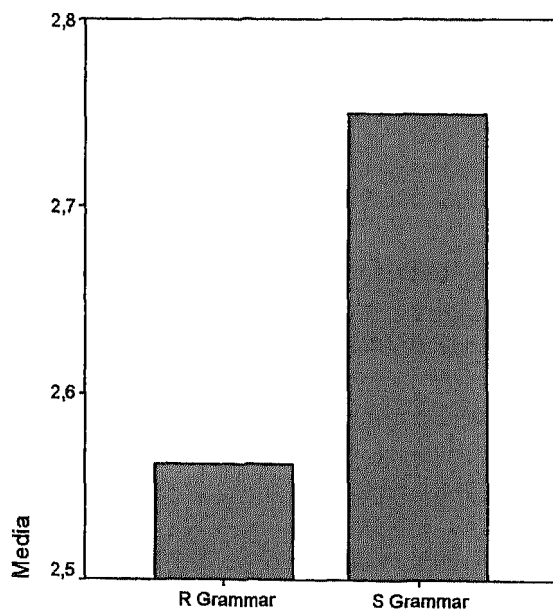
As mentioned above, the 'Group Speaking Test' is characterised amongst other things by the fact that it only provides data for the specific scores given by the rater and each student, since here the interlocutor only gives a global assessment on a 0 – 5

scale with descriptors. However, the rater categories for assessment ('Grammar and Vocabulary', 'Pronunciation', 'Discourse Structure' and 'Interaction') are the same as in the 'Individual Oral Proficiency Interview'. We will therefore now proceed to present the data obtained in these categories for the 'Group Speaking Test'.

Grammar and Vocabulary

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. t.íp.
R Grammar	78	1,0	5,0	2,562	,8013
S Grammar	78	1,0	5,0	2,750	,7151
N válido (según lista)	78				



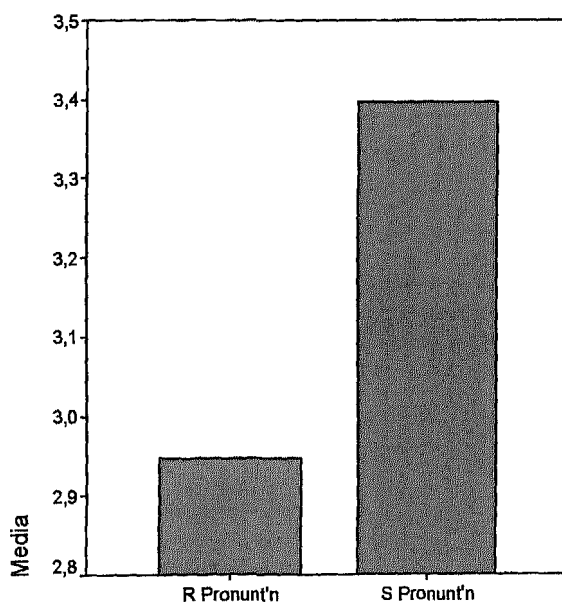
The statistical comparison (paired sample *t*-test) between the mean score of the rater and that of the student results in a *t* value of -1.944 with a probability of 0.055. Since this is greater than 0.05 and the confidence interval (see Appendix 16) includes

0, we can affirm that there is no statistical difference between the mean score of the rater and that of the student in the 'Grammar and Vocabulary' category. This is in contrast to the first test, where the difference between the rater and student scores was statistically significant, but with the rater giving the higher score.

Pronunciation

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. ttp.
R Pronunt'n	78	1,0	5,0	2,946	,7272
S Pronunt'n	78	2,0	5,0	3,397	,7786
N válido (según lista)	78				



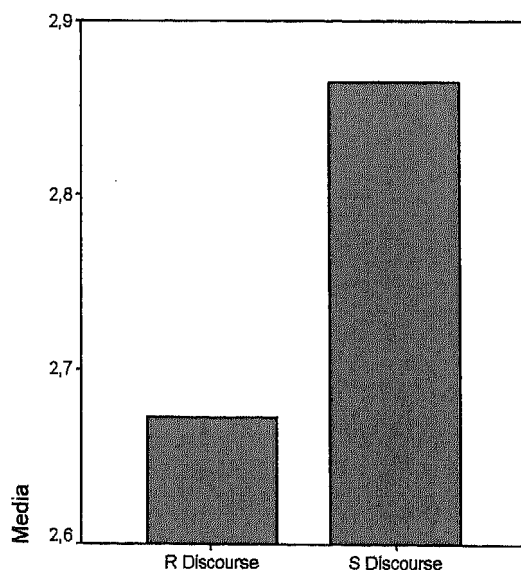
The Paired Samples *t*-Test gives a result of -4.696 which corresponds to a probability of .000. The confidence interval does not include 0 and thus demonstrates that the results, as in the one-to-one interview, show a highly significant statistical

difference. However, on this occasion, the scoring trends are reversed, with the student self-assessment mark being the higher of the two. A possible explanation for this is the presence of a more balanced power situation, where the students are comparing their own pronunciation to that of their peers and probably finding more similarities than with that of the interviewer in the one-to-one format. The fact that the students all have the same L1 also probably means that the non-native like variations in their pronunciation are fairly similar and are less likely to be perceived than in a group with different L1 speakers.

Discourse Structure

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Discourse	78	1,0	5,0	2,673	,7929
S Discourse	78	1,0	4,0	2,865	,6917
N válido (según lista)	78				

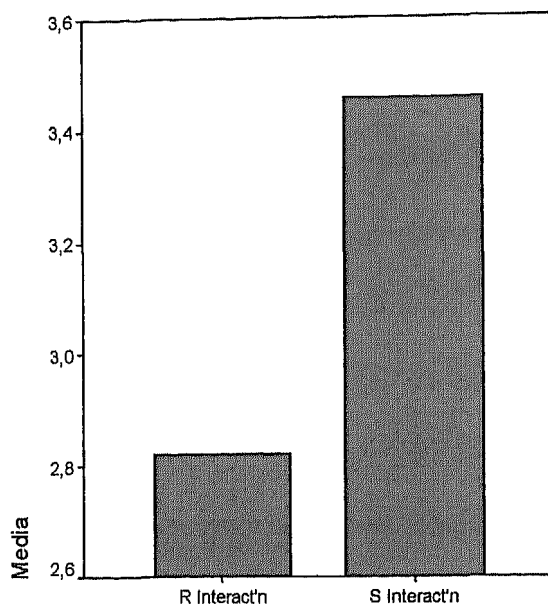


The statistical comparison (paired sample *t*-test) between the mean score of the rater and that of the student in the category of 'Discourse Structure' results in a *t* value of -1.899 with a probability of 0.61. This is greater than 0.05 and if we also take into account that the confidence interval (see Appendix 16) includes the value 0 we can affirm that there is no significant statistical difference between the mean scores of the rater and the student. This contrasts with the findings in the previous test where Discourse Structure was the only category where there was a significant difference between the scores of both rater and student and interviewer and student. Again, we can see here how the group test format seems to give students more confidence in their speaking ability, leading them to feel much more in control of this complex aspect of structuring speech and which they felt to be so lacking in the first test.

Interaction

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Interact'n	78	1,0	5,0	2,821	,9999
S Interact'n	78	1,0	5,0	3,455	,9194
N válido (según lista)	78				



The paired sample *t*-test gives a result of -5.248, corresponding to a probability of .000. Moreover, the confidence interval does not include 0, so here we can appreciate a highly significant statistical difference between the scores, which is in complete contrast to the first test where no significant differences were found between any of the scores. It is possible to infer from these results that the students feel much more comfortable interacting with one another in the group speaking test than they do with just the interviewer/candidate interaction in the one-to-one format, and they therefore give themselves a much higher score than the rater and also a much higher score than they gave themselves in the previous test.

b) Correlation Study

On comparing the mean scores for each scoring category we have seen that there are significant differences between the rater and student scores for two aspects of the test. For this reason we subsequently decided to verify whether these differences

were limited to the students awarding themselves a higher score or whether there was also a lack of correlation between the two sets of scores. In order to do this, we again used the Pearson Correlation Test.

With the results obtained from the Pearson Test (see Appendix 17) it can be seen that although there are some significant differences between the scores given by the rater and the student, there is also a very high level of correlation between the scores in the categories of 'Grammar and Vocabulary', 'Pronunciation' and 'Interaction', which can be considered significant at the value of 0.01 (very much higher than 0.05, the minimum for statistical significance to be considered). In the marks awarded for 'Discourse Structure', there is also a significant correlation, although at a lower level (0.05). We can therefore conclude that the differences noted between the rater and student scores are limited to the fact that the students award themselves higher marks in all areas, since the positive correlation is consistently at a statistically significant level. This mirrors the results of the correlation study for the 'Individual Oral Proficiency Interview' test and lends support to the hypothesis that students are, in fact, quite accurate in their self-assessment, at least on a level of establishing a pattern of relative strengths and weaknesses.

IV.2.2 Data from Questionnaire 3 (Student)

As we explained above, the last 17 variables (15 – 31) correspond to the 17 items on Questionnaire 3. In order to present the data obtained we will analyse the descriptive data from each item separately. It is also important to remember that these variables are quantitative ordinals since they correspond to items on a Likert scale of 1

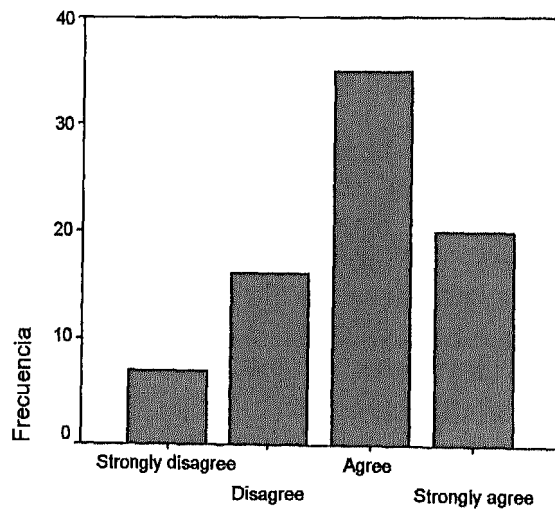
– 4.

Question 1: *I felt nervous throughout the test.*

Quest-III 1

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	7	9,0	9,0	9,0
	Disagree	16	20,5	20,5	29,5
	Agree	35	44,9	44,9	74,4
	Strongly agree	20	25,6	25,6	100,0
	Total	78	100,0	100,0	

Quest-III 1



Quest-III 1

Here, we can observe that the group test format may have an influence on reducing anxiety, since although 70.5% of students stated that they felt nervous throughout the test, this contrasts with the 82.3% who responded in the affirmative for the interview format. It is not possible to establish whether this difference is statistically significant because the two groups of students taking the tests were not exactly the same, but the students' self-assessment scores for the 'Group Speaking

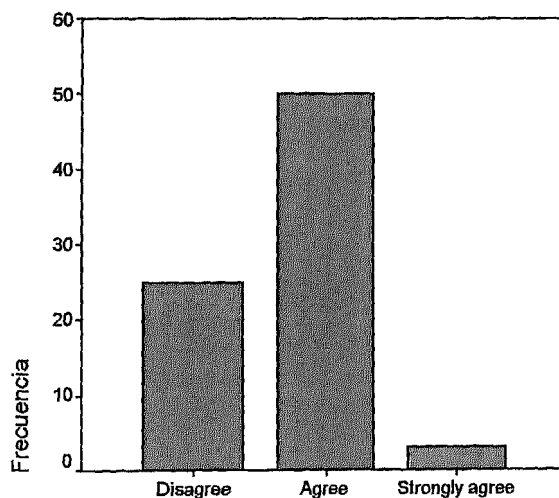
Test', as we have seen above, are consistently higher than for the individual interview test which may also indicate a greater degree of self-confidence in this format.

Question 2: *I think I did well in the test.*

Quest-III 2

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	25	32,1	32,1	32,1
	Agree	50	64,1	64,1	96,2
	Strongly agree	3	3,8	3,8	100,0
	Total	78	100,0	100,0	

Quest-III 2



Quest-III 2

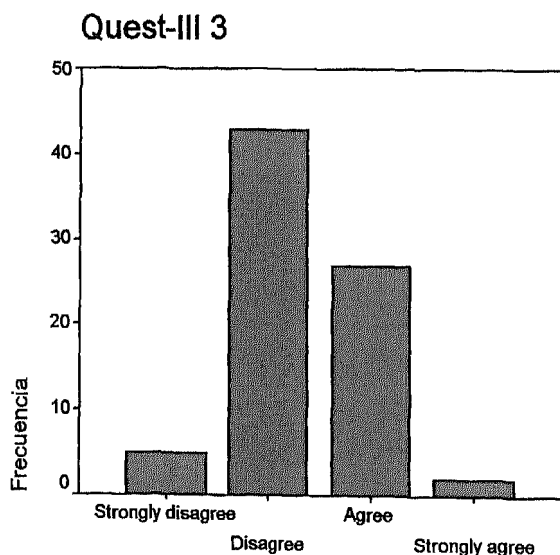
As we can see in the graph above, there was no strong disagreement with this statement and most of the students felt that they had done well in the test. The percentage is very similar to that found in the 'Individual Oral Proficiency Interview' (67.9% vs. 68.7%) which would seem to indicate that, while the group speaking test

was well received, the one-to-one test format did not have as negative an impact on the perceived performance of the test takers as we had expected.

Question 3: *I think I performed to the best of my ability in the test.*

Quest-III 3

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	5	6,4	6,5	6,5
	Disagree	43	55,1	55,8	62,3
	Agree	27	34,6	35,1	97,4
	Strongly agree	2	2,6	2,6	100,0
Total		77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		



Quest-III 3

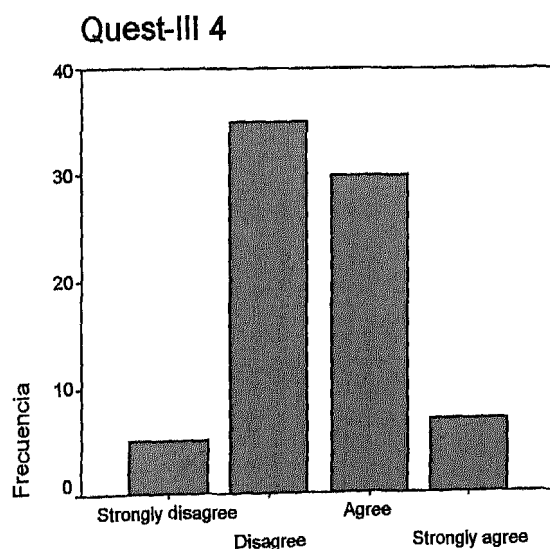
Here we can see that most students continued to feel that the test had not given them the opportunity to perform to the best of their ability, although the percentage was not as high as for the first test (82.3% vs. 62.3%). This may indicate that,

generally, students feel that exams do not allow them to show the full extent of their capabilities and we could therefore consider that the much lower percentage of students who felt they had not performed as well as they thought they were capable of here in comparison with the first test, reflects positively on the group test format in the provision it makes for samples of performance.

Question 4: *I think I spoke enough for the examiner to judge my ability.*

Quest-III 4

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	5	6,4	6,5	6,5
	Disagree	35	44,9	45,5	51,9
	Agree	30	38,5	39,0	90,9
	Strongly agree	7	9,0	9,1	100,0
	Total	77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		



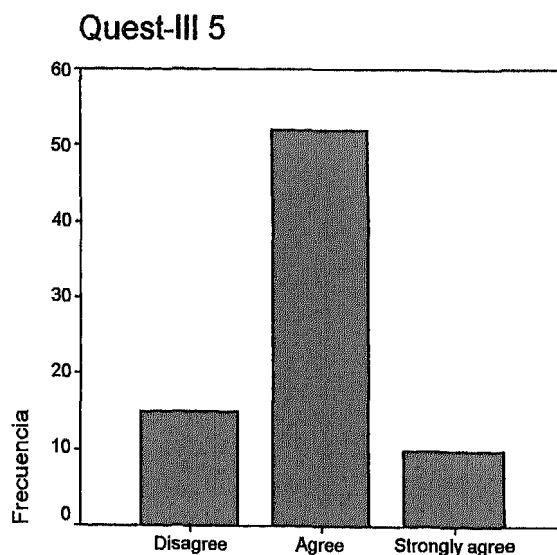
Quest-III 4

The students' impressions here are almost equally divided, with just under half agreeing that they had produced enough speech during the test for the examiner to make an appropriate judgement on their speaking ability. These results are almost the same as for the 'Individual Oral Proficiency Interview' test and we are therefore unable to draw any inferences about the influence of the test format on students' perception of the size of the speech sample they produced.

Question 5: *I felt comfortable with the procedure of the test.*

Quest-III 5

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	15	19,2	19,5	19,5
	Agree	52	66,7	67,5	87,0
	Strongly agree	10	12,8	13,0	100,0
	Total	77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		



Quest-III 5

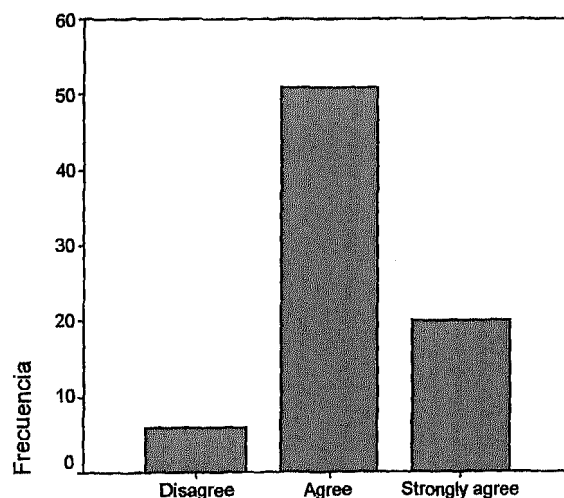
The results for this item are practically identical to the corresponding item in the first test (78%): it seems that the great majority of students (80.5%) were happy with the test format, with no candidate expressing strong disagreement. Although these items do not show any strong preference for either one of the test formats, they do show that the group speaking test was at least well-received among students who had previously been used to taking speaking tests in a one-to-one format.

Question 6: *I knew exactly what I had to do.*

Quest-III 6

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	6	7,7	7,8	7,8
	Agree	51	65,4	66,2	74,0
	Strongly agree	20	25,6	26,0	100,0
	Total	77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		

Quest-III 6



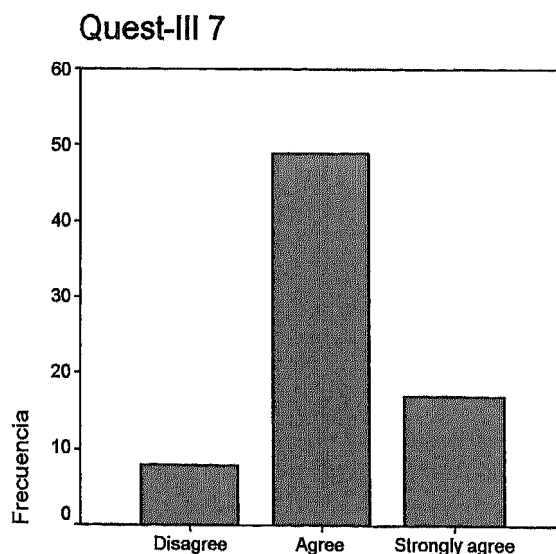
Quest-III 6

It can be seen here that the overwhelming majority of students (92.2%) understood what was expected of them in the test, with only six students disagreeing and none showing strong disagreement. This again supports the use of this test format and procedure since students found it easy to understand and follow. It is our belief that this will contribute to lowering anxiety which will, in turn, have a positive effect on the perception of performance, and perhaps on performances themselves.

Question 7: *The test was similar to the kind of task practised in class.*

Quest-III 7

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	8	10,3	10,8	10,8
	Agree	49	62,8	66,2	77,0
	Strongly agree	17	21,8	23,0	100,0
	Total	74	94,9	100,0	
Perdidos	Sistema	4	5,1		
Total		78	100,0		



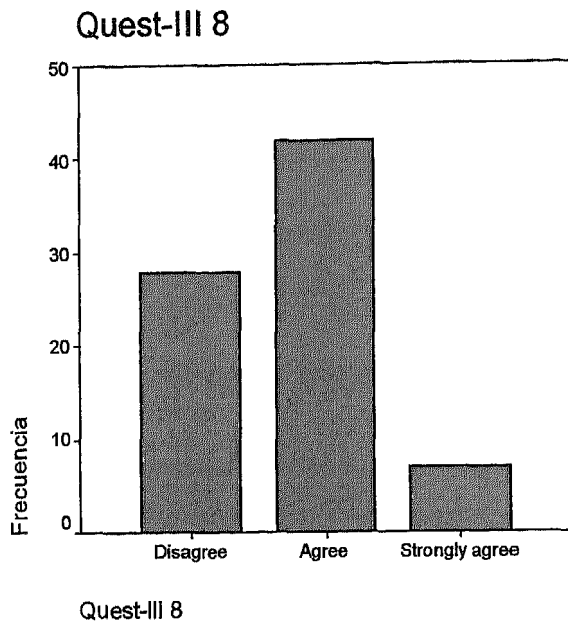
Quest-III 7

Here we can see that, again, the majority of students (89.2%) found the test task similar to the kind of task practised in class. This is in complete contrast to the response to the same item for the 'Individual Oral Proficiency Interview', where only 37.5% of the students thought the test tasks bore any resemblance to classroom activities. One of our original aims was to produce a test that was related to classroom activity and practice and here we can see that the 'Group Speaking Test' accomplishes this goal by providing speaking tasks that are similar to those carried out in the teaching/learning programme. The fact that students had practised this test format in the classroom prior to the test should also have had an influence on the fact that they perceived it as an integral part of the subject syllabus.

Question 8: *I could answer the questions without difficulty.*

Quest-III 8

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	28	35,9	36,4	36,4
	Agree	42	53,8	54,5	90,9
	Strongly agree	7	9,0	9,1	100,0
	Total	77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		

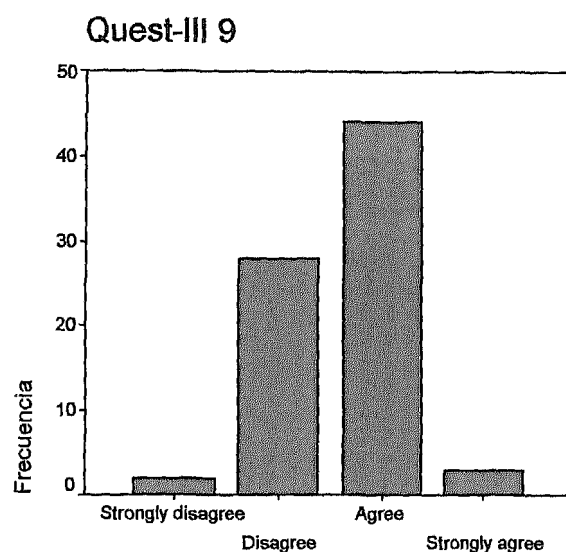


Most students (63.6%) also indicated that they did not find difficulty in answering the questions, and again there were no cases of strong disagreement. These figures are 10% higher than for the 'Individual Oral Proficiency Interview' (where 53% of the test takers stated they could answer the questions without difficulty) and therefore reflect positively on the group test format and procedure.

Question 9: *I had enough to say about the topic.*

Quest-III 9

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	2	2,6	2,6	2,6
	Disagree	28	35,9	36,4	39,0
	Agree	44	56,4	57,1	96,1
	Strongly agree	3	3,8	3,9	100,0
	Total	77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		



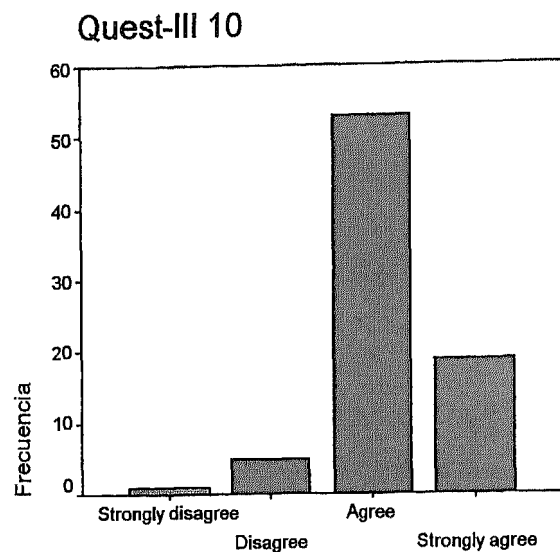
Quest-III 9

Sixty-one per cent of the students reported that they had enough to say about the topic of their test. This is a much larger number than for the 'Individual Oral Interview', where only 43% of students answered this question affirmatively. It may be possible to infer from this that interaction which is co-constructed between the group members is more easily produced and also that this test format leads to the use of supportive conversation strategies among candidates, allowing them to develop their ideas in a collaborative way.

Question 10: *I understand what my mark means.*

Quest-III 10

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	1	1,3	1,3	1,3
	Disagree	5	6,4	6,4	7,7
	Agree	53	67,9	67,9	75,6
	Strongly agree	19	24,4	24,4	100,0
	Total	78	100,0	100,0	



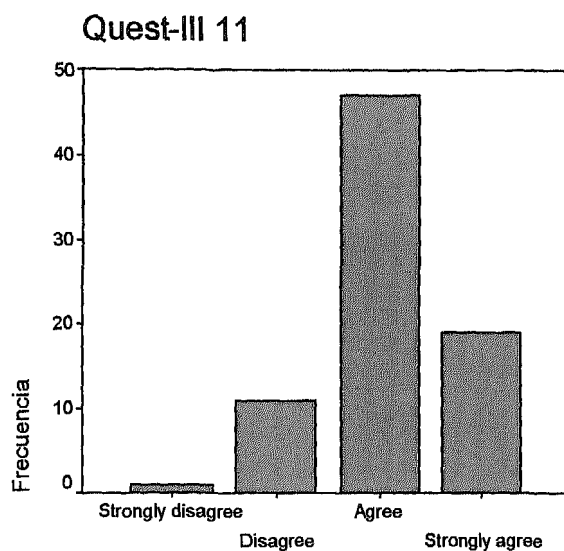
Quest-III 10

We can observe from the graph that the vast majority (92.3%) of students indicated that they understood the mark they had received for the 'Group Speaking Test'. These results do not differ substantially from those obtained for the 'Individual Oral Proficiency Interview' and continue to reflect positively on the use of the analytic scoring scale.

Question 11: *I know what I need to do in order to improve my speaking.*

Quest-III 11

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	1	1,3	1,3	1,3
	Disagree	11	14,1	14,1	15,4
	Agree	47	60,3	60,3	75,6
	Strongly agree	19	24,4	24,4	100,0
	Total	78	100,0	100,0	



Quest-III 11

For this item, almost 85% of students agreed or strongly agreed that the scoring procedure for this test was useful as a means of indicating what they needed to do in order to improve their speaking skills. This is comparable with the results obtained for the interpretation of the analytic score on the first test, where 94% of the candidates indicated that their mark helped them to understand how to improve. In contrast, the global mark out of 10 was seen as useful for this purpose by just 66% of students in the Individual ‘Oral Proficiency Interview’; it was not used in the ‘Group Speaking Test’.

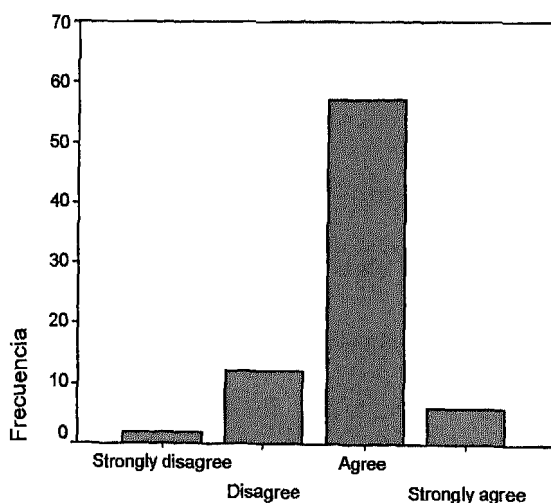
Again, we can observe that students believe the analytic scale provides them with more information on which to base further study and practice than the traditional 0 – 10 norm-referenced scale. These findings provide us with an argument in favour of continuing to work on the development of this type of scoring system in order to improve our teaching and assessment methods.

Question 12: *I think that my general self-assessment was a true reflection of my speaking ability in English.*

Quest-III 12

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Strongly disagree	2	2,6	2,6	2,6
	Disagree	12	15,4	15,6	18,2
	Agree	57	73,1	74,0	92,2
	Strongly agree	6	7,7	7,8	100,0
	Total	77	98,7	100,0	
Perdidos	Sistema	1	1,3		
Total		78	100,0		

Quest-III 12



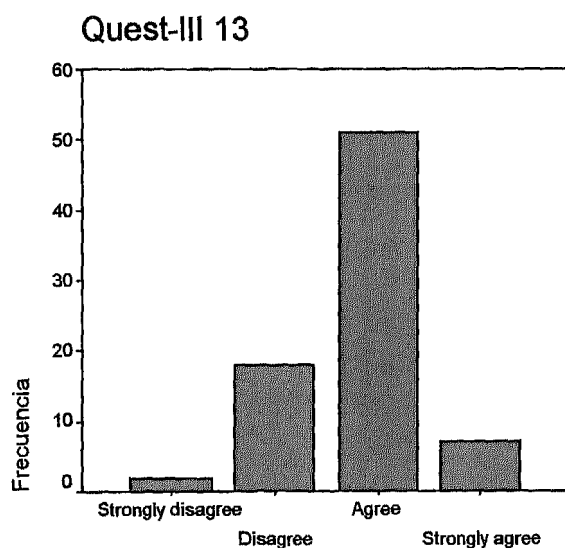
Quest-III 12

With regard to self-assessment procedures, once again the vast majority of students (81.8%) were of the opinion that they could give an accurate reflection of their own general speaking ability using the assessment criteria provided. The accuracy of this opinion is further corroborated by the high correlation values obtained for the rater and student scores on the test (see Appendix 17).

Question 13: *I think that my self-assessment in the group oral test was a true reflection of my speaking ability.*

Quest-III 13

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	2	2,6	2,6	2,6
Disagree	18	23,1	23,1	25,6
Agree	51	65,4	65,4	91,0
Strongly agree	7	9,0	9,0	100,0
Total	78	100,0	100,0	



Quest-III 13

Slightly fewer candidates thought that their self-assessment of their speaking ability in the context of the test was an accurate reflection of their speaking ability (74.4%). From this it is possible to interpret that, for the most part, students actually distinguish between giving themselves a score for test performance and evaluating their ability in speaking generally, and that they are objective enough to be able to do this. It seems that if we consider the similarity between the patterns of correlation of

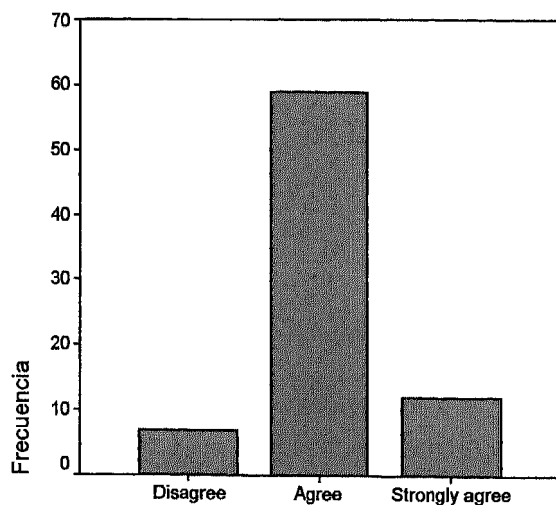
the scores awarded by the rater and the student, this is in fact the case, although the students under-score their performance in the 'Individual Oral Proficiency Interview' and over-score it in the 'Group Speaking Test', possibly for reasons which we have mentioned above.

Question 14: *I think self-assessment can play a useful role in learning generally.*

Quest-III 14

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	7	9,0	9,0	9,0
	Agree	59	75,6	75,6	84,6
	Strongly agree	12	15,4	15,4	100,0
Total		78	100,0	100,0	

Quest-III 14



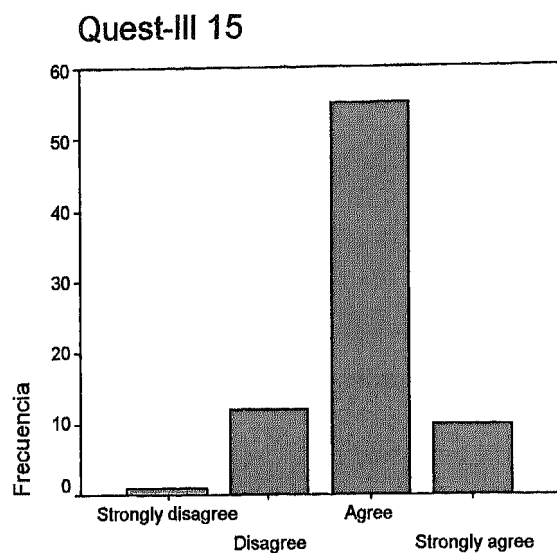
Quest-III 14

The above graph shows that ninety-one per cent of students thought that self-assessment can play a useful role in learning, with no respondents strongly disagreeing with this item. The positive attitude towards self-assessment observed here provides a sound basis for attempting to develop ways of incorporating it into our teaching/learning programme as a form of motivation for our students to engage in further study and encourage improvement. The 7 students who disagreed may have done so due a lack of familiarity with the procedure or to a deeply-rooted belief that assessment is the teacher's job.

Question 15: *I think my self-assessment should be taken into consideration in my overall grade for the subject Lengua B II.*

Quest-III 15

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Strongly disagree	1	1,3	1,3	1,3
Disagree	12	15,4	15,4	16,7
Agree	55	70,5	70,5	87,2
Strongly agree	10	12,8	12,8	100,0
Total	78	100,0	100,0	



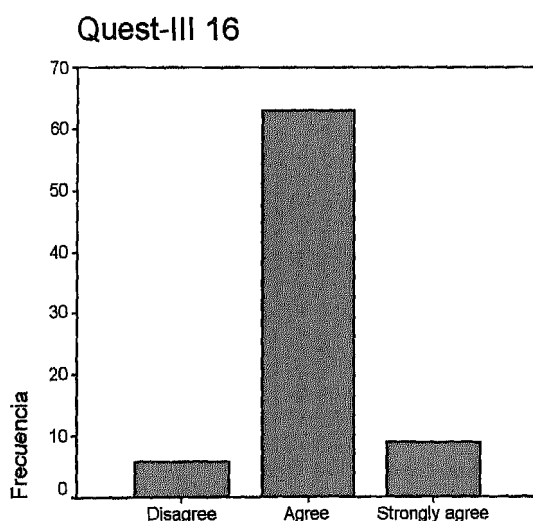
Quest-III 15

Again, we can observe that the overwhelming majority of learners (almost 84%) thought that their self-assessment scores should be taken into consideration as part of the overall grade they received for the subject *Lengua BII*. This would indicate that students believe themselves to be accurate in their perception of their own ability and performance, and this belief is supported by the data from our results in the study of the test scores themselves. It is likely that if students know that their self-assessment scores will form part of their final mark, they will be motivated to monitor their progress and to apply strategies that will help them to improve in the areas where they judge their ability to be weaker. Our challenge would then lie in designing a system of implementation of self-assessment procedures which provide this motivation, while at the same time safeguarding fairness and avoiding openness to abuse.

Question 16: *We should be given the opportunity to use self-assessment more frequently in this subject.*

Quest-III 16

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Disagree	6	7,7	7,7	7,7
	Agree	63	80,8	80,8	88,5
	Strongly agree	9	11,5	11,5	100,0
Total		78	100,0	100,0	



Quest-III 16

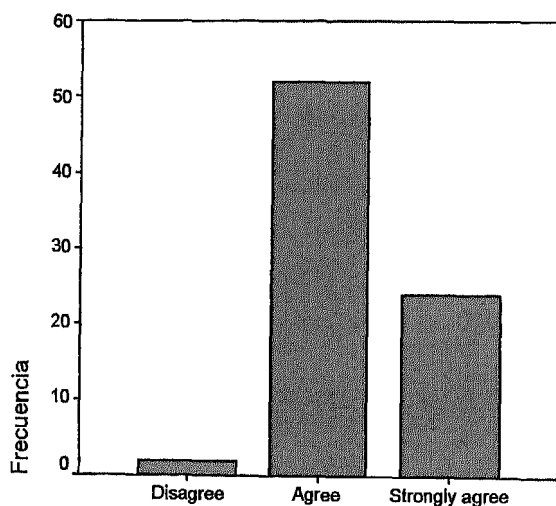
Practically all the students in this sample expressed a positive opinion about carrying out self-assessment, agreeing that they should be given the opportunity to use it more frequently, and again supporting our initial hypothesis that it can have a positive impact on learning. Those that disagree, the very small minority, might do so for reasons of unfamiliarity and traditional views of teaching roles.

Question 17: *We should be trained in how to assess our language skills in this subject.*

Quest-III 17

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Disagree	2	2,6	2,6	2,6
Agree	52	66,7	66,7	69,2
Strongly agree	24	30,8	30,8	100,0
Total	78	100,0	100,0	

Quest-III 17



Quest-III 17

Finally, only 2 out of 78 respondents were in disagreement with the idea that it would be useful to train students to assess their own English language skills. This training might involve, as a first step, providing students with similar criteria to the ones we use as teacher/examiners to assess them and practice in using them. At the same time as helping us to clarify our criteria for assessment, this would also assist

students in focusing on their own strengths and weaknesses, thus providing a more personalised, learner-centred approach to learning which does not necessarily involve the teacher in an excessive workload when dealing with large numbers of students.

IV.2.3 Data from Questionnaire 4 (Interviewer)

The same four interviewers who took part in the first testing session ('Individual Oral Proficiency Interview') also conducted the 'Group Speaking Tests'. The results of the questionnaires they filled in at the end of the second testing session are presented in the same way as for the first test, in the form of a table representing the original questionnaire, with the numbers indicating how many respondents chose each of the answers.

QUESTIONNAIRE 4 (Group Oral Test)

	Strongly disagree	Disagree	Agree	Strongly agree
1. I was able to manage the interview and give each student a score at the end of the test using the rating scale provided			3	1
2. I was more focused on managing the interview than on the rating criteria		3		1
3. I felt comfortable with the test procedure			2	2
4. The students produced a large enough speech sample for assessment			2	2
5. It was difficult to manage the test with three students participating	2	1	1	
6. I felt comfortable in the dual role of interviewer and global rater			1	3
7. I knew what features to focus on while assessing the candidates			3	1
8. It was easy to assess how well the candidates were interacting			3	1
9. It was useful to have a rating scale to refer to when giving the global score				4

10. It was easier to assess students who expressed an opinion similar to mine on the topic	4			
11. It was easier to use a scale from 0-5 than one from 1-10		2	2	
12. It was easier to assign meaning to a scale of 0-5 than to one of 1-10		1	3	
13. I think that I awarded the students a fair score			4	
Reason:				
14. I think that students can give a true reflection of their general speaking ability using the criteria provided		1	3	
15. I think that students can give a true reflection of their performance in the group oral test using the criteria provided		1	3	
16. Self-assessment is a useful tool for helping students to know how improve their speaking ability in English				3*
17. Self-assessment can play a useful role in learning generally			1	3
18. Self-assessment should be taken into consideration in the students' overall mark for English Language subjects at the ULPGC		1		1*

*In Questionnaire 4, one interviewer did not answer item 16 and only two answered item 18.

As in the previous Section, in order to contrast the perspectives of the rater/interviewers with those of the students and to obtain an overall view of the effects of test format, scoring procedures, and the role of self-assessment in teaching, learning and assessment, the results obtained from this questionnaire will be discussed with reference to the research questions they address in Section 3 below.

IV.3 DISCUSSION

In the following section, we will discuss the results we have obtained from our analyses of the data presented above in order to discover whether they contribute to our understanding of what candidates and interviewers believe is taking place during the different speaking tests, and whether or not this is reflected in the scores obtained. We will also attempt to answer the research questions set out in Chapter 3 *Research Design* in order to ascertain whether the design of our research project was adequate for our purposes and, in the cases where it appears to be so, to draw some preliminary conclusions about the procedure and scoring methods of speaking tests in our university teaching and learning context. First, we will analyse the data collected with reference to the two test formats used (Section IV.3.1), followed by a consideration of the scoring procedures (Section IV.3.2), together with their effectiveness and usefulness, and finally we will look at the possible pedagogical implications of self-assessment and the role it may play in testing (Sections IV.3.3 and IV.3.4).

IV.3.1 Test Format

We shall consider test format from both the student and interviewer/rater perspectives.

a) Student Perspective

The first research question we wished to address with regard to test format was:

1. *Does taking a speaking test in a group reduce the anxiety inherent to speaking tests in general and, if so, is anxiety lower than in a one-to-one interview situation?*

On designing the 'Group Speaking Test', our initial hypothesis was that this format would reduce the candidates' level of anxiety due to aspects of the test such as peer support and familiarity with the test procedure, resulting from its similarity to classroom tasks, and hence it would lead them to feel more confident both with the test itself and in expressing their opinions. In relation to familiarity with the test procedure, 88.2% of students agreed that the tasks in the 'Group Speaking Test' were similar to those practised in class, while only 37.5% found a similarity between the 'Individual Oral Proficiency Interview' and classroom procedure. Similarly, 92.2% of those questioned responded affirmatively to the question number 6 "I knew exactly what I had to do" in the group test. It is questionable, however, whether these positive responses translate into reduced anxiety in any significant way, since 70.5% of candidates continued to report that they felt nervous throughout the test, compared with 82.3% in the one-to-one interview. However, since the one-to-one interview took place as a mock speaking exam, whilst the group oral test was an obligatory part of the final assessment for *Lengua BII*, we would have expected greater, rather than lower, anxiety on the part of the students. The fact that fewer students reported extreme nervousness in the group test format may be considered positive evidence in favour of implementing this kind of test.

In both the one-to-one interview and the group test, almost the same percentage (78% and 80.5%) of the students reported that they were comfortable with the procedure of the test. This is surprising, since we would have anticipated a much higher degree of discomfort with the one-to-one interview due to the imbalanced situation with regard to power structure (even more so in the tests carried out for our research, since both an interviewer and a rater were present in the first test creating a

2-1 examiner/candidate ratio) and hence the type of discourse expected. We could hypothesise several reasons for this; it may be that in general, the social or power structure at university, or even throughout the education system, is such that our students are happy to accept any kind of imposed test as reasonable, whether or not it bears relation to the teaching/learning syllabus.

Alternatively, students may see the interview as a valid means of testing spoken language simply from a traditional point of view, as do many teachers and examiners; it is simply the way it has always been done and is therefore accepted without question. They may even perceive it as a useful form of gaining experience for their later lives in the world of work, where they will almost certainly have to take part in face-to-face interviews in unbalanced power situations. It is also possible that more confident candidates in fact enjoy the relatively infrequent opportunity of a one-to-one conversation in the second language and do not feel intimidated by the situation, and also that our interviewers were extremely skilled in putting candidates at their ease, which they would not have had very much opportunity to do in the group test since control of the interaction is relinquished to the students themselves.

Our second question was concerned with the relationship between test procedure and performance:

2. *Does familiarity with the task and/or test type have a bearing on performance?*

In contrasting the scores given in both types of test across the four different aspects that we have identified as part of the speaking construct that can be objectively evaluated ('Grammar and Vocabulary', 'Pronunciation', 'Discourse Structure' and 'Interaction'), we can see that in the first test, the 'Individual Oral Proficiency Interview', the students gave themselves the lowest mark of the three assessments in

every category, while, in contrast, in the 'Group Speaking Test' they consistently awarded themselves the highest score in all the categories. Therefore, we may tentatively take this evidence to mean that familiarity with both the task and test format leads to an improved perception of test performance, and also possibly that this, together with the group speaking test format, reduces test-taking anxiety.

It is difficult to know what steps we may take as teachers/examiners in order to reduce anxiety further during tests, and perhaps it is even questionable as to whether this is desirable. It may be that the degree of nervousness experienced by test-takers in the majority of cases is necessary in order to enhance performance. The percentages of students who thought they did well in the test are almost the same in both cases: 67.9% for the group test and 70% for the individual interview, showing that, in fact, their impression of how they performed in the test seems not to be linked to the degree of anxiety experienced.¹⁰

However, with regard to our third question concerning test format,

3. *Do students feel that the test format allows them to demonstrate their speaking ability?*

the group test format did seem to lead more students to believe that they had performed to the best of their ability during the exam. Twenty per cent more candidates answered this question affirmatively than for the oral interview (37.7% compared to 17.7%). This may be related to the ease with which students were able to express an opinion on the topics they were asked to speak about (in the group test,

¹⁰ In the 'Individual Oral Proficiency Interview' 72% passed (rater score), while 70% reported they felt they had done well in the test and 82.3% felt nervous throughout. In the 'Group Speaking Test' 67% passed (rater score), while 67.9% thought they had done well and 70.5% felt nervous throughout the test.

61% said that they had enough to say about the topic, compared to only 43.1% in the interview) and to the perceived difficulty of the questions (in the group test 63.6% reported being able to answer the questions without difficulty, compared to 52.9% in the interview). In the latter case, the difference would seem to be too small to indicate that students find it substantially easier to answer the questions if they are written (as was the case in the 'Group Speaking Test') rather than just spoken by the interviewer, although for the group test procedure this would still seem to be the most practical way of providing the candidates with a spring-board for discussion and maintaining minimal involvement of the interlocutor in the interaction.

b) Interviewer/Rater Perspective

From the perspective of the interviewer/rater, our first research question regarding test format was:

1. *Do examiners feel that they can manage test materials and interaction, as well as give accurate and objective scores for candidates' speaking performance at the same time?*

In the 'Individual Oral Proficiency Interview', the interviewer carried out the dual role of managing the interview and used the analytic rating scale to score students on the different categories of the speaking construct, as well as giving a traditional 0 – 10 mark. For the 'Group Speaking Test', the interlocutor was only required to give a global score on the 5-point descriptive scale specially designed for that purpose. In both tests, the rater used the 0 – 5 analytic rating scale designed for the *Lengua BII* speaking test.

The data collected in Questionnaire 2 for the one-to-one interview format with regard to interview management and simultaneous scoring on the traditional, intuitive 0 – 10 scale, show that all the interviewers were confident that they were in control of the situation. The same was true for the interlocutor managing the group speaking test and giving all three candidates a score using the 5-point descriptive analytic rating scale (Questionnaire 4), with three interviewers agreeing and one strongly agreeing that they were able to manage interaction and rating in both cases. One examiner did, however, comment that “the role of interviewer was easier than that of the rater, since you didn’t need to consider the different aspects to be assessed, but the general performance.”

In the case of the one-to-one interview, three interviewers felt that they could manage the interaction of the interview and then give a fair score from the analytic scale at the end of the test, while one felt unsure of both things, alleging that s/he had devoted more attention to conducting the interview than to focusing on details of the students’ speaking competence. The same interviewer, however, did feel that s/he had managed the interview and given the global 0 – 10 mark competently, although later in the same questionnaire doubted that the mark s/he had awarded was a fair one. It is precisely these contradictions that should lead us to question whether what we think we are doing is what we are *really* doing in carrying out tests and assessing them in real time.

In the individual interviews, just two of the four interviewers felt happy with the test procedure, and all agreed that they were more focused on managing the interaction than on the rating criteria, yet three still thought that they had given an accurate and fair mark according to both the intuitive and to the analytic scale. These

replies also seem to be contradictory, since if our attention is focused on one thing (naturally, in this case, on managing the immediate task at hand, i.e. the interaction) by definition we cannot be focused on something else at the same time. In order for two actions to take place simultaneously in this way, one has to be automatised, that is, our actions take place without us having to think consciously about them like changing gear and steering when driving a car. We do not have to focus on these actions; we carry them out automatically, freeing up our attention for what is happening on the road in front of us. In the case of interviewing (where necessarily the interlocutor needs to attend to the content of what the candidate is saying, as well as to directing the interaction), it cannot be possible to notice and, at the same time, to competently measure, distinct features of the speaking construct demonstrated in the performance sample.

Our question referring to the difficulties of test management with a group of test takers rather than with an individual,

2. Is managing a test with three students easier than managing and simultaneously scoring an individual test?

can only be partially answered, due to a failing in the questionnaire design which did not make a specific reference to comparison between the two tests. Three of the four interviewers disagreed or strongly disagreed with the statement “It was difficult to manage the test with three students participating” while one agreed. However, one examiner commented at the end of Questionnaire 4 that “It was much easier to handle a one-to-three interview than a one-to-one oral exam; [there was] more interaction, less tension for the students and the interviewer (but more for the rater!), and [it was] closer to a real situation of communication”. These findings indicate that, generally, the

interviewers were able to manage the group test without difficulty, but we cannot make any global statement about their preferences for one test format or the other since the question did not focus accurately on this point.

However, two of the interviewers did indicate that they were uncomfortable with the dual role of interviewer and rater, while the remaining two reported finding no difficulty in carrying out both tasks simultaneously. In the group test, all four interviewers felt comfortable in the dual role of interlocutor and global rater, but in order to draw any definitive conclusions from this difference, we would need to run trials involving more examiners. However, we could consider these results to indicate that it is a less complex task to manage the interaction and to give a global mark at the end of the test using a descriptive rating scale than to try to judge candidate ability using an analytic rating scale and manage the test at the same time.

Only one of the interviewers found difficulty in assessing the interactive ability of the candidate in the one-to-one situation, perhaps more aware that the candidates were acting only in a respondent role and that they had no opportunity to initiate or change a topic or ask a question, and probably did not feel confident enough to disagree. We could postulate that this tendency to feel that it is possible to judge interactive ability in an interview situation by the readiness to respond to directed dialogue corresponds to the lack of a readily available definition of what the speaking construct involves. A test where a candidate responds willingly and with relative ease is likely to be perceived as demonstrating a 'good' interactive performance without consideration of all the features that make up communication in authentic situations; the interview is a relatively uncommon communicative situation and does not require a wide range of communication strategies. Another interviewer, although agreeing that

it was easy to assess how well the candidate was interacting, commented that “Interaction can be rated, but it is not natural because only the Interviewer asks questions. Interaction is the ability to maintain conversation.” Here we can see an attempt at defining just what speaking involves, and this is a necessary step to take if we are to make confident and reliable judgements about performance.

Our final question regarding features of test format from the perspective of the examiner was:

3. *Does the test format influence the size of the speech sample produced by candidates, either facilitating or hindering assessment?*

All four raters considered that the speech sample produced by the candidate was sufficient for them to assess their speaking ability on both tests, and two interviewers marked the option ‘Strongly agree’ for this question (Item 6) in the questionnaire concerning the group test, indicating that candidates possibly spoke more in that test format. This contrasts markedly with the students’ impressions, with just 45% of candidates considering that they spoke enough during the oral interview and 48% with the same opinion in the group test. Although this difference between students’ perceptions of how much they spoke in the two tests is not significant, it does appear to indicate that students are not disadvantaged with regard to the amount of time available for speech by taking a speaking test with their peers, since the one-to-one format does not necessarily mean that they speak more.

IV.3.2 Scoring

a) Interviewer/Rater Perspective

Our first question concerning scoring procedures from the point of view of the interviewer/rater was:

1. *Do examiners feel more confident in awarding scores when using a descriptive scoring scale than when using a traditional 0 – 10 scale?*

Here, our hope was that by reducing the numerical range of scores (0 – 5) and providing definitions for each one, examiners would be able to identify features of performance that would allow them to award scores more objectively, hence giving them greater confidence in the testing procedure. However, it was also possible that interviewers involved in simultaneous test management and rating would continue to think that it was easier to give a mark on the 0 – 10 scale, which has been internalised over years of use and which, generally, they feel themselves to be in control of, than to use the new analytic scale. The results show a balance for preference between the two types of scoring procedure, with just two of the interviewer/raters preferring the traditional scale, while the other two stated that they found the descriptive 5-point scale easier to use.

Our next question addressed the important issue of what our speaking test scores actually mean:

2. *How do interviewers/raters interpret the meaning of these two types of rating scale?*

With regard to the rating procedure for the one-to-one interview, the interviewers had varying opinions on their understanding of the two rating scale types (the global intuitive scale, and the detailed analytic scale). Only one of the

interviewer/raters said that s/he was not sure exactly what was being scored when using the traditional 0 – 10 scale, while the remaining three were all confident that they knew what they were assessing in awarding the global mark. Since this scale provided no written criteria to which examiners could refer, we had initially suggested that the raters' perception of the 0 – 10 scale was likely to be norm-referenced (performances are compared to one another and scored accordingly) rather than criterion-referenced (scores are given according to previously agreed criteria of level and expectations, and are not dependent on contrasting student performances). In fact, a failing of the first questionnaire was that we did not ask this question explicitly and none of the examiners reflected on it in their general comments about the interview experience. However, in the comments on the group oral test, one of the examiners observed that “The reason why I found it a bit difficult to assess their oral skills was that you had to pay attention to three people at the same time and, sometimes, instead of giving an objective mark, I compared their performance.” Since a major component of all spheres of our daily life involves decision-making based on comparison, it may be that unconsciously we continually compare students' performance in our assessment of all their language skills, and that in order to make a criterion-referenced judgement about these skills we need to heighten our awareness of our procedures and approaches in assessment.

All four interviewers/raters agreed or strongly agreed that they understood the features of speaking that were being assessed when using the analytic scale. However, it is interesting to notice the way in which they interpreted and adapted the analytic scales according to their own internalised understanding of them. All the raters used a modified scale to assess the candidates, rather than adhering to the one that had been

provided. Two of the raters commented on this tendency: “Sometimes I felt the need to use marks such as 1.5, 1.75...”; “Sometimes I gave half-points, maybe because I am used to the scale from 0 – 10”. In fact, we had removed .5s from the initial scale design in order to make it easier to use (they were originally included, but with no descriptor). Our examiners not only re-incorporated them of their own volition, but also extended the scale to include very fine tuning by means of adding values of .25 and .75.

Just what features of performance they were focusing on in order to necessitate these extremely precise scores is unclear. It is possible that rather than distinguishing actual criterion-referenced features of performance, they were attempting to make a distinction between student performances in a norm-referenced manner in the way described above. The traditional 0 – 10 scoring system tends to spread students along an achievement scale with fewer at the extremes (especially at the higher end of the scale) and a greater number in the middle. We seem to function with the idea that achieving high marks is unusual and that the student who does so must be truly outstanding; in effect, some way beyond the level we actually require or expect at a given moment and a given stage of the learning process. These are two radically different approaches to interpreting assessment and we believe that they require discussion, debate and consideration in their application. We can see here that, even when descriptors of levels and features of performance are provided, the over-riding tendency is for examiners to internalise and interpret these in their own way, subsequently leading to their adaptation according to previously assimilated models of assessment.

Our third question in the area of scoring refers to the way in which a descriptive scale may guide rating procedure by aiding raters to focus on a range of discrete features of the construct:

3. *Do interviewers/raters focus on a wider range of features of speaking when using a descriptive scoring scale?*

In both types of test, the results show that the rater gave a higher score than the interviewer in all the aspects of the speaking skill which were assessed. Although not statistically significant, these results are surprising since we would have expected the person who has constant simultaneous access to the rating scale descriptors, and thus a necessarily more objective view of the speech samples of the candidates, to be a stricter marker. It may be the case that the interviewer over-compensates for the fact that s/he scores the test retrospectively and somehow fixes on errors (especially in form) as salient in candidate speech and consequently awards marks with a 'negative intent' rather than reflecting a balanced overview. Three of the interviewers said that grammatical accuracy was not the most important part of their assessment when giving either the global mark or the detailed score, while one reported that s/he had focused on grammatical accuracy in order to give the global mark, but not for analytic scoring. This may indicate that it is actually very difficult to focus on the structural features of what a candidate is saying at the same time as trying to set them at ease, guide them through the interview, and elicit as much language production as possible. It may be that the rater who only has to focus on students' performance can notice discrete features of language, together with both the strengths and weaknesses of candidates, through referring constantly to the descriptive scales throughout the test, and therefore

they can compensate or balance out the positive and negative aspects of test performance.

Although our results do not directly indicate that the rater is focusing on a wider range of features of speaking when using the descriptive scale, they may go some way towards explaining why raters gave higher scores than interviewers. Further and more specific research would be necessary in this area to be able to draw more definitive conclusions.

Common to both test formats was the conviction that it was just as easy to mark a student who expressed an opinion contrary to that of the interviewer as it was to score a student whose views coincided with theirs. However, it may be of interest to note that three interviewers marked 'disagree' to the relevant items in the first questionnaire (the 'Individual Oral Proficiency Interview'; Items 14 and 15) and all four marked 'strongly disagree' for the 'Group Speaking Test' (Item 10). This may indicate that, in fact, although interviewers believe themselves to be objective at all times, there is a difference between being involved in the interaction and being detached from it in the role of rater; two of the interviewers remarked on this fact in their general comments on the test procedure: "It was much easier to assess the students as the Rater"; "Evidentemente, es más fácil evaluar al alumno ejerciendo de *rater* que de *interviewer*" [Obviously, it is much easier to assess the student in the role of 'Rater' than in that of 'Interviewer']. The role of a rater who is not involved in the interaction seems to give more confidence in the scoring procedure and a greater conviction that the score finally given is objective.

b) Student Perspective

With regard to the rating scales themselves and to how students understand and respond to the marks they receive, the first question we addressed was:

1. *Is an analytic score, which relates to a set of descriptors, more meaningful than a mark received on the traditional 0 – 10 scoring system?*

In the oral interview, 90% of students responded affirmatively to the questions “I understand what my global mark means” and “I understand what my analytic mark means”. However, in the one-to-one interview, 55% of students stated that the global mark they received was easier to understand than the analytic mark. Given the effort we had made to adapt and clarify the details of the descriptive rating scale, this is somewhat disappointing and certainly not what we expected. In part, it may be due to the fact that students, like teaching staff, are accustomed to the 0 – 10 scale and that, in essence, from a student perspective, it *is* clear; 4 means “I didn’t pass”, 4.5 means “Why didn’t the teacher pass me?”, 5 means “I passed” and beyond that there is probably a kind of self-placement within the peer group, a user’s norm-referenced judgement on their own ability.

Our second question attempted to find out if there was a relation between test scores and the language learning process from the students’ perspective:

2. *Do analytic scores indicate areas of strength and weakness to students, and hence have a pedagogical value?*

Here, we found more encouraging results in the replies to the questions “The global mark helped me to understand what steps I need to take in order to improve my speaking” (67%; presumably ‘I need to get better’), compared to a resounding 94%

who considered that the analytic mark helped them to understand the steps they needed to take in order to improve their speaking (Items 14 and 15 in Questionnaire 1).

Our results for the group oral test, where only analytic scales were used, were also favourable: 92% of the students said that they understood their mark, and 85% that they understood what they needed to do in order to improve. The reason for a lower percentage answering the same question in the affirmative for the group oral test as opposed to the one-to-one interview is unclear, since the same scales were used for both tests. It may be that, since the students consistently and significantly rated themselves higher than the rater in the 'Interaction' category, and also felt that they had interacted much better in this test than in the one-to-one interview, they did not understand how they could improve in this aspect of their speaking.

IV.3.3 Self-Assessment

a) Student Perspective

Since the analytic rating scale attempts to define some of the features that speaking is composed of, and it is implemented as the measurement scale for assessing student performance and subsequently generalising to ability, we decided that it would be useful for the students to see and use these scales as way of assessing themselves and perhaps of focusing on areas for improvement. We also believe that students can focus better on task requirements and reduce their anxiety if they know the criteria that are being used to assess them. This approach (at least in our context) is relatively unusual; students are rarely asked to evaluate their own abilities and performance, and would almost never have this taken into account as part of their overall assessment. Generally, assessment is something that comes from outside and which is thus felt to

be an objective report of their abilities. However, as we have already seen, this is not necessarily the case and in any case, learning to be objective about ourselves and our strengths and weaknesses, and using this as an aid to focus on areas that require attention and subsequently attempting to improve them, is a positive and necessary skill for life. As such, we feel that the use of self-assessment is justified in our educational programme, both pedagogically and socially. Data concerning student and teacher opinions on self-assessment was collected in Items 12-17 in Questionnaire 3 (Student) and Items 14-17 in Questionnaire 4 (Interviewer).

Our first question concerned the pedagogical role of self-assessment:

1. How useful is self-assessment in learning and improving?

The responses demonstrated an extremely positive attitude towards its use; the vast majority of students (91%) thought that self-assessment could play a useful role in learning generally and an overwhelming 97.5% felt that they should be trained in the use of self-assessment in their language skills in order to enhance learning.

With regard to our second question,

2. Should self-assessment be taken into account as part of the final mark for the subject Lengua BII?

eighty three per cent of the students who filled in the questionnaire considered that these scores should be taken into account in their overall grade for the subject *Lengua BII* and 92% thought that self-assessment should be incorporated into the programme of study throughout the syllabus.

The third question we tried to answer from the students' point of view was

3. Can self-assessment give an accurate reflection of speaking ability?

Our data for the results here comes from the responses to items 12 and 13 in Questionnaire 3. Eighty two percent of students thought that their self-awarded score for speaking outside the test situation was a true reflection of their ability, contrasted with 74.5% who felt that their self-assessment for the speaking test, while accurate, did not reflect their ability to speak in English. This is interesting, since it also indicates that some students think that their test performance does not reflect their underlying ability, perhaps due to the effects of anxiety on performance or the tendency to perform less effectively in limited or restricted time contexts.

These results provide abundant evidence of a strong desire on the part of students to be actively involved in the processes that affect and evaluate their progress in learning and their final results for the subject. They also suggest that self-assessment and motivation are closely linked and that it would be possible to increase our learners' intrinsic motivation by introducing methods of self-assessment into our study programme.

b) Teacher/Examiner Perspective

From the point of view of teachers/examiners, the opinion of the incorporation of self-assessment into our study programmes appears to be in total opposition to that of the students. In answer to our first question,

- 1. Should self-assessment be incorporated into our teaching programmes and testing procedures?*

only one of the raters (also teachers) felt that self-assessment should form part of a student's overall assessment; one disagreed, one expressed no opinion, and the other gave an opinion based on certain conditions which are reproduced below:

La auto-evaluación precisa de un entrenamiento de años de práctica por parte del alumno para que pueda tener un valor real en lo que se refiere a la medición de su progreso. El alumno encuentra dificultad para discernir y no tener en cuenta otros factores personales como el interés, el esfuerzo, el trabajo desplegado, la afectividad, etc. Por ello, en las preguntas 16 y 18 no pongo respuesta. Si el alumno estuviera convenientemente entrenado, estaría de acuerdo en los dos casos.

[Self-assessment requires years of training on the part of the student in order to be able to measure progress accurately. The student finds difficulty in distinguishing and leaving aside personal considerations such as interest, effort, the amount of work carried out, affective issues, etc. For these reasons I have not responded to questions 16 and 18. If students were given adequate prior training, I would agree in both cases.]

In effect, it is difficult to consider this to be a positive answer; “years of training” hardly seems a realistic proposition in order to introduce an innovation in our study programme, although we do agree that some training in self-assessment procedures is necessary, and the teacher’s doubts about the students’ capacity for objectivity are a salient feature of the comments. To some extent, it corroborates our original idea that, as teachers or ‘outside observers’, we believe that we are capable of objectivity, reliability and consistency in our measurement of language ability, despite our lack of reference to a construct definition, whilst our students, who are engaged in the learning process, are seen as unable to evaluate their progress objectively. We would argue that, as humans, we are *all* subject to the influence of “personal

considerations” in the judgements we make, and that often experience can accentuate, rather than reduce, this.

In looking at our second question from the teacher/examiner perspective,

2. *How accurate can students be in their self-assessment?*

we find that the same rater also disagreed that students’ self-assessment according to the criteria provided could give a true reflection of either their general speaking ability or of their test performance. This was in contrast to the other three raters who, rather surprisingly, all agreed that these things were possible. In view of this, it is difficult to understand why they would not have agreed that the self-assessment scores should be taken into account for students’ final grades in the subject.

Our final question in this section again addressed the pedagogical role of self-assessment:

3. *Can self-assessment be useful in helping students to improve their language skills?*

Coinciding with student opinion, three out of the four raters strongly agreed that self-assessment is a useful tool for helping students to improve their speaking ability in English, and all four agreed that self-assessment can play a useful role in learning generally. Again, these results are contradictory and confusing; if teachers/examiners do not believe that learners can be objective and accurate in assessing their performance or competence, it is difficult to understand why they think that improvement through self-assessment is possible.

Despite these conflicting views and uncertainties as expressed by the interviewers/raters, there would seem to be some basis for attempting to include it in our teaching and learning programme in the future in accordance with the positive

reception from the students, and certainly for more serious investigation of its effects and consequences.

IV.3.4 Empirical Evidence Regarding Self-assessment

In an attempt to answer our final research question:

1. *Is there any objective evidence to support an argument for introducing self-assessment into our study programme for the subject Lengua BII?*

we will compare only the scores the scores awarded by the rater with those of the student in the different categories that make up the analytic rating scale, since in the group test, the interviewer only gave a global score on a 0 – 5 scale.

In the category of the scale that corresponds to ‘Grammar and Vocabulary’, in the individual oral proficiency interview we find that there is a significant difference between the scores of the rater and the student which is not repeated in the group oral test. With reference to ‘Discourse Structure’, we find that in the interview there is also a statistically significant difference between the rater and student scores which again is not repeated in the group test format. In the one-to-one interview, in both categories, the students’ impression is that their score is much lower than the one the rater assigns them. This contrasts with the group test format where the reverse happens (the student scores are higher than those of the rater), although in this case the differences between the rater and student scores are not statistically significant.

In ‘Pronunciation’ there is a highly significant statistical difference between the scores in both tests, but with the higher score being given alternately by the rater and the student in the two types of test; in the interview, the students give themselves a much lower score in this aspect of speaking, while in the group test their self-

assessment scores are much higher than those of the rater. This may be because, in the group test, they tend to compare themselves with their peers and feel that their own pronunciation compares favourably with that of the others in the group, whereas in the interview they feel inferior to the interviewer in this aspect of their speaking skills for socio-cultural reasons. If this is the case, the cause can only be that of the power structure inherent to the interview test type, since none of the interviewers was a native speaker of English, although they were much more proficient than the candidates they were assessing.

In the scores for 'Interaction' we find that, while there are no significant differences in the results for the individual interview format, there is a very highly statistically significant difference in the group speaking test, where students seem to have the impression that they are interacting in a far more positive way than the rater perceives. This is an interesting issue, since it may be that the judgements made in the instance of the one-to-one interview-type test are quite accurate but about an imbalanced interactional situation, that is, the students' interactive abilities are restricted by the format of the test itself, but the rater recognises this and compensates for it in the scores. In the group test, the students may perceive themselves to be interacting in a more natural way, and hence they give themselves a higher mark. However, the raters award them not only a lower score than they give themselves, but also a lower mean score than they gave in the one-to-one interview. It is difficult to postulate a reason for this; perhaps in the one-to-one interview situation where candidates only have to respond in already-initiated dialogue there is less demand on students' interactive competence and they seem to be interacting with greater ease or spontaneity. In the group test, they need to be much more aware of turn-taking

strategies, the need to include others in the conversation, changing the direction of the conversation or initiating new topics. The raters' scores may reflect the fact that indeed our students are not very skilled at these things in English due to lack of practice or awareness, and if this is the case, we may identify a need for them to be included as specific learning targets in our study programmes.

Our results show that in the 'Group Speaking Test' there is a significant difference between the mean scores of both the interviewer and the rater when compared with the self-assessment scores of the students which may be due to various factors. The students had received scores from the first testing session which revealed that the rater had awarded them higher scores than they had given themselves in all the categories assessed. This may have led them to give much higher self-assessment scores in the second session. The group test format may also have contributed to a more positive self-perception than the interview with a socially superior examiner, and thus to higher self-assessment scores. However, we should also note that there is a significant difference between the mean score of the rater and the interviewer in the group test format, with the rater's score being the higher of the two. This shows that in the group test the students' own perception of their speaking competence is closer to that of the objective observer (rater) than to that of the interviewer.¹¹

More importantly, despite these differences in the scores themselves, we can observe a very strong positive correlation pattern of strengths and weaknesses between the rater and student self-assessment scores in all the categories of the group test,

¹¹ In the results for the one-to-one interview, the student self-awarded scores are closer to those of the examiner who is directly engaged in dialogue with them, but are significantly different (lower) in every category except Interaction.

which indicates that the students actually have a very clear and accurate perception of their own speaking performance. This may provide a sound basis in favour of the argument of including self-assessment criteria in our curricula and for training students in their use. Beyond the scope of the present study are the questions of motivation and metacognitive learning strategies that are involved in this issue, but if research into these areas indicates that self-assessment enhances motivation by allowing students to self-monitor progress and that this can lead to motivation for learning, then the argument for their inclusion in our study programmes would be further strengthened.

In the following chapter, *Conclusions*, we will give a general summary of these results in relation to the three main areas of our research project, test format, scoring and rating scales, and self-assessment. We will compare the two test formats, evaluate the rating scales and scoring procedure, and consider the impact our study may have on our current practice, including future decisions for the inclusion of self-assessment in our teaching and learning syllabus for the subject *Lengua BII*.

V. CONCLUSIONS

Whilst the present investigative study has been limited to a relatively small sample of students, and an even smaller number of examiners, it has highlighted some interesting general tendencies in our current practice of oral examining and, at the same time, suggested areas for further research which could continue to contribute to our understanding of some of the relevant issues in the field of testing speaking skills. This final chapter will summarise the findings of the present study, propose some changes to our current testing methods based on these results, and also describe the contributions and limitations of our investigation. Finally, we will propose some paths for future research that we believe will improve those aspects of oral examining that our results indicate may be desirable for our testing practices to become more effective.

V.1 SUMMARY OF RESEARCH RESULTS

V.1.1 Test format

From the perspective of both examiners and students, the results which we have analysed in the current research project appear to indicate that the group speaking test procedure is more effective than, and preferable to, the individual oral interview format in the context of our teaching and learning programme in the Faculty of Translation and Interpreting at the University of Las Palmas de Gran Canaria. Student responses to the questionnaires seem to indicate that their test-taking anxiety is reduced by the group test format. This is also reflected in their consistently higher self-assessment scores in this test which may have been brought about by enhanced self-esteem and a greater confidence in their own abilities when

involved in group interaction with their peers rather than with a 'high status' interviewer.

The fact that the task is similar to our classroom practice also probably helps to reduce nervousness in the test situation, since students who have attended classes with regularity are familiar with the procedure of group discussions and expressing opinions on topics which are in some way controversial. This could also have added to the high degree of acceptance of the test procedure and to an improved perception of performance as reflected in the questionnaire responses discussed in the previous chapter.

The examiners who took part in our study, while reticent to admit that they found difficulties maintaining the dual role of interviewer and rater in the 'Individual Oral Proficiency Interview', gave contradictory answers in the questionnaires which lead us to question whether they were actually carrying out the interview management and scoring in the way they believed. Even when they recognised that they had paid more attention to conducting the interview, they still affirmed that the marks they had awarded were accurate and fair according to both the global and analytic scales. This would lead us to conclude that, for the most part, examiners are probably unaware of their limitations in their own objective scoring ability, strengthening the argument in favour of conducting speaking tests with two examiners in different roles (interlocutor and rater) and providing descriptive scoring criteria to ensure greater standardisation of the test.

In the group test format, all the examiners reported feeling happy with the dual responsibility of managing the test and giving a global mark at the end based on a 5-point descriptive scale, although there were still questions as to the ease with which three candidates could be simultaneously assessed even by a rater whose

only role is to focus on assessing performance rather than interlocuting. There was a suggestion that this procedure could lead to the rater comparing performances rather than adhering to the scale and its descriptors. Individual examiner comments, however, emphasise that it is easier to score candidates from an objective rater position and that these scores are more accurate than those awarded by simultaneous rater/interviewers, reinforcing the argument in favour of using both in a speaking test.

Evidently, our findings are not conclusive with regard to the size of the speech sample produced by students in each test format, but they do suggest that at the very least candidates speak as much in the group test format as they do in the individual interview. It is possible that in fact they produce a larger sample of speech for assessment, which in turn would lead to the scores having greater validity and reliability, but we would need to conduct further empirical investigation to attempt to verify this.

Although these results do not conclusively show the group test format to be the most desirable test format available, they do suggest that there are grounds to favour it over the individual interview since both candidates and examiners seem to have a more positive attitude towards this type of testing procedure with regard to their involvement in the test situation, the scoring of the test, understanding those scores, and to the way in which the procedure reflects classroom practice and learning objectives as previously established in the teaching and learning programme and test preparation tasks.

V.1.2 Rating scales and scoring

From the analysis of the responses to the items addressing rating scales and scoring in our questionnaires, we can see that the traditional method of scoring on a scale from 0 – 10 is perceived by teacher/examiners to be a fair and objective way of assessing students despite the fact that it does not describe in any accurate or detailed way the levels of proficiency which are being measured. Over time, firstly as students and then as teachers, we have internalised the meaning of this marking system and believe we know how it corresponds to achievement. One of our objectives was to question this method of assessment and to attempt to introduce a new procedure which we believed would be both more meaningful and more useful to students, teacher/examiners and which could finally, with the passage of time, provide more useful information to prospective employers and other educational institutions.

Evidence for this internalisation process was revealed almost accidentally in the study in the way in which the interviewers and raters adapted the analytic scales to suit their own needs or beliefs about assessment. By including .5 and .75 in the scales they seem to be indicating a need to somehow distinguish between a performance that is somewhere between two descriptors, or which doesn't quite fulfil the description of the score provided in the scale. This would seem to reflect a tendency to unconsciously revert to the 0 – 10 scale with the intention of making a greater distinction between student performances according to norm-referenced criteria. If we attempt to write descriptors for ten (or more) levels of performance, we will find it extremely difficult, if not impossible, to precisely define and distinguish between so many distinct features of the speaking skill construct and our question therefore continues to be, how can we measure a language skill if we

cannot first describe or define it? Also, if we do persist in ‘measuring’ it, what does that measurement mean? The score given in the interview on a 0 – 10 scale, by definition, can only be used as an indicator of performance on that particular test since there are no fixed definitions of what it means beyond the test situation. In contrast, the analytic score should be generalisable to other situations and to underlying ability because it describes components of the speaking skill and to what extent these have been demonstrated as an aspect of test performance.

As teachers and assessors we need to ask ourselves continually what our criteria for assessing level and evaluating progress are, how we can assign meaning to them, and how they can be of use to those whose lives they have an impact on. If we are not engaged in this process, our professional development and contribution to the field is stagnant and serves only to perpetuate an existing, accepted system which is based on traditional use and habit rather than on theories of skill construct and measurement.

When attempting to achieve objective, consistent, and standardised rating, it is obviously necessary to train and standardise raters much more than we did for the trials we conducted as part of this study. It was partly our intention to explore the way in which raters apply rating scales and therefore we did not wish to ‘indoctrinate’ or influence them with our own ideas prior to the testing sessions. If we find sufficient evidence to justify the implementation of the group test as our test procedure for final assessment for the subject *Lengua BII*, we would attempt to achieve a more uniform application and understanding of this testing and scoring procedure by means of further exposure to and explanation of the aims of the descriptors, the scale and the rationale behind the group test format in order to bring about a greater standardisation of criteria for the raters involved.

The raters consistently gave higher scores than the interviewers in both tests. This may be because when they are also carrying out the role of interlocutor, the examiners are, for the most part, concerned with interview management, so scoring tends to take place retrospectively. If we are unsure when carrying out an assessment in a 'live' test situation, we may tend to be wary of being too generous if we are unsure of how to assess and manage the interview at the same time. By assigning the raters a more objective role which does not require them to take part in the interaction, and by providing them with a more informative descriptive scale, they may be able to focus on a wider range of features that make up the speaking construct, which in turn leads to greater confidence in awarding scores.

An interesting aspect to point out here is that the scores given by the raters in the group test are lower in every category assessed than those they gave in the one-to-one interview. These comparisons have not been made statistically in Chapter 4 because the candidates for the two tests were not exactly the same and therefore a direct comparison between the mean scores in each category assessed is not strictly possible in this type of study. However, almost two thirds (65%) of the students who took the one-to-one interview test also took the group speaking test, so we would expect their results to be representative of the whole group registered in the subject *Lengua BII* during that period. These findings may indicate that, in fact, in the one-to-one interview situation, examiners are more favourably biased with respect to the scoring of candidates than in the group test format (where they remain at a distance from the interaction for the major part of the test) through empathy with the anxiety that the students experience. Again, this may be an argument in favour of the apparently greater objectivity in scoring provided by the group test format, and the time and attention that is made available for it.

As regards the students' view of the scoring procedure, we have to admit to some disappointment in discovering that they continued to understand the meaning of a mark on a non-descriptive 0 – 10 scale better than a mark awarded using the analytic scale that described the features it attempted to measure. However, on reflection, it may be that this was inevitable; after so many years of receiving marks in this way, they too will have internalised an interpretation of their meaning and the descriptive scale may have even seemed to complicate the issue far more than necessary. Yet if we focus our attention on the students' opinions elicited by the questionnaire concerning their understanding of the steps they need to take to improve their speaking skills in English, the results are much more encouraging. It may be that, with further training in interpreting the meaning of scales such as these, together with the introduction of self-assessment techniques, we can highlight areas of strength and weakness for our learners and enhance motivation for learning.

V.1.3 Self-Assessment

The student reaction to both the concept and the procedure of self-assessment was extremely positive; the majority reported that they were capable of making accurate statements about their own speaking ability in English and also about their performance on the tests. It is interesting to note that they did not feel that their self-assessment scores for the tests were a true reflection of their speaking ability although they stated that they did accurately reflect their performance in the test, from which we can deduce that students believe that the test situation does not allow them to demonstrate their skills fully. This is probably due to the anxiety caused by the speaking test situation, which is generally felt to be at a higher level

than for written tests, since by nature it necessitates very rapid cognitive organisation of language and also involves a social situation where it is easy to 'lose face'.

Student reactions towards the questions addressing the possible effects of self-assessment on learning are also highly encouraging; almost all those questioned claimed that they considered that self-assessment could play a useful role if included in their learning programmes and the majority would also have liked it to be taken into account in their final mark for the entire subject. This reflects the positive attitude towards learning which we are currently engaged in promoting by means of the proposed changes that will take place on adapting our study programmes to the requirements of the European Higher Education Area by the end of the decade; it indicates a shift away from teacher-centredness towards more learner-centred modes of study and, by implication, assessment. These changes in our teaching/learning curricula will necessarily bring about the need for a change in assessment methods, since the two are linked and interdependent. Our current tendency is to see assessment as a final step in, or the product of, the teaching/learning process, something which happens as a conclusion to the teaching programme. Even what we call 'continuous assessment' is usually a kind of mini-testing or assessment of end-products throughout a course of study, rather than the evaluation of a process. What it may be possible to bring about by including self-assessment techniques throughout a teaching/learning programme is an on-going process of enhanced motivation and awareness of progress which will necessarily be centred on the individual in a far more personalised way than any teacher or tutor could ever design. By incorporating assessment techniques based on a learner-centred perspective (rather than an external one) at the starting point of

the programme of study (instead of at the finishing line), it may be possible to bring about a true change in our approach to teaching and learning that reflects these European aims, and not simply to turn around what we currently do on paper so that it looks like something innovative.

This will be a major challenge for teachers, not least for the reason that they are at present caught up in a torrent of paperwork, deadlines, mathematical calculations, and unclear instructions from the Spanish Ministry for Education and the EHEA about how to adapt what they currently do in their own teaching and assessment procedures to the new system without actually having to provide any clear rationale for doing so. It is the latter aspect which makes the process of real change so difficult, and which to some extent we can see reflected in the results from our teacher/examiner questionnaires. As teachers, we feel that the domain of assessment is necessarily external and believe that the fact that it comes from outside means that it is inherently objective. As far as we are concerned, students are not capable of being objective about themselves; the results of the present study, on the contrary, show that this belief is unfounded. Although there were significant differences between the self-assessment scores of the students and the test scores awarded by the raters in the categories of 'Pronunciation' and 'Interaction', these did not occur for 'Grammar and Vocabulary' or 'Discourse Management'. This, together with the very high level of correlation between the scores in all the categories assessed, indicates that, in fact, the students in this project had a very similar perception of their own performance on the test to that of the objective observer/rater although it was influenced slightly by the affective features of the test format (in the individual interview situation students gave themselves lower scores than the rater and in the group test, higher ones). These

findings warrant at least further investigation into self-assessment as a tool which can be used for both motivating learners by means of developing their metacognitive self-monitoring strategies and improving techniques for evaluating achievement.

V.2 PRACTICAL IMPLICATIONS OF THE STUDY

Having taken into consideration the results that we have presented in this study, we consider it appropriate to make some proposals for change to our current programme for the subject *Lengua BII (inglés)* in the degree programme of Translation and Interpreting at the University of Las Palmas de Gran Canaria. Although these will require further investigation and validation, as preliminary steps towards improvement in our approach we believe them to be justified by our initial findings and worthy of future implementation.

Our first proposal is to use the group speaking test format as the procedure for final assessment in the speaking component of the course. We believe that the scores it provides are a more accurate measure of our students' speaking ability than those yielded by the one-to-one interview format due to our attempt to define the construct and describe what we intend to measure with the rating scales. The effective implementation of these scales requires the presence of an objective rater who is not involved in the interaction and in order to make this an acceptable proposal in socio-affective terms of power structure, we would need to have at least two candidates in each test to balance the candidate-examiner ratio. For administrative reasons, since the number of examiners available is limited to those teachers involved in the subject during any given year and student numbers are relatively large, it is more practical to conduct the tests in groups of three rather

than in pairs. We also believe that the conversational interaction in a group rather than in a pair provides more scope for assessing a wider range of facets of interactive ability, such as turn-taking, encouraging another participant, involving other speakers in the conversation by asking questions, and so on.

We also propose a change in the scoring system we use to evaluate speaking skills in order to incorporate the descriptive rating scale we have designed. At present, there is no alternative other than to convert the scores from 1 to 5 it provides us with to a mark out of 10, but even so the score still has meaning since it refers to specific descriptions that correspond to a numerical mark on the scale. It may be possible to accommodate rater requirements for greater fine-tuning by adding in the .5 scores, but without a descriptor. The scale could be adapted by including a column between each whole number and descriptor for the .5 values which simply reads “some features of 3 and some of 4” as happens in some of the ‘Cambridge ESOL’ scales. Further empirical research and consultation with raters would be necessary to confirm that this improves the usability of the scale and allows for more accurate scoring.

Finally, our encouraging results on the possible roles and impact of self-assessment techniques in enhancing motivation, heightening awareness of learning processes, and monitoring progress lead us to propose that we include them as an important innovation in our programme of study and learning objectives for the subject *Lengua BII*. We believe that by providing students with criteria that are similar to the ones we ourselves use to assess their language skills, and by asking them to judge for themselves the extent to which they are able to achieve what the programme requires, we can effectively enhance their learning and raise metacognitive awareness. Again, fundamental in this assumption is the need to

define just what it is we wish to measure before attempting to measure it. We will need to turn our attention to construct definition in other areas of language learning, although based on the findings of the present research study these probably do not present as great a complexity and challenge as defining the construct of speaking.

It is our belief that this change in our approach to language teaching and learning could have a much-needed impact on our students' attitudes towards learning. In a society where almost everything is available on demand if you pay for it, our younger generations have grown up in a world of service-providers where practically all our needs and our leisure time are covered by monetary exchange. Education is no exception, and it is common to find that, as a sector, our students equate class attendance with learning and are generally unaware of the fundamental part they themselves play in the learning process or how much the responsibility for progress lies with themselves both inside and outside the classroom. This can also place considerable demands on teachers, who find themselves in constant search of ways to motivate and engage learners in the hope that they will command their attention for long enough to allow the objective of the lesson to be achieved. If we can involve students in their own learning process by highlighting areas and ways in which they can improve, we may go some way towards solving this problem and rendering the learning process more effective.

One of the more useful aspects of setting out what learners can do in descriptions of achievement for self-assessment protocols is that it focuses attention on positive attributes and achievements; as teachers we often tend to employ 'constructive criticism' in our assessments, highlighting learner errors in the hope that they will remedy them. However, criticism, by definition, is de-

motivating: it can only point out what we did wrong, or the goals we failed to achieve. In order to be constructive, we need to provide *observations* that will describe both what we can do or have mastered to date, and make suggestions for further steps towards improvement. A challenge for the future will be to provide marking scales with descriptors that can aid in this process and that are familiar pedagogical tools shared by teacher/examiners and learners.

V.3 CONTRIBUTIONS AND LIMITATIONS OF OUR STUDY

As we have mentioned above, one of the major limitations of our study is the reduced number of subjects we have worked with which necessarily limits the impact of the study and the generalisability of the results and conclusions. Especially in terms of the number of examiners involved, we would need to carry out trials with much larger numbers of raters and interviewers in order to confirm that the tendencies we have observed here in areas such as scale adaptation to suit internalised concepts, confidence in the tasks of interview management and simultaneous scoring, or the features of speaking focused on when awarding scores are common to all examiners in speaking test situations.

Additionally, we should have also included more specific questions in the Interviewer survey concerning the use of the 0 – 10 scale in order to establish at the very least what raters *think* they do when they apply it in the assessment of speaking performance. Since we did not make direct reference to this activity in our questionnaire, we can only hypothesise about the way in which the raters appear to use the scale in a norm-referenced way.

A further limitation was brought about by the voluntary nature of participation in the first test ('Individual Oral Proficiency Interview'). This meant

that the students who took the two tests were not exactly the same which translated into the impossibility of directly comparing certain statistics such as the scores given by the rater in each category of the speaking construct across the two tests. Due to this problematic aspect of our investigation, we can only comment on general tendencies rather than make affirmations about rater objectivity and the range of features of the speaking construct focused on during assessment.

Although we may not have made a significant contribution to the field of language learning and testing, we have found some grounds for change in our own context that have enabled us to propose a way forward for improving our teaching/learning objectives for the subject *Lengua BII* in the degree programme of Translation and Interpreting based on empirical evidence rather than intuition. This could possibly influence practice in other subjects with similar characteristics, at least in our own academic context, which may have the effect of enhancing co-ordination and continuity between study programmes, another aim of the EHEA modifications to all European university degree programmes.

We believe that our results provide sound evidence of the pedagogical value of self-assessment and hence support its inclusion in second or foreign language learning programmes. We have also shown that students *do* assess themselves objectively and in a similar way to raters when working with the same scales and descriptors. This is a good reason to continue to explore self-assessment as a valuable tool in our assessment procedures and we are confident in including it as a proposal for modification to our own context.

V.4 POSSIBLE AREAS FOR FUTURE RESEARCH

Naturally, the more we explore areas of our knowledge, the greater our awareness of how little we still know becomes. The present study has suggested other gaps in our understanding of the speaking construct and the appropriateness of our measurement tools which we consider to require further consideration and investigation.

First, we need to continue to develop our definitions of the construct in relation to oral testing in order to feel more confident that we are conducting measurements that correspond precisely to the skill of speaking. This is especially true in the area of interactive competence which was the category that produced major discrepancy between the student and rater scores in the group test, and which is currently the least documented in the literature. Further investigation into why our students perceived themselves to be interacting much more competently than the judgments made by the raters reflected may throw some light on just what communicative features spoken interaction involves and whether these differ across languages and cultures. If this is the case, then we would also need to consider its inclusion in our teaching programmes so that our students are informed about what we consider to be positive and negative contributions to group or paired interaction.

As far as the scoring of tests is concerned, we would like to suggest a future study that will address the cognitive processes that take place when we use the 0 – 10 rating scale and whose results, we hope, would support our theories of its norm-referenced nature and hence give us more grounds for requiring careful description in assessment. This would provide greater transparency and credibility for our actions as educators and also for our academic learning context.

We also propose to initiate a study to investigate the effectiveness of the introduction of self-assessment procedures in our teaching and learning context. In order to be able to confirm their positive impact on learning, we would need to design instruments for measurement and data collection with reference to the possible effects we expect it to have on both motivation and learning. This would also require making a distinction between these two aspects of the teaching/learning process and attempting to verify through objective data whether they are as directly related as we intuitively believe.

There are also studies that we could initiate with the data that we have collected for our present research project. By analysing the tapes and video recordings of all the interviews that took place, it would be possible to compare the average time available for candidate speech production in the individual interview and the group test formats and thus determine whether the group test, as we feel it should do, allows more time for students to speak because the interviewer is not taking up speaking time in most of the interaction. It may also be interesting to try to discover with this data whether the size of the sample produced by the candidate is related positively to the final mark they receive (i.e. the more they speak, the higher the mark). Were this to prove to be the case, it may indicate that raters focus more on interaction or at least 'willingness to participate', since another intuitive feeling here is that the more a candidate speaks, the more mistakes they are likely to make.

Finally, during the testing sessions the materials packs used were recorded on the candidate mark sheets. By analysing the scores awarded on the tests in relation to the test pack, we could also attempt to determine whether a relationship exists between discussion topics and scores. This would assist us in the design of

test materials and the production of comparable test packs which will guarantee equal conditions for all candidates. It would also hopefully be possible to discern which topics students find more interesting to express an opinion on or which are closer to their sphere of experience, and may also throw light on whether or not these topics of greater interest or ease for discussion coincide with those that have been included during the course of study. This might also allow us to make tentative deductions about the effects of some classroom practice on learning, or at least on the confidence produced in students by prior presentation of topics or vocabulary.

Whilst it is hoped that our investigative study has contributed in some small way to informing our current teaching and assessment practices, it is clear from this brief summary of our proposed areas for possible further investigations that our quest in finding a valid and reliable way of testing and assessing competence in speaking is only just beginning. We will thus endeavour to continue our investigation into the testing of speaking skills in order to provide empirical data on which to base our decisions for educational development and classroom practice as well as find ways to enable our learners to become more involved in the language learning process and assessment procedures in line with current learner-centred thinking in second language acquisition research.

BIBLIOGRAPHY

- Adams, M. L.** (1980). "Five cooccurring factors in speaking proficiency," in J.R. Frith (ed.), *Measuring Spoken Language Proficiency*. Washington DC: Georgetown University Press, 1-6.
- Alderson, C.** (1981). "Report of the discussion on general language proficiency," in J.C. Alderson and A. Hughes (eds.), *Issues in Language Testing*. London: The British Council, 187-94.
- Alderson, C.** (1991a). "Dis-sporting life. Response to Alastair Pollitt's paper: 'Giving students a sporting chance'," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 60-70.
- Alderson, C.** (1991b). "Bands and scores," in C. Alderson, and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 71-86.
- Alderson, J.C., Clapham, C. and Wall, D.** (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- Allan, D.** (1999). "Testing and assessment," *English Teaching Professional*, 11: 20.
- Angiolillo, P.** (1947). *Armed Forces Foreign Language Teaching*. New York: Vanni.
- Bachman, L. F.** (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L. F. et al** (1995). "Investigating variability in tasks and rater judgements in a performance test of foreign language speaking," *Language Testing*, 12, 2: 238-258.
- Bachman, L.F.** (2002a). "Alternative interpretations of alternative assessments: some validity issues in educational performance assessments," *Educational Measurement: Issues and Practice*, 21: 5-18.
- Bachman, L.F.** (2002b). "Some reflections on task-based language performance assessment." *Language Testing*, 19, 4: 53-76.
- Bachman, L.F. and Palmer, A.S.** (1996). *Language Testing in Practice*. Oxford: Oxford University Press.
- Bachman, L. F. and Savignon, S.J.** (1986). "The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview," *Modern Language Journal*, 70: 380-90.

- Barnwell, D.** (1989). "Naïve native speakers and judgements of oral proficiency in Spanish," *Language Testing*, 6, 2: 152-63.
- Batstone, R.** (1994). *Grammar*. Oxford: Oxford University Press.
- Bejar, I.I.** (1985). *A Preliminary Study of Raters for the Test of Spoken English*. Princeton, NJ: Educational Testing Service.
- Brindley, G.** (1991). "Assessing achievement in a learner-centred curriculum," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 153-166.
- Brown, A.** (2003). "Interviewer variation and the co-construction of speaking proficiency," *Language Testing*, 20, 1: 1-25.
- Brown, G. and Yule, G.** (1983). *Teaching the Spoken Language: An Approach Based on the Analysis of Conversational English*. Cambridge: Cambridge University Press.
- Brown, J. D.** (2001). *Using Surveys in Language Programs*. Cambridge: Cambridge University Press.
- Camacho Rosales, J.** (2000). *Estadística con SPSS para Windows*. Madrid: RA-MA.
- Canale, M. and Swain, M.** (1980). "Theoretical basis of communicative approaches to second language teaching and testing," *Applied Linguistics*, 1: 1-47.
- Canale, M.** (1983). "From communicative competence to communicative language pedagogy," in J.C. Richards and R.W. Schmidt (eds.), *Language and Communication*. London and New York: Longman, 2-27.
- Candlin, C.N.** (1987). "Towards task-based language learning," in C.N. Candlin and D.F. Murphy (eds.), *Language Learning Tasks (Lancaster Practical Papers in English Language Education)*, Vol. 7. Englewood Cliffs, NJ: Prentice Hall International, 5-22.
- Carpenter, K., Fyji, N. and Kataoka H.** (1995). "An oral interview procedure for assessing second language abilities in children," *Language Testing*, 12, 2: 157-175.
- Carroll, B. J.** (1991a). "Resistance to change," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 22-27.
- Carroll, B. J.** (1991b). "Response to Don Porter's paper: 'Affective factors in language testing'," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 41-45.

Chalhoub-Deville, M. (1995). "Deriving oral assessment scales across different tests and rater groups," *Language Testing*, 12, 1: 16-30.

Chalhoub-Deville, M. (2001). "Task-based assessment: a link to second language instruction," in M. Bygate, P. Skehan, and M. Swain (eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. Harlow, England: Longman, 210-228.

Chalhoub-Deville, M. (2003). "Second language interaction: current perspectives and future trends," *Language Testing*, 20: 369-383.

Chambers, F. and Richards, B. (1993). "Oral assessment: the views of language teachers," *Language Learning Journal*, 7: 22-27.

Chambers, F. and Richards, B. (1995). "The 'free conversation' and the assessment of oral proficiency," *Language Learning Journal*, 11: 6-11.

Chapelle, C. (1998). "Construct definition and validity inquiry in SLA research," in L. Bachman and A.D. Cohen (eds.), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, 32-70.

Chapelle, C. (1999). "Validity in Language Assessment," *Annual Review of Applied Linguistics*, 19: 254-272.

Clapham, C. (2000). "Assessment for academic purposes: where next?" *System*, 28, 4: 511-521.

Chipere, N. (2001). "Native speaker competence. Implications for first-language teaching," *Language Awareness*, 10, 2: 107-124.

Cranfield, S. and Clouet, R. (2006). "Using a real-life project in the Translation classroom," in S. Bravo Utrera y R. García López (eds.), *Estudios de Traducción: Problemas y Perspectivas*, Las Palmas: University of Las Palmas de Gran Canaria, Servicio de Publicaciones, 649-689.

Cristóbal Ruano, M.L. (1992). *El ordenador como fuente de estímulo y motivación en el aprendizaje del inglés*. PhD Thesis, Universidad Complutense, Madrid.

Davidson, F. and Lynch, B.K. (2002). *Testcraft. A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT and London: Yale University Press.

Davies, A. (1978). "Survey article: language testing," *Language Testing and Linguistics Abstracts*, 11, 3: 145-159 and 4: 215-229.

Davies, A. (1991). "Language testing in the 1990s," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 136-149.

- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T.** (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.
- Davis, P. and Rinvoluceri, M.** (1988). *Dictation: New Methods, New Possibilities*. Cambridge: Cambridge University Press.
- Dean Brown, J. and Rodgers, T.** (2002). *Doing Second Language Research*. Oxford: Oxford University Press.
- Douglas, D.** (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
- Elder, C., Iwashita, N., and McNamara, T.** (2002). "Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?" *Language Testing*, 19, 4: 347-368.
- Ellis, R.** (1985). *Understanding Second Language Acquisition*. Oxford: Oxford University Press.
- Ellis, R.** (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Folland D. and Robertson, D.** (1976). "Towards objectivity in group oral testing," *English Language Teaching Journal*, 30, 2: 156-67.
- Fulcher, G.** (1987). "Tests of oral performance: the need for data-based criteria," *English Language Teaching Journal*, 41, 4: 287-91.
- Fulcher, G.** (1994). "Some priority areas for research in oral language testing," *Language Testing Update*, 15: 39-47.
- Fulcher, G.** (1995). "Variable competence in second language acquisition: a problem for research methodology?" *System*, 23, 1: 23-51.
- Fulcher, G.** (1996). "Testing tasks: issues in task design and the group oral," *Language Testing*, 13, 1: 23-51.
- Fulcher, G.** (2003). *Testing Second Language Speaking*. Great Britain: Pearson Education.
- Fulcher, G. and Márquez Reiter, R.** (2003). "Task difficulty in speaking tests," *Language Testing*, 20, 3: 321-344.
- He, A.W. and Young, R.** (1998). "Language proficiency interviews: a discourse approach," in R.Young and A.W. He (eds.), *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam: John Benjamins, 1-24.

Hilsdon, J. (1991). "The group oral exam: advantages and limitations," in C. Alderson, and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 189-197

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hymes, D. (1979). "On communicative competence." Philadelphia, PA: University of Philadelphia Press. Reprinted extracts in C.J. Brumfit and K. Johnson (eds.), *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

ICEC. (1998). *Evaluación de la Enseñanza del Inglés. Educación Primaria*. Las Palmas de Gran Canaria: ICEC (Instituto Canario de Evaluación y Calidad Educativa).

INCE. (2002). *Evaluación de la Enseñanza y el Aprendizaje de la Lengua Inglesa. Educación Primaria*. Madrid: INCE (Instituto Nacional de Calidad y Evaluación).

INECSE. (2004). *Evaluación de la Enseñanza y el Aprendizaje de la Lengua Inglesa. Educación Secundaria Obligatoria*. Madrid: INECSE (Instituto Nacional de Evaluación y Calidad del Sistema Educativo).

Iwashita, N., McNamara, T. and Elder, C. (2001). "Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design," *Language Learning*, 5, 13: 401-436.

Kerlinger, F. N. and Lee, H. B. (2000). *Foundations of Behavioural Research. Fourth Edition*. Orlando, FL: Harcourt Brace.

Lado, R. (1961). *Language Testing*. London: Longman.

Lantolf, J.P. and Frawley, W. (1985). "Oral proficiency testing: a critical analysis," *Modern Language Journal*, 69, 4: 337-45.

Lantolf, J.P. and Frawley, W. (1988). "Proficiency: understanding the construct," *Studies in Second Language Acquisition*, 10, 2: 181-195.

Lazaraton, A. (1992). "The structural organization of a language interview: a conversation analytic perspective," *System*, 20, 3: 373-386.

Lazaraton, A. (1996). "Interlocutor support in oral proficiency interviews. The case of CASE," *Language Testing*, 13, 2: 151-172.

Lewkowicz, J.A. (2000). "Authenticity in language testing: some outstanding questions," *Language Testing*, 17, 1: 43-64.

Liskin-Gasparro, J.E. (1984). "The ACTFL proficiency guidelines: gateway to testing and curriculum," *Foreign Language Annals*, 17, 5: 375-389.

- Long, M. H.** (1989). "Task, group and task-group interactions," *University of Hawai'i Working Papers in ESL*, 8, 2: 1-26.
- Lowe, P.** (1987). "Interagency Language Roundtable Proficiency Interview," in J.C. Alderson, K. Krahnke, and C. Stansfield (eds.), *Reviews of English Language Proficiency Tests*. TESOL, 43-47.
- Lumley, T., and McNamara, T.F.** (1995). "Rater characteristics and rater bias: implications for training," *Language Testing*, 12, 1: 54-71.
- Lunz, M.E., Wright, B.D. and Linacre, J.M.** (1990). "Measuring the impact of judge severity on examination scores," *Applied Measurement in Education*, 3: 331-345.
- Lynch, B. K.** (2001). "The ethical potential of alternative language assessment," *Studies in Language Testing 11*. Cambridge, Cambridge University Press.
- Markee, N.** (2000). *Conversation Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Masden, H. S.** (1983). *Techniques in Testing*. Oxford: Oxford University Press.
- Matthews, M.** (1990). "The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations," *English Language Teaching Journal*, 44, 2: 117-121.
- 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: 446-465.
- Meara, P. and Buxton, B.** (1987). "An alternative to multiple choice vocabulary tests," *Language Testing*, 4, 2: 142-151.
- Messick, S.** (1996). "Validity and washback in language testing," *Language Testing*, 13, 3: 241-256.
- Morrow, K.E.** (1979). "Communicative language testing: revolution or evolution?" in C.J. Brumfit and K. Johnson (eds.), *The Communicative Approach to Language Teaching*, Oxford: Oxford University Press, 143-159.
- 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. and Schneider, G.** (1998). "Scaling descriptors for language proficiency scales," *Language Testing*, 15, 2: 217-63.
- Nunn, R.** (2000) "Designing rating scales for small group interaction," *English Language Teaching Journal*, 54, 2: 169-178.

- Nunan, D.** (1989). *Designing Tasks for the Communicative Classroom*. Cambridge, Cambridge University Press.
- Orr, M.** (2002). "The FCE speaking test: using rater reports to help interpret test scores," *System*, 30: 143-154.
- O'Loughlin, K.** (2001). "The equivalence of direct and semi-direct speaking tests," *Studies in Language Testing 13*. Cambridge: Cambridge University Press.
- O'Sullivan, B.** (2002). "Learner acquaintanceship and oral proficiency test pair-task performance," *Language Testing*, 19, 3: 277-295.
- Oxford, R.** (1990). *Language Learning Strategies: What Every Teacher Should Know*. Rowley, Mass.: Newbury House.
- Peinemann, M., Johnston, M. and Brindley, G.** (1988). "Constructing an acquisition-based procedure for second language assessment," *Studies in Second Language Acquisition*, 10: 217-234.
- Pollitt, A.** (1991a). "Giving students a sporting chance," in C. Alderson and B. North, (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 46-59.
- Pollitt, A.** (1991b). "Response to Charles Alderson's paper: 'Bands and scores'," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 87-94.
- Porter, D.** (1991a). "Response to Brendan Carroll's paper: 'Resistance to change'," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 28-31.
- Porter, D.** (1991b). "Affective factors in language testing," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 32-40.
- Porter-Ladousse, G.** (1993). *Language Issues: A Course for Advanced Learners*. Longman Group UK.
- Reed, D. and Cohen, A.** (2001). "Revisiting raters and ratings in oral language assessment," *Studies in Language Testing 11*. Cambridge: Cambridge University Press.
- Reves, T.** (1980). "The group oral test: an experiment," *English Teachers' Journal*, 24: 19-21.
- Reves, T.** (1991). "From testing research to educational policy: a comprehensive test of oral proficiency," in C. Alderson and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 178-188.

Richards, J. C. and Rodgers, T. S. (1986). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

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

Ross, S. (1998). "Divergent frame interpretations in language proficiency interview interaction," in R. Young, and A.W. He (eds.), *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam: John Benjamins, 333-353.

Saville, N. and Hargreaves, P. (1999). "Assessing speaking in the revised FCE," *ELT Journal*, 53, 1: 42-51.

Sacks H., Schegloff, E. and Jefferson, G. (1974). "A simplest systematics for the organisation of turn-taking for conversation," *Language*, 50, 4: 696-735.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Sfard, A. (1998). "On two metaphors for learning and the dangers of choosing just one," *Educational Researcher*, 27: 4-13.

Shohamy, E. (1983a). "Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew," in J.W. Oller (ed.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 229-236.

Shohamy, E. (1983b). "The stability of oral proficiency assessment in the oral interview procedure," *Language Learning*, 33, 4: 527-40.

Shohamy, E., Reves, T. and Bejerano, Y. (1986). "Introducing a new comprehensive test of oral proficiency." *English Language Teaching Journal*, 40: 212-220.

Skehan, P. (1988). "State of the art article: language testing. Part I." *Language Teaching*, 21, 4: 211-221.

Skehan, P. (1989). "State of the art article: language testing. Part II." *Language Teaching*, 22, 1: 1-10.

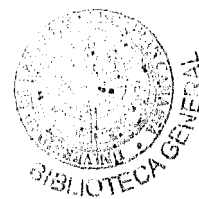
Skehan, P. (1991). "Progress in language testing: the 1990s." In C. Alderson, and B. North (eds.), *Language Testing in the 1990s*. Macmillan (reprinted under the Phoenix imprimatur, 1995), 3-21.

Skehan, P. (1996). "A framework for the implementation of task-based instruction," *Applied Linguistics*, 17: 38-62.

Skehan, P. (1998a). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

- Skehan, P.** (1998b). "Processing perspectives to second language development, instruction, performance and assessment," *Thames Valley Working Papers in Applied Linguistics*, 4: 70-88.
- Spolsky, B.A.** (1986). "What does it mean to know how to use a language? An essay on the theoretical basis of language testing," *Language Testing*, 2, 2: 180-191.
- Sollengberger, H.E.** (1978). "Development and current use of the FSI oral interview test," in J.L.D. Clark (ed.), *Direct Testing of Speaking Proficiency: Theory and Application*, Princeton, NJ: Educational Testing Service, 1-12.
- Swain, M.** (2001). "Examining dialogue: another approach to content specification and to validating inferences drawn from test scores," *Language Testing*, 18, 3: 274-302.
- Swain, M. and Lapkin, S.** (2001). "Focus on form through collaborative dialogue: exploring task effects," in M. Bygate, P. Skehan, and M. Swain (eds.), *Researching Pedagogic Tasks: Second Language Acquisition in Context*, London: Longman, 99-118.
- Tarone, E.** (1988). *Variation in Interlanguage*. London: Edward Arnold.
- Tarone, E.** (1998). "Research on interlanguage variation: implications for language testing," in L. Bachman and A.D. Cohen (eds.), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, 71-89.
- Taylor, L.** (2000a). "Investigating the paired speaking test format," *Research Notes 2*. Cambridge: University of Cambridge Local Examinations Syndicate, 14-15.
- Taylor, L.** (2000b). "Approaches to rating scale revision," *Research Notes 3*. Cambridge: University of Cambridge Local Examinations Syndicate, 14-16.
- Tomlinson, B.** (ed.) (1998). *Materials Development in Language Teaching*. Cambridge: Cambridge University Press.
- Underhill, N.** (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques*. Cambridge: Cambridge University Press.
- Upshur, J. A. and Turner, C. E.** (1995). "Constructing rating scales for second language tests," *English Language Teaching Journal*, 49, 1: 3-12.
- Upshur, J. A. and Turner, C. E.** (1999). "Systematic effects in the rating of second-language speaking ability: test method and learner discourse," *Language Testing*, 16, 1: 82-111.
- Wall, D.** (2000). "The impact of high-stakes testing on teaching and learning: can this be predicted or controlled?" *System*, 28, 4: 499-509.

- Weir, C.** (1990). *Communicative Language Testing*. Prentice Hall.
- Weir, C.** (1993). *Understanding and Developing Language Tests*. New York and London: Prentice Hall International.
- Widdowson, H.G.** (1978). *Teaching Language as Communication*. Oxford: Oxford University Press.
- Widdowson, H.G.** (1990). *Aspects of Language Teaching*. Oxford: Oxford University Press.
- Williams, M. and Burden, R. L.** (1997). *Psychology for Language Teachers: A Social Constructivist Approach*. Cambridge: Cambridge University Press.
- Wigglesworth, G.** (1997). "An investigation of planning time and proficiency level on oral test discourse," *Language Testing*, 14: 85-106.
- Wood, M. et al.** 2007. *FreconWin: Corpus Canario de Inglés Oral (Colección Evaluación e Investigación Educativa)*. Consejería de Educación, Cultura y Deportes del Gobierno de Canarias (Instituto Canario de Evaluación y Calidad Educativa).
- Xiaoju, L.** (1990). "In defence of the communicative approach," in R. Rossner and R. Bolitho (eds.), *Currents of Change in English Language Teaching*. Oxford: Oxford University Press.
- Young, R. and He, A.W.** (1998). *Talking and Testing. Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam: John Benjamin.



Appendix 1

FSI Rating Scale

Pronunciation

1. Often unintelligible.
2. Usually foreign but intelligible.
3. Sometimes foreign but always intelligible.
4. Sometimes foreign but always intelligible.
5. Native

Grammar

1. Accuracy limited to set expressions; almost no control of syntax; often conveys wrong information.
2. Fair control of most basic syntactic patterns; conveys meaning accurately in simple sentences most of the time.
3. Good control of most basic syntactic patterns; always conveys meaning accurately in reasonably complex sentences.
4. Makes only occasional errors and these show no pattern of deficiency.
5. Control equal to that of an educated native speaker.

Vocabulary

1. Adequate only for survival, travel and basic courtesy needs.
2. Adequate for simple social conversation and routine job needs.
3. Adequate for participation in all general conversation and for professional discussions in a special field.
4. Professional and general vocabulary broad and precise, appropriate to occasion.
5. Equal to vocabulary of an educated native speaker.

Fluency

1. Except for memorized expressions, every utterance requires enormous, obvious effort.
2. Usually hesitant; often forced to silence by limitations of grammar and vocabulary.
3. Rarely hesitant; always able to sustain conversation through circumlocutions.
4. Speech on all professional matters as apparently effortless as in English: always easy to listen to.
5. Speech at least as fluent and effortless as in English on all occasions.

Comprehension

1. May require much repetition, slow rate of speech; understands only very simple, short, familiar utterances.
2. In general understands non-technical speech directed to him, but sometimes misinterprets or needs utterances repeated. Usually cannot follow conversation between native speakers.
3. Understands most of what is said to him; can follow speeches, clear radio broadcasts and most conversation between native speakers, but not in great detail.
4. Can understand all educated speech in any moderately clear context; occasionally baffled by colloquialisms and regionalisms.
5. Equal to that of a native speaker.

Appendix 2

FSI Absolute Ratings

Level 1: Elementary Proficiency. *Able to satisfy routine travel needs and minimum courtesy requirements.*

Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are 'very familiar' and elementary needs vary considerably from individual to individual, any person at Level 1 should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases and tell time.

Level 2: Limited Working Proficiency. *Able to satisfy routine social demands and limited work requirements.*

Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family and autobiographical information; can handle limited work requirements needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e. topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though quite often faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

Level 3: Minimum Professional Proficiency. *Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social and professional topics.*

Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.

Level 4: Full Professional Proficiency. *Able to use the language fluently and accurately on all levels normally pertinent to professional needs.*

Can understand and participate in any conversation within the range of his experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

Level 5: Native or Bilingual Proficiency. *Speaking proficiency equivalent to that of an educated native speaker.*

Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers in all its features, including breadth of vocabulary and idiom, colloquialisms and pertinent cultural references.

FOREIGN SERVICE INSTITUTE: ABSOLUTE LANGUAGE PROFICIENCY RATINGS

The rating scales described below have been developed by the Foreign Service Institute to provide a meaningful method of characterizing the language skills of Foreign Service personnel of the Department of State and of other Government agencies. Unlike academic grades, which measure achievement in mastering the content of a prescribed course, the S-rating for speaking proficiency and the R-rating for reading proficiency are based on the absolute criterion of the command of an educated native speaker of the language.

The definition of each proficiency level has been worded so as to be applicable to every language; obviously the amount of time and training required to reach a certain level will vary widely from language to language, as will the specific linguistic features. Nevertheless, a person with S-3's in both French and Chinese, for example, should have approximately equal linguistic competence in the two languages.

The scales are intended to apply principally to government personnel engaged in international affairs, especially of a diplomatic, political economic and cultural nature. For this reason heavy stress is laid at the upper levels on accuracy of structure and precision of vocabulary sufficient to be both acceptable and effective in dealings with the educated citizen of the foreign country.

As currently used, all the ratings except the S-5 and R-5 may be modified by a plus (+), indicating that proficiency substantially exceeds the minimum requirements for the level involved but falls short of those for the next higher level.

The Measurement of Speaking [and Reading] Proficiency in a Foreign Language

With an ever-increasing demand for multi-lingual Americans to fill business, government, and academic positions both at home and abroad, a need has developed for a meaningful and efficient way to describe proficiency in a foreign language. Such terms as 'good', 'fluent', or 'bi-lingual', whether applied by teachers or supervisors to the competence of their students or employees or used as self-appraisal designations, have proved to be vague, un-measurable, and open to many interpretations.

Since 1956 the Foreign Service Institute of the Department of State has been rating government employees on a simple numerical scale which succinctly describes speaking and reading proficiency in a foreign language. This scale has become so widely known and well understood that statements like 'The consul has an S-2 R-3 in Thai' or 'That position requires someone with S-4 R-4 in French' are immediately intelligible within meaningful limits of accuracy to everyone concerned with personnel assignments in the numerous government agencies who use the FSI testing facilities.

The usefulness of the system is based on careful and detailed definition in both linguistic and functional terms of each point on the scale.

This paper is principally concerned with the description of the testing procedures and evaluation techniques whereby the rating system is currently applied at the Foreign Service Institute.

BACKGROUND

Prior to 1952 there was no inventory of the language skills of Foreign Service Officers and, indeed, no device for assessing such skills. In that year, however, a new awareness of the need for such information led to preliminary descriptions of levels of proficiency and experimental rating procedures. By 1956, the present rating system and testing methods had been developed to a practicable degree.

Both the scope and the restrictions of the testing situation provided problems and requirements previously unknown in language testing. The range of these unique features is indicated below:

- 1) The need to assess both speaking and reading proficiency within a half-hour to an hour. The requirement was imposed principally by the limited time available in the examinees' crowded schedule.
- 2) The need to measure the complete range of language competence, from the skill acquired in 100 hours of training or a month of experience abroad to the native facility of someone who received his entire education through the foreign language.
- 3) A population consisting of all the kinds of Americans serving the United States overseas: diplomats (from career ambassadors to visa officers), secretaries, agricultural specialists, Peace Corps Volunteers, soldiers, tax experts, and many others. They may have learned their language skills at home or on the job or through formal training, in any combination and to any degree. Generally no biographical information is available before the test.
- 4) The necessity for a rating system applicable to any language, easy to interpret by examiners, examinees, and supervisors, and immediately useful in decisions about assignments, promotions, and job requirements.
- 5) The need for unquestioned face validity and reputation of high reliability: those using the test results make decisions about the careers of others are themselves examinees and must have faith in the accuracy of their own ratings.

With these restrictions there was, from the beginning, very little choice in the kind of test that could be given. A structured interview custom-built to fit each examinee's experience and capabilities in the language promised to use the time allowed for the test with maximum efficiency. A rating scale with units gross enough to ensure reasonable reliability was developed on the basis of both linguistic and functional analyses.

Although both the testing procedure and the rating scale were first put to use more or less in their present form in 1956, a real measure of their effectiveness began in the summer of 1958, when the Department of State instituted a mandatory testing program. This lead was quickly followed by the U.S. Information Agency, the Agency for International Development, and then a number of other Agencies; and the Language Testing Unit now administers over 3,000 tests a year in approximately 40 languages. As a consequence, both testing techniques and rating criteria have been refined and elaborated to the point where they can be quickly and reliably learned by qualified linguists, language teachers, and native speakers.

The Examiners

Whenever possible, the testing team consists of a native speaker of the language being tested and a linguistic scientist thoroughly familiar with the language. Ideally, the native speaker is an experienced teacher of English speakers, has reasonably well-informed interest in current events throughout the world, and has a warm, friendly, and tactful curiosity about all kinds of people. The linguist need not speak the language fluently but should have very high aural comprehension and acute sensitivity to phonological, structural, and lexical errors. The more all these attributes are shared by both examiners, the smoother and more reliable the test.

It is sometimes the case that the only native speaker available has none of the desired characteristics. If this is true, the linguist will need to alter his own role to much more active one, as will be seen in the description of speaking test procedure below. If the linguist does not meet the normal requirements, he is forced to depend heavily on the native speaker and on his own experience in testing other languages.

Speaking Test: Procedure

The usual speaking test is conducted by the native speaker, with the linguist observing and taking notes. To the greatest extent possible, the interview appears as relaxed, normal conversation in which the linguist is a mostly silent but interested participant.

The test begins with simple social formulae: introductions, comments on the weather, questions like 'Have you just come back from overseas?', 'Is this the first time you've taken test here?', 'Did we keep you waiting long?'

The examinee's success in responding to these opening utterances will determine the course of the rest of the test. If he fails to understand some of them, even with repetition and rephrasing, or does not answer easily, at least a preliminary ceiling is put on the level of questions to be asked. He will be asked as simply as possible to talk about himself, his family, and his work; he may be asked to give street directions, to play a role (e.g., renting a house), to act as interpreter for the linguist on tourist level. Rarely, he may handle these kinds of problems well enough to be led on to discussions of current events or of detailed aspects of his job. Usually he is clearly pegged at some point below the S-2.

The examinee who copes adequately with the preliminaries generally is led into natural conversation on autobiographical and professional topics. The experienced

interviewer will simultaneously attempt to elicit the grammatical features that need to be checked. As the questions increase in complexity and detail, the examinee's limitations in vocabulary and structure normally become apparent quite rapidly. (A well-trained team usually can narrow the examinee's grade to one of two ratings within the first five or ten minutes; they spend the rest of the interview collecting data to verify their preliminary conclusion and make a final decision).

If the examinee successfully avoids certain grammatical features or if the opportunity for him to use them does not arise, or if his comprehension or fluency is difficult to assess, the examiners may use an interpreting situation appropriate to the examinee's apparent level of proficiency. In languages in which testing is relatively infrequent or in role-playing, a set of bi-lingual dialogs is often prepared and written down for this purpose. If the situation is brief and plausible and the interchange yields a sufficient amount of linguistic information, his technique is a valuable supplement.

A third element of the speaking test, again an optional one, involves instructions or messages which are written in English, given to the examinee to be conveyed to the native speaker, (e.g. 'Tell your landlord that the ceiling in the living room is cracked and leaking and the sofa and rug are ruined.'). This kind of task is particularly useful for examinees who are highly proficient on more formal topics or who indicate a linguistic self-confidence that needs careful exploration.

In all aspects of the interview an attempt is made to probe the examinee's functional competence in the language and to make him aware of both his capacities and limitations.

The speaking test ends when both examiners are satisfied that they have pinpointed the appropriate S-rating, usually after half to two-thirds of the allotted time for the whole test.

Speaking Test Evaluation

[...] A weighted scoring system [...] has been derived from a multiple correlation with the over-all S-rating assigned. Partly because the sample was based mainly on tests in Indo-European languages, partly because of a widespread initial suspicion of statistics among the staff, use of the scoring system has never been made compulsory or even urged, though the examiners are required to complete the check-list. The result has been that most examiners assign the S-rating on the basis of experienced judgment and compute the check-list score only in cases of doubt or disagreement. Nevertheless, the occasional verifications of the check-list profiles seem to keep examiners in all languages in line with each other (in the sense that an S-2 in Japanese will have much the same profile as an S-2 in Swahili); and those who once distrusted the system now have faith in it.

To the trained examiner, each blank on each scale indicates a quite specific pattern of behavior. The first two scales, 'Accent' and 'Grammar', obviously indicate features that can be described most concretely for each language. The last three ('Vocabulary', 'Fluency' and 'Comprehension') refer to features that are easy to equate from language to language but difficult to describe except in functional terms, speech on a scale more refined than these six-point ones.

[...]

When the mandatory testing program began in 1958, ratings were given on the basis of one-sentence definitions of each level. These definitions were written to be broad enough to cover every language but were so unspecific that standards varied from language to language and examinees often felt they were under-rated. As the number of tests and types of examinees increased, it was possible to set more and more concrete criteria. For S-ratings there are now three principal sources of clarification: the official amplified definitions, which describe both the linguistic and functional characteristics for each level; a chart called 'Factors in Speaking Proficiency' which specifies the minimum criteria at each level for the five check-list factors, and is therefore primarily linguistic in focus; and a questionnaire called 'A Checklist for Self-Appraisal of Speaking Proficiency'. This last is based almost entirely on functional skills and is designed not only to permit someone to appraise himself accurately but to give new or would-be examiners insight into the range and depth of mastery demanded at each level. These three documents provide sufficient information to enable even the most inexperienced tester (if he meets the basic qualifications) to determine an S-rating within a point of the rating given by trained examiners after he has observed three or four tests, and within half a point (that is, a 'plus') after he has observed ten or twelve.

[...]

Validity

Aside from recently developed tests (not yet standardized) in some European languages, there are no measures currently available beside those of FSI which test the full range of speaking [and reading] ability from beginner to native competence. The measures that exist are oriented toward the college student or literary specialist rather than the adult conducting his business or profession in a foreign language. As a consequence the only criterion available to us has been the acceptability of the S- [and R-] rating definitions and their application to actual performance for both the examinees and the end-users of test results.

By this criterion it seems safe to say that the tests are valid. Almost all the Government agencies using FSI testing facilities have established language policies which depend heavily on test scores for assignment to specific positions, for promotion, and for incentive awards. The examiners are made continuously aware that test ratings are commitments on the examinee's linguistic capacity to perform certain functions, and it is obvious that these commitments are being met with sufficient consistency to enable many different groups to rely on them without question.

The examiners themselves have generally accepted both procedure and rating system as valid measures of their competence. The complaints about results are now limited almost entirely to those who disregard basic grammatical features of the language or are unaware of them. Those who are both familiar with the rating criteria and aware of own limitations, now a very great majority, not only concede the accuracy of their scores but have become increasingly competent at self-appraisal.

Appendix 3

***Lengua BII* Rating Scales**

SELF-ASSESSMENT CRITERIA: SPEAKING (LENGUA BII)

	1	2	3	4	5
GRAMMAR AND VOCABULARY	My grammar and vocabulary is very limited and I can only speak about familiar topics. I use a limited range of grammar structures and am not very correct.	Somewhere between 1 and 3	My grammar and vocabulary is adequate to deal with any topic of discussion. I can usually communicate my message successfully, although I make mistakes.	Somewhere between 3 and 5	I use a wide range of grammar forms and vocabulary items and I can discuss almost any topic without difficulty. I make very few grammar mistakes which do not interfere with communication.
PRONUNCIATION	I am not confident that people understand my pronunciation.	Somewhere between 1 and 3	People can generally understand what I say although I have an L1 accent.	Somewhere between 3 and 5	People always understand what I say although I may have a slight L1 accent
DISCOURSE STRUCTURE	I find it difficult to order my ideas and connect them coherently. Sometimes I can't finish what I am saying.	Somewhere between 1 and 3	I can usually organize what I want to say well, although sometimes I need to start again to get it right.	Somewhere between 3 and 5	I can nearly always organize what I want to say coherently and I do not need to hesitate often.
INTERACTION	I find it difficult to participate in a conversation and I am not confident of either answering or asking questions.	Somewhere between 1 and 3	I get involved in the conversation, although sometimes I answer questions rather than ask them. I wait for my turn to listen and speak.	Somewhere between 3 and 5	I get fully involved in the conversation, both by asking and answering questions. I invite other people to join in and let them speak too.

SELF-ASSESSMENT CRITERIA: SPEAKING (LENGUA BII) - INTERVIEW

	1	2	3	4	5
GRAMMAR AND VOCABULARY	My grammar and vocabulary was very limited and I could only speak about familiar topics. I used a limited range of grammar structures and I am not confident that I was very correct.	Somewhere between 1 and 3	My grammar and vocabulary was adequate to deal with most of the topics of discussion. I communicated my message successfully, although I probably made some mistakes.	Somewhere between 3 and 5	I used a wide range of grammar forms and vocabulary items and I could discuss almost any topic without difficulty. I think I made very few grammar mistakes which did not interfere with communication.
PRONUNCIATION	I am not confident that the interviewer understood my pronunciation.	Somewhere between 1 and 3	The interviewer generally seemed to understand what I said although I have a foreign accent.	Somewhere between 3 and 5	The interviewer always understood what I said, although I may have a slight accent
DISCOURSE STRUCTURE	I found it difficult to order my ideas and connect them coherently. Sometimes I couldn't finish what I was saying.	Somewhere between 1 and 3	I was able to organise what I wanted to say well, although sometimes I needed to start again to get it right, or pause briefly to think.	Somewhere between 3 and 5	I nearly always organized what I wanted to say coherently and I did not need to hesitate often.
INTERACTION	I found it difficult to participate in the conversation and I was not confident of either answering or asking questions.	Somewhere between 1 and 3	I got involved in the conversation, although mostly I answered questions since I did not feel confident enough to ask any.	Somewhere between 3 and 5	I got fully involved in the conversation and I felt confident enough to ask questions as well as answer them.

SELF-ASSESSMENT CRITERIA: SPEAKING (LENGUA BII) – GROUP TEST

	1	2	3	4	5
GRAMMAR AND VOCABULARY	My grammar and vocabulary was very limited and I could only speak about familiar topics. I used a limited range of grammar structures and was not very correct.	Somewhere between 1 and 3	My grammar and vocabulary was adequate to deal with all the topics of discussion. I communicated my message successfully, although I made some mistakes.	Somewhere between 3 and 5	I used a wide range of grammar forms and vocabulary items and I could discuss almost any topic without difficulty. I made very few grammar mistakes and these did not interfere with communication.
PRONUNCIATION	I am not confident that the others understood my pronunciation.	Somewhere between 1 and 3	The others generally understood what I said although I have an accent.	Somewhere between 3 and 5	The others always understood what I said, although I may have a slight accent
DISCOURSE STRUCTURE	I found it difficult to order my ideas and connect them coherently. Sometimes I couldn't finish what I was saying.	Somewhere between 1 and 3	I structured what I wanted to say well, although sometimes I needed to start again to get it right.	Somewhere between 3 and 5	I nearly always organized what I wanted to say coherently and I did not need to hesitate often.
INTERACTION	I found it difficult to participate in the conversation and I was not confident of either answering or asking questions.	Somewhere between 1 and 3	I got involved in the conversation, although mostly I answered questions rather than asked them. I waited for my turn to listen and speak.	Somewhere between 3 and 5	I got fully involved in the conversation, both by asking and answering questions. I invited others to join in and let them speak too.

ORAL TEST ASSESSMENT CRITERIA (LENGUA BII) - RATER

	1	2	3	4	5
GRAMMAR AND VOCABULARY	Limited range of vocabulary. Has difficulty dealing with unfamiliar topics. Frequent grammatical inaccuracies.	Somewhere between 1 and 3	Adequate range of vocabulary to deal with most topics of discussion. Some grammatical inaccuracies, but message is transmitted successfully.	Somewhere between 3 and 5	Wide range of vocabulary sufficient to deal with discussion of almost any topic. Few grammatical inaccuracies.
PRONUNCIATION	Intrusive L1 accent makes understanding difficult. Inadequate rhythm and stress cause strain on the listener.	Somewhere between 1 and 3	Apparent foreign accent, but can generally be understood without too much strain on the listener. Attempts to use appropriate rhythm and stress.	Somewhere between 3 and 5	Slight foreign accent. Rhythm and stress close to native patterns. No instances of incomprehensibility.
DISCOURSE STRUCTURE AND DEVELOPMENT OF IDEAS	Speech disconnected and some utterances left unfinished. Repetition of simple structures and vocabulary.	Somewhere between 1 and 3	Speech is generally coherent, although there are some hesitations and restructuring. Adequate range of structures to deal with most topics of conversation.	Somewhere between 3 and 5	Nearly all utterances are coherently produced with a varied range of structures.
INTERACTION	Does not contribute to the interaction or attempt to develop the conversation either by answering or asking questions.	Somewhere between 1 and 3	Becomes involved in the conversation, although this may be more by responding than by initiating interaction. Is aware of turn-taking strategies.	Somewhere between 3 and 5	Plays a full role in the interaction, both by initiating and responding. Is able to involve others in the conversation and allows them to develop their turns without dominating.

0 = Sample insufficient for assessment.

ORAL TEST ASSESSMENT CRITERIA (LENGUA BII)
GLOBAL MARK – INTERVIEWER

1	2	3	4	5
Inadequate competence at this level. Frequent pauses and hesitations to search for language, and abandoned utterances cause excessive strain on the listener. Fails to interact appropriately.	Weak performance. Only attempts to construct utterances at a very basic level. Takes little part in the interaction or interacts inappropriately. Pronunciation difficult to understand.	Adequate control of spoken language skills. Reasonable grammatical accuracy throughout the test. May take a limited part in the interaction, but reacts appropriately. Pronunciation may cause some strain on the listener.	Good performance. Is able to use a range of grammar and vocabulary accurately. Interacts appropriately throughout and may encourage the participation of others. Pronunciation understood with relative ease.	Very good spoken skills. Often uses more complex structures with reasonable accuracy. Participates well in the interaction, both as listener and speaker and involves others in the conversation. Pronunciation understood with ease.

0 = Sample insufficient for assessment.

Appendix 4

Materials Packs for the ‘Individual Oral Proficiency Interview’ and the ‘Group Speaking Test’

'Individual Oral Proficiency Interview'

Introductory Phase

- What will you be doing this summer?
- What did you do last summer?
- Have you ever been to England / an English speaking country?
- Have you ever travelled abroad? **If 'no'** – If you had the opportunity, where would you go?
- Are you going on an Erasmus exchange next year?
- **(If 'no')** – Are many of your friends going? How will you feel without them?
- Did you vote in the elections? Was it the first time you had voted? What motivated you to vote?
- Do you think many young people voted? Why (not)?
- What kind of things do you do in your spare time?
- Tell us about a film you have seen recently. Did you enjoy it? Why (not)?

Salmon pink turns a safer shade of grey

Andrew Osborn in Brussels
and James Meikle

Its orangey-pink flesh glistens from countless supermarket shelves across Britain, but the highly prized salmon is about to undergo a change of colour thanks to a new European Union food safety edict.

Concerned that canthaxanthin, a chemical fed to farmed salmon to give them their bright hue, may also be harming people's eyesight, the maximum permitted amount of artificial colouring allowed in the fish by the EU is to be slashed by a factor of three. "Brighter eyesight or brighter salmon?" was how the European Commission described the stark choice this week.

The pigment is also fed to chickens to make their skin and eggs a brighter yellow; the maximum levels for poultry will also be cut. However, three-quarters of the eggs sold in Britain do not contain the chemical, and the levels fed to poultry are said to be well within the EU's new limits.

"Scientific assessments have shown that a high intake of canthaxanthins produces an accumulation of pigments in the retina, affecting the sight," David Byrne, the EU food safety commissioner, said. "The use of this feed additive is purely cosmetic, to colour food, and reduced levels of the additive will not adversely affect taste or quality."

The flesh of wild salmon is naturally pink because the fish consume large amounts of shrimps. However, almost 90% of the salmon sold in supermarkets is farmed, and there is no obligation to state on labelling that canthaxanthin has been used.

Salmon farmers feed large doses of the additive to fish because, they argue, consumers find the greyer shade that farmed salmon would naturally have to be a turn-off. "It's appealing to the eye," said Julie Edgar, communications director at Scottish Quality Salmon, Scotland's main trade body. "People traditionally associate salmon with pink and red." But using canthaxanthin carries a risk. Beate Gminde, a commission spokeswoman, said: "There's no such thing as zero risk. We do know that there's a possible impact in the long term [on eyesight], but it's impossible to quantify it."

Salmon pink turns a safer shade of grey

1. Were you surprised to discover that animals are fed substances to make their end-products look more appetising?
2. Do you think that the food we buy in supermarkets is 100% safe to eat?
3. Do you think the public should be given more information about the contents of the food they buy?
4. Would you be happy to eat food that didn't look so good, but that you knew had not had chemical products added to it?

Doctors say immigrants not diseased

Gaby Hinsliff, Jo Revill
and Martin Bright

Doctors have called for voluntary health checks to be offered to all refugees as the medical establishment hit back against claims that immigrants are responsible for spreading infectious disease.

Department of Health officials said they did not recognise figures purporting to show immigration was doubling the rate of HIV and increasing the risk of hepatitis B twentyfold, describing statistics used in the Sun newspaper and Spectator magazine as "manipulated".

Vivienne Nathanson, head of ethics at the British Medical Association, said that, far from importing disease, many asylum-seekers' health was damaged by coming to Britain and living on the breadline.

Voluntary screening, however, would be an acceptable move, she said: "We are in favour of having screening, voluntary screening, so that they can be checked and any health needs identified."

As the asylum debate intensified, Tony Blair's claims that Britain could drop its obligations to refugees under European human rights legislation were dismissed by legal experts as unworkable. The Prime Minister said concerns about terrorism had led the Government to consider withdrawing from the European Convention on Human Rights.

Doctors say immigrants not diseased

1. Why do you think a newspaper would publish manipulated statistics about immigrants spreading life-threatening diseases?
2. Do people in the Canary Islands believe this to be true of illegal immigrants who arrive here in small boats?
3. Do immigrants face a better or a worse life in their destinations?
4. Are you worried about immigration in the Canary Islands?

And 3m little piggies went to play...

Britain's 3 million pigs have been guaranteed a playful future following the introduction last week of welfare rules that make "toys" compulsory in their pens, writes *Martin Wainwright*.

Experiments in grunt-filled sheds such as Stuart Rowntree's, on the Pennines in Yorkshire, have convinced Brussels officials that fun and games are what a potential pork chop or bacon rasher needs. The Europe-wide requirements, which the Department of the Environment, Food and Rural Affairs (Defra) circulated to farmers last week, define pig toys as "manipulable materials".

A spokesman for Defra explained the theory behind the reforms: "For many years now vets have been suggesting that you put a football or something to kick around into the stall with a horse if it is restless," he said.

"Basically, the same is true for pigs. If you put in a football or you dangle a chain, they can nose it around and play with it. It's good for them and helpful."

Farmers who flout the rules can be fined up to £2,500.



Porkers playtime ... pigs on Stuart Rowntree's farm play ball with pigman Paul Fradgley Photograph: Don McPhee

And 3m little piggies went to play ...

1. Do you think it is possible that happy pigs will make better bacon?
2. Would you feel better about eating meat from animals that had lived in good conditions?
3. Would you ever consider becoming a vegetarian?
4. How much does our psychological well-being affect our physical health?

UK News

Brown tells drug giants: sort out Aids row

Patent rules must be relaxed to allow sale of generic medicines to poorer countries

Sarah Boseley

Gordon Brown has sent an uncompromising message to the multinational drug companies to recognise their responsibilities to help save millions of people dying of Aids and other diseases in poor countries.

In an interview with the *Guardian* the Chancellor of the Exchequer said that Britain was looking for a rapid settlement of trade negotiations aimed at loosening the patent rules to allow developing countries to buy cheap medicines for Aids and other diseases.

But the World Trade Organisation talks, which resumed this week, are in a quagmire. The US, which supports the interests of multinational drug companies with the tacit support of some European countries, insists that only the very poorest countries should be allowed to buy generic copies of patented drugs and then only for

HIV/Aids, malaria and tuberculosis.

Mr Brown said: "Nobody can stand outside the need for action here and nobody can claim special interests or special privileges when people are dying unnecessarily. It's time that all recognise the responsibilities to help avoid unnecessary deaths, and that means we've got to get an agreement for the trade round."

Brown tells drug giants: sort out Aids row

1. Is it immoral for giant drug companies to retain high prices for medicines which could save millions of lives and prevent disease from spreading across the world?
2. Do you think medicines really cost as much to produce as drug companies sell them for?
3. What measures could be taken to ensure that people all over the world can have access to the medicines they need?
4. What is more important – money or health?

20 NGO OPPORTUNITIES

As controversy surrounds the use of 'emotive appeals' by relief agencies, **Gary Younge** reports on the scramble to get ahead

Taking a first aid course

ONE lunchtime last month, TV news in Britain carried a story on the famine in southern Sudan. Within seconds the phonelines to Unicef were jammed. People were sending money. Lots of it.

On May 6 Unicef raised \$400,000 in one day. "If it had not been for the powerful visuals I doubt we would have raised even half that much," a Unicef spokeswoman said. The public sees the pictures, reaches for its credit cards and dials the numbers. It wants something done. It wants to put food in children's mouths and it does not care how it gets there.

The main aid agencies want the same thing with one key exception — each agency also wants to be the one to deliver the goods. Most are eager to get there first. That is what they are there for. They are charities but they are run like companies, with huge turnovers, marketing strategies and revenue targets.

They are in the market of misery and their currency is public sympathy. They find misery, they alleviate it, and then they collect more money so that they can start all over again. If they cannot tap the nation's compassion then they cannot do their job.

"The non-governmental organisation which is the first there gets to stake out the territory and is often the only one allowed by the local

government to work in a particular area," says Roger Riddell, a research fellow at the independent Overseas Development Institute in the UK. "But there may be another NGO that is cheaper, better qualified to do the job and with better programmes that encourage greater participation."



Powerful images are crucial to aid agencies' appeals to the public for funds

PHOTO: MARIANTONIETTA PERU

NGOs

1. What do you think about the way some NGOs use high-impact images to raise money?
2. Should NGOs be run like large commercial companies?
3. Do you support/Have you ever supported or given money to an NGO? Which one(s)
4. Today it is normal for advertisements and television programmes to use very powerful or shocking images. How do you think this affects generally?

Drugs giants seek key to life

Tim Radford

TEN pharmaceutical giants and five world-famous laboratories are to join forces in a £30 million attempt to create a new picture of humanity and a new kind of medicine.

Everyone shares 99.9 per cent of the 100,000 genes that make up humans, and the remaining 0.1 per cent accounts for all variety in people. These differences are hall-marked as SNPs or "snips" — single nucleotide polymorphisms — tiny changes in the genetic code. The consortium plans to identify at least 300,000 such changes, many of which could be markers for a propensity to diabetes, asthma, or other hereditary diseases.

Knowledge of these could lead to medical treatment specifically tailored to individuals.

Ultimately, using a new kind of diagnostic tool — a DNA chip — doctors should be able to tell, from a simple saliva test, whether a patient is likely to be allergic to penicillin, or could be at higher than usual risk of cancer or Alzheimer's disease, and produce the right treatment.

The Wellcome Trust is investing £9 million in the project. Partners in the SNP consortium are the Sanger centre in Cambridge and four top US laboratories, along with Astra-Zeneca, Bayer, Bristol-Myers Squibb, Hoffman La-Roche, Glaxo Wellcome, Hoechst Marion Roussel, Novartis, Pfizer and SmithKline Beecham.

Drugs giants

1. What do you think about scientists interfering in our genetic make-up?
2. Do you think medical research will really be able to “produce the right treatment” even if it manages to identify the propensity of a person to a particular disease?
3. Are you happy about genetic manipulation in crops and fruit and vegetables?
4. How far do you think scientific research will be able to solve our problems in the future?



Sarah Thomas: turning away from 'cattle market' PHOTO: CHRIS MOORE

Model quits, jaded by waif worship

Amelia Gentleman

THE fashion industry's obsession with abnormally slender physiques has come under fire again, this time from within.

Teenage model Sarah Thomas has announced her decision to quit the catwalk in Paris, New York and Milan this year because she can no longer tolerate the fashion world's compulsive worship of the skinny form.

These views, from someone who has witnessed the industry's flaws first hand, have fuelled the long-running criticism of a business that promotes unattainable and unhealthy ideals as the norm.

Described as one of Britain's modelling success stories, Ms Thomas, aged 18, already feels jaded by the international fashion world and is rejecting the chance to earn up to £6,500 a day in the autumn shows because of the "ghastly cattle market". She said: "I had begun to dislike putting up with the pressure to be thin. People want you to be skinnier all the time."

The model, now the "face" of the cosmetics company Cover Girl, also said the drink, drug and eating problems of other models horrified her.

Model quits

1. Why do you think models might suffer from eating disorders and drink and drug problems?
2. What influence do you think very thin models have on society?
3. How responsible is advertising for the way people behave today?
4. Do you think competitiveness is a positive or a negative quality?

Techno-babble enters lexicon

John Ezard

SOBBER English dictionaries already fight to record the newest trendy phrases of the present day. But last week Guinness Publishing went one better by compiling a handbook of buzzwords and techno-babble that will fall from the lips of tomorrow's teenagers.

The book, Guinness Amazing Future, forecasts a world divided between a "cosmetic underclass" and "surgiholics", between "screenagers" and youngsters desperate for "meatspace". Among office workers it predicts outbreaks of "prairie-dogging" in "cube farms", followed by stampedes for "break-out space".

Today's young middle-aged trendies are unlikely to mature with age, the book says. Two years ago the Oxford English Dictionary recorded "adulescent" — "a 30-35-year-old who has interests typically associated with youth culture". By 2020 Guinness expects this generation to have

turned into "adulescents" — old people addicted to youth culture.

The firm said its forecasts were not wild guesses. "We found all the phrases we list already being used in magazines and other specialised fields which affect our lifestyle," said a spokesman, Jon Cunningham. "We expect them to spread into general use as such phrases do."

Staff in call centres already call them cube farms — "open plan offices based around cubicles", say the editors. Prairie-dogging is "a sudden commotion in an open plan office which causes all other workers to look up from their desks".

Screenagers are "streetwise, techno-wired youths, born and raised in the digital age". They will scorn television and newspapers at The Outernet — "traditional media not on the Internet".

And by 2020 many of them will join a drive to rejoin meatspace — "the opposite of cyberspace, the real world".

Techno-babble

1. Do you agree that dictionaries should record newly invented language?
2. What might be some of the problems with doing this?
3. Why do you think these new words continue to appear?
4. What are some examples of words like this in your own language and what do they mean?

Protecting children from TV violence

Editorial

The question of violence and pornography on French television has caused much heated debate over the past six months. It was certainly right for the culture minister, Jean-Jacques Aillagon, in September to ask a panel of 38 experts — headed by the philosopher Blandine Kriegel — to make an accurate and dispassionate diagnosis of the situation, which could then form a basis for public discussion.

No one expected the panel to pull definitive solutions out of a hat. It was simply hoped that it would come up with wide-ranging information and ideas that would help decisions to be made on the issue. In this respect there can be no doubt that the report the panel has just completed will enable progress to be made in understanding the problems and will assist in the search for possible solutions.

One of the report's main conclusions is that violence and pornography on television have had an effect on what Kriegel describes as "the undifferentiated and ill-defined increase in violence and delinquency in every sector of our society".

There are those who deny the existence of a direct link between the content of television programmes and the state of our society, and who argue that what is shown on our screens only reflects what actually goes on in the world.

The panel counters their argument by attempting to gauge the effects of "the upsurge of violence on television", and it concludes that "the broadcasting of violent programmes clearly affects the behaviour of young people".

The panel suggests that pornography should be made inaccessible to children through various devices, and that television channels should be urged to set up codes of ethics to improve the way that they indicate the nature of the programmes they broadcast, and to stop broadcasting violent or pornographic programmes during the day.

The proposals are a step in the right direction. But they cannot spirit away all the ambiguities that are inherent in the very nature of the issue under examination, because there exists a form of violence in the world today that it will always be the duty of television news programmes to cover. *November 15*

Protecting children from TV violence

1. Do you agree that violence on television contributes to the increase in violence in today's society?
2. What could be done to restrict the amount of violence children watch on television?
3. What other reasons do you think there might be for the existence of 'cultures of violence' such as street-gangs and skin-head groups?
4. Do you think children watch more violent programmes on television now than when you were younger?

Pluperfect Virus Bugs E-Mailers

Bob Hirschfeld

AN INSIDIOUS new computer virus is spreading throughout the Internet. Named Strunkenwhite after the authors of a classic guide to good writing, it returns e-mail messages that have grammatical or spelling errors. It is deadly accurate in its detection abilities, unlike the dubious spell checkers that come with wordprocessing programs.

The virus is causing panic throughout corporate America, which has become used to the typos, misspellings, missing words and mangled syntax so acceptable in cyberspace. The CEO of LoseItAll.com, an Internet startup, said the virus had rendered him helpless. "Each time I tried to send one particular e-mail this morning, I got back this error message: 'Your dependent clause preceding your independent clause must be set off by commas, but one must not precede the conjunction.' I threw my laptop across the room."

A top executive at a telecommunications and long-distance company, 10-10-10-10-10-123, said: "With the number of e-mails I crank out each day, who has time for proper grammar? Whoever created this virus should have their programming fingers broken."

A broker at Begg, Barow and Steel speculated that the hacker who created Strunkenwhite was a "disgruntled English major who couldn't make it on a trading floor. When you're buying and selling on margin, I don't think it's anybody's business if I write that 'i meetinged through the morning, then cinched the deal on the cel phone while bareling down the xway.'"

The virus has left government e-mail systems in disarray. Officials at the Office of Management and Budget can no longer transmit electronic versions of federal regulations because their highly technical language runs foul of Strunkenwhite's dictum that "vigorous writing is concise." The White House speechwriting office reported that it had received the same message, along with a caution to avoid phrases such as "the truth is . . ." and "in fact . . ."

Strunkenwhite is particularly difficult to detect because it doesn't come as an e-mail attachment but is disguised within the text of an e-mail entitled "Congratulations on your pay raise."

Pluperfect virus

1. Do you think incorrect grammar and spelling are acceptable in e-mail and text messages, or on the Internet?
2. How important is it to be correct in other circumstances?
3. Why do you think someone would create a computer virus of the kind mentioned in the text?
4. Do you feel strongly about how people use your language?
Why do you think many people feel strongly about their language?

Study finds oceans of old plastic

Tim Radford

Humans are smearing the oceans with plastic, according to British scientists who sifted shorelines to find microscopic fragments of stockings, yoghurt pots, rope, shopping bags and bleach bottles everywhere they looked.

The spread of polymer waste has been reported before: researchers have surveyed beaches on uninhabited islands in Antarctica and found plastic cups, polymer sandals and drinks' bottles.

But Richard Thompson and colleagues at the University of Plymouth reported in *Science* last week that they looked at apparently clean sand and mud on British beaches, in intertidal estuaries and even under 9m of water for evidence of invisible pollution. "We found microscopic fragments almost from the first sample. Since then we have looked at more than 20 sites around the UK and this material has been present at all of them, from Land's End to the north of Scotland," he said. "We are finding just as much in remote parts as we are nearer the big centres."

Plastics wash up on beaches to be repeatedly broken by the pounding waves. The team searched for nylon, polyester, acrylic and six other kinds of polymer with a clear chemical "signature". But they believe that they have underestimated the spread of human debris.

They could not identify plastics produced more than 20 years ago, and they could not pick up evidence of particles smaller than 20 microns. But they have clear evidence that long after plastic bags, nylon ropes and Tupperware boxes have vanished, their constituent fragments remain. Nobody knows whether this material can get into the food chain: that is the next line of research.

"If we look at the larger plastic debris accumulating on the shoreline, the most common items are things like plastic bags and boxes and packaging and, ironically, they are all items that needn't be there," Dr Thompson said. "So there is a challenge to all of us to reduce the amount of disposable plastic we use, to recycle things as much as possible."

Study finds oceans of old plastic

- Are people aware of how much plastic they use and throw away?
 - What are the advantages and disadvantages of using plastic?
 - Is plastic waste a problem in the Canary Islands?
-
- How much of your rubbish do you separate and recycle?

'Group Speaking Test'

Introductory Phase

- What will you be doing this summer?
- What did you do last summer?
- Have you ever been to England / an English speaking country?
- Have you ever travelled abroad? **If 'no'** – If you had the opportunity, where would you go?
- Are you going on an Erasmus exchange next year?
- **(If 'no')** – Are many of your friends going? How will you feel without them?
- Did you vote in the elections? Was it the first time you had voted? What motivated you to vote?
- Do you think many young people voted? Why (not)?
- What kind of things do you do in your spare time?
- Tell us about a film you have seen recently. Did you enjoy it? Why (not)?

Teachers demand remedy for violence

In a chilling new insight into rising violence in British schools, teachers have called for airport-style security checks to identify pupils carrying concealed weapons. They also want compulsory behaviour management courses for parents of unruly children.

The National Association of Schoolmasters/Union of Women Teachers, whose members called for tough anti-violence measures at their annual conference, have conducted surveys indicating that teachers are frequently abused, both physically and verbally. One delegate told of being hit in the back by a ball-bearing fired from a gun. Although police said the weapon was potentially lethal, the pupil was excluded from school for just three days. The conference voted for permanent exclusion of all pupils found with weapons.

The conference was just part of the traditional spring season for teacher union conventions. This year there were clear signs that the profession is becoming more outspoken. In spite of the billions being poured into education, many teachers still feel underpaid, overworked and unappreciated. Doug McAvoy, the retiring general secretary of the biggest union, the National Union of Teachers, used his final speech to conference to deliver a blistering attack on the Government, which he said was hell-bent on dismantling state education. He accused ministers of wanting "schools to be run like Tesco stores", offering two lessons for the price of one.

Teachers demand remedy for violence

- What might explain the rise in violent behaviour among school pupils?
 - Is this kind of problem common in schools in the *Canary Islands*?
 - What measures could be taken to protect teachers from violent pupils?
-
- Would you consider becoming a school teacher?

Parking wars: wardens want body armour

Hugh Muir

London's parking wardens are used to the withering stare, the verbal abuse, even the odd shove. It comes with the job. But as hostility between those who police London's parking spaces and the drivers who seek to occupy them grows, one set of wardens have appealed to be given anti-stabbing security vests.

Officials of the union Unison are demanding police-style protection on behalf of members in Hammersmith, west London, because they say they need protection from the daily attacks they suffer from enraged motorists.

They, and colleagues across London, are being kicked, punched and threatened with knives. Occasionally they are shoved into the path of oncoming traffic. But if tempers are rising it is because both wardens and drivers both feel they are operating under incredible pressure, and both sides claim that they are in the right.

London's drivers paid more than £162m last year in parking tickets. Westminster council made the most, from issuing 976,476 fines which raised about £39m — 7% more than the year before.

Geoff Martin of Unison said such



A traffic warden in Westminster, which raised £39m from parking tickets last year Photograph: Frank Baron

frenetic activity was taking its toll. "At the moment these wardens have the worst jobs in London. A lot are on performance-related pay. We are hearing of people being pressurised and humiliated if they don't issue enough tickets. The whole thing stinks."

A union organiser in southwest London, Aiden Grimes, said: "The turnover among the wardens here is phenomenal because they are continually caught between highly stressed managers demanding more tickets and the abuse they get on the street.

Wandsworth needs a complement of 84 people but they have had over 500 under contract in four years. That tells you something."

One warden, too frightened to be named, said the hostility was causing people to rethink: "Right now we are being told, 'Do your job safely. Only ticket when it is safe to do so. And at the first sign of trouble call for back-up right away'."

But motorists feel that they are under attack from the congestion charge, bus-lane restrictions, road

cameras and over-zealous wardens. A survey by the Freight Transport Association showed that the number of parking tickets issued to delivery drivers in London last year rose by an average of 78%.

Earlier this month Westminster council tried to calm tempers. It has told wardens to adhere to the protocol that gives illegally parked drivers two minutes' grace for dropping off passengers. Delivery drivers who park on yellow lines must be given a 20-minute window to load or unload.

Parking wars: wardens want body armour

- Are drivers justified in arguing with traffic wardens who give them a parking fine?
 - Do local councils make too much easy money from giving drivers traffic fines for minor offences?
 - Would you pay a parking fine if you got one? Why/Why not?
-
- Do you think city centres should be closed to traffic?

International News

Superweed warning as GM soya 'miracle' in Argentina turns sour

Paul Brown

Seven years after GM soya was introduced to Argentina as an economic miracle for poor farmers, researchers claim it is causing an environmental crisis, damaging soil bacteria and allowing herbicide-resistant weeds to grow out of control.

Soya has become the cash crop for half of Argentina's arable land, more than 11m hectares, most on fragile pampas lands. After Argentina's economic collapse, soya became a vital cash export providing cattle feed for Europe and elsewhere. Now researchers fear that heavy reliance on one crop may bring economic ruin.

The GM soya, grown and sold by Monsanto, is the company's great success story. Programmed to be resistant to Roundup, Monsanto's

patented glyphosate herbicide, soya production increased by 75% over five years to 2002, and yields increased by 173%, raising more than \$5bn profits for farmers who had been hard hit financially.

However, a report in *New Scientist* magazine says that, because of problems with the crops, farmers are now using twice as much herbicide as in conventional systems.

Soya is so successful it can be viewed as a weed itself: soya "volunteer" plants, from seed split during harvesting, appear in the wrong place and at the wrong time and need to be controlled with powerful herbicides, since they are already resistant to glyphosate.

The control of rogue soya has led to a number of disasters for neighbouring small farmers who have lost

their own crops and livestock to the drift of herbicide spray.

So keen have big farmers been to cash in on the soya bonanza that 150,000 small farmers have been driven off the land so that more soya can be grown. Production of many staples such as milk, rice, maize, potatoes and lentils has fallen.

One of the problems in Argentina is the rapid spread of weeds with a natural resistance to Roundup. Such weeds, say opponents of GM, could develop into a generation of "superweeds" impossible to control. The chief of these is *equisetum*, known as mareetail or horsetail, a plant that can rapidly choke fields of soya.

Superweed warning as GM soya 'miracle' in Argentina turns sour

- Do you agree with the genetic manipulation of crops?
- What are some of the consequences of the genetic manipulation of food?
- Do you think genetic manipulation should be used in humans for medical purposes?

- Do you think the food you buy in supermarkets is totally safe to eat?

Tobacco subsidy to end

The EU is to withdraw its massive subsidies to tobacco growers following a bitter battle among agricultural ministers in Brussels, *write Paul Brown and Ian Black in Brussels.*

The withdrawal of payments for what is the most subsidised crop in Europe reflects unease about helping tobacco farmers while EU states campaign against smoking. The UK, which pays \$155m of the \$1.4bn annual subsidy, was among a group of northern European states that demanded an end to the payment.

The EU has 1,000 tobacco growers

and is the world's fifth largest tobacco producer, with 75% of its crop being grown in Greece and Italy.

Smoking kills an estimated 500,000 Europeans a year, yet EU farmers are paid \$9,274 a hectare to grow tobacco. Wheat farmers receive \$424 a hectare.

In the UK alone the health service spends \$2.6bn a year treating people with smoking-related diseases. The government spends around \$53m on anti-smoking education campaigns and another \$70m helping people to stop smoking.

Tobacco subsidy to end

- Do you think the EU should subsidise tobacco farmers?
- Do anti-smoking campaigns or warnings on cigarette packets really have any effect?
- Should smoking be allowed in public places?
- Do you think it is easy to give up smoking?

Plan to make maths count for pupils

A catastrophic shortage of qualified maths teachers could force the Government to waive university fees, and even to pay students for taking the subject. That drastic solution, costing about £100m a year, has emerged from an inquiry into the slump in the number of young people taking maths at GCSE and A-level, and going on to study it for a degree.

The author of the government report, Professor Adrian Smith, describes "a dire, catastrophic, crisis-level shortage of specialist maths teachers".

More than one in four maths lessons in secondary schools are taught by under-qualified teachers. There is an estimated shortage of 3,500 maths teachers, and at the same time the number of pupils taking A-level maths has slumped by 20%. "We seriously have to look at financial incentives, either for the kids to take maths, or maybe the Government really has to look at a bung, a fee waiver, if you go to university to take maths," said Prof Smith. "The other thing you could do is pay universities to make maths a prerequisite for entry into certain popular areas. I think it's so serious, you've got to reach for the levers."



Plan to make maths count for pupils

- What might be some of the reasons for young people not wanting to study maths?
 - How would you feel if you knew that a student doing a degree in Maths did not have to pay, while you were paying to do a language-related degree?
 - If you were good at maths, would you consider becoming a maths teacher?
-
- Did you like maths at school? Why/why not?

France to ban pupils' religious dress

Jon Henley in Paris

Muslim headscarves and other religious symbols are almost certain to be banned from French schools and public buildings after a special commission told the government last week that legislation was needed to defend the secular nature of the state.

The 20-member group, appointed by President Jacques Chirac and headed by the national ombudsman, Bernard Stasi, recommended that all "conspicuous" signs of religious belief — including Jewish skullcaps, oversized Christian crosses and Islamic headscarves — be outlawed in state-approved schools.

The report, compiled after six months of study and more than 120 hearings, also recommended that the laws should include a clause requiring "the strict neutrality of all public service employees". Some Muslim women had reportedly been insisting that their husbands accompany them at all times in hospital and would

accept only female doctors. The report said the legislation must remind all health service users that "it is forbidden to reject a healthcare worker, and that the rules of hygiene must be respected".

In a gesture of respect to "all spiritual options", the report said the Jewish and Muslim holy days of Yom Kippur and Eid should be made official school holidays, and companies should consider ways of allowing their employees to take off the religious holiday of their choice.

Mr Chirac, who hinted last week that he favoured a law protecting France's secular republic, said he would make his decision known this week. "I will be guided by respect for republican principles and the demands of national unity and the rallying of the French people," he said.

The question of whether a "secularism law" is desirable or necessary — particularly to deal with the increasing number of Muslim girls wanting to wear headscarves at

school — may seem abstract, or even absurd, to those used to British or US notions of multiculturalism. In France, where secularism is a constitutional guarantee and everyone, in the eyes of the republic, is supposed to be equally French regardless of ethnic or religious differences, the issue has dominated media and political debate for several months.

Mr Stasi said the proposed law aimed to preserve constitutional secularism and counter "forces trying to destabilise the republic", a clear reference to Islamic fundamentalism. But he stressed that the law was not directed at the mainly moderate Muslim community of 5 million. "Muslims must understand that secularism is a chance for Islam," Mr Stasi said. "Secularism is the separation of church and state, but it is also the respect of differences."

The main teachers' union, the SNES, said that the proposals did not go far enough to promote secularism in schools.

France to ban pupils' religious dress

- Was the French government right to ban all symbols of religious belief from state schools and jobs?
 - Does this kind of legislation lead to fewer problems in society?
 - Do you think religion should be an assessed subject in the school curriculum?
-
- Should immigrants be expected to assimilate a new country's customs?

Deafening message for clubbers

It's official: the infernal racket made by nightclub music is seriously damaging the health of clubbers. A covert study commissioned by the RNID charity for deaf people, found that the decibel count on the noisiest dance floors was as high as 110 — about the same as an aircraft taking off or a pneumatic drill operating 3m away, and 20 decibels higher than the level recommended for workplaces.

“Someone who goes clubbing once a week could potentially be putting their hearing at risk, even if they only spend a few minutes on the dance floor on each occasion,” said a RNID statement.

Deafening message for clubbers

- Why do you think young people are attracted to the clubs that play the loudest music?
- Should there be controls on how loud nightclub music is played?
- In your opinion, what are the features of a good nightclub?

- Are there any pubs or clubs in Gran Canaria that play the music too loud.

Naomi Campbell wins media privacy fight

Steven Morris, Claire Cozens
and Owen Gibson

The supermodel Naomi Campbell last week won a landmark privacy ruling against the Daily Mirror that could have implications for the way the media deals with public figures.

In the most important privacy case since the implementation of the Human Rights Act in 2000, Ms Campbell was awarded £3,500 damages after the Mirror revealed that she was a drug addict.

The law lords ruled that though the tabloid was entitled to reveal that Ms Campbell was an addict, because she had always made a point of distancing herself from drugs, it had committed a breach of confidence by revealing details of her treatment and printing a photograph of her outside a meeting of Narcotics Anonymous.

The Mirror's editor, Piers Morgan, led criticism of the law lords' ruling, describing it as a "backdoor privacy law". He said: "This is a very good day for lying, drug-abusing prima donnas who want to have their cake with the media and the right to then shamelessly guzzle it with their Cristal champagne."

Other tabloid executives said they feared that the ruling could hamper exposés because they could be sued



Naomi Campbell outside the House of Lords Photo: David Bebber/Reuters

for revealing intimate details that backed up true stories.

Ms Campbell, 33, had conceded that the Mirror was within its rights to reveal in February 2001 that she was receiving treatment for drug addiction. But she claimed that the paper had overstepped the mark by revealing details and printing the photograph.

The five law lords overturned a decision in the Court of Appeal last year that publication was justified in the public interest because Ms Campbell had courted publicity and claimed that she did not take drugs. Three of the lords ruled in the supermodel's favour, while two backed the newspaper.

Naomi Campbell wins media privacy fight

- Should public figures who exploit the media to earn money, have the right to privacy when they want it?
 - How much does the general public have the right to know about famous people?
 - Are you interested in the life of any famous or popular person?
-
- Should the media be allowed to make public details of important people's private lives?

Mourning sickness marks a selfish culture

Matthew Taylor

People who wear ribbons to show empathy with worthy causes and mourn in public for celebrities they have never met are part of a growing culture of "ostentatious caring which is about feeling good, not doing good", according to a new study.

The report, *Conspicuous Compassion*, was published this week by the thinktank Civitas. It argues that the trend towards public outpourings of compassion reveals not how altruistic society has become, but how selfish. Its author, Patrick West, said: "We sport countless empathy ribbons, send flowers to recently deceased

celebrities, weep in public over the deaths of murdered children, wear red noses for the starving in Africa, go on demonstrations to proclaim Drop the Debt or Not in My Name . . . [but] they do not help the poor, diseased, dispossessed or bereaved. Our culture of ostentatious caring concerns, rather, projecting one's ego, and informing others what a deeply caring individual you are."

Mr West says that public displays of grief have spiralled out of control in the past decade: "We live in a post-emotional age, one characterised by crocodile tears and manufactured emotion . . . Mourning sickness is a religion for the lonely crowd that no

longer subscribes to orthodox churches. Its flowers and teddies are its rites, its collective minutes' silence its liturgy and mass. But these bonds are phoney, ephemeral and cynical."

Mr West concludes that instead of "piling up damp teddies and rotting flowers to show how nice they are" people should try to do some "un-ostentatious good".

Meanwhile, speaking last week at the conference of the National Council of Voluntary Organisations, the Chancellor, Gordon Brown, issued a "call to service" to boost volunteering, and proposed the extension of a scheme for young people to spend a gap year working in the community.

Mourning sickness marks a selfish culture

- Do you agree that public displays of grief, sadness or indignation are really selfish?
- Have you ever been on a demonstration? Why did you go?
- How far do you think public opinion can influence government decisions?

- Have you ever been affected by a tragedy that has happened to someone else?

Female builders could bridge gender gap

- Can women be good builders, plumbers, electricians, etc.?
- Why do you think there are so few women in the construction trade/industry?
- Would you be happy for a woman to mend your kitchen sink or fly a plane you were travelling on, or for a man to look after your children?
- Are men and women naturally better at different jobs (or can they learn to do any kind of work)?

Appendix 5

Questionnaires

NAME _____

QUESTIONNAIRE 1 – STUDENT

Please fill in the questionnaire about the test you have just done. Put a **circle around the number** of the answer that most accurately corresponds to what you think about the statement. If you make a mistake or change your mind, put a cross through the incorrect answer and circle the one you have chosen.

	Strongly disagree	Disagree	Agree	Strongly agree
1. I felt nervous throughout the whole test	1	2	3	4
2. I think I did well in the test (Give yourself a mark from 1-10):	1	2	3	4
3. I performed to the best of my ability in the test	1	2	3	4
4. I think I spoke enough for the tester to judge my ability	1	2	3	4
5. I was happy about the procedure of the test ..	1	2	3	4
6. The test was similar to the kind of task done in class	1	2	3	4
7. I could answer the questions without difficulty	1	2	3	4
8. I could find enough to say about the topic ...	1	2	3	4
9. The global mark I received was a fair mark ..	1	2	3	4
10. The analytic mark I received was a fair mark	1	2	3	4
11. I understand what my global mark means ..	1	2	3	4
12. I understand what my analytic mark means	1	2	3	4
13. The global mark I received was easier to understand than the analytic mark	1	2	3	4
14. The global mark helped me to understand what steps I need to take in order to improve my speaking	1	2	3	4
15. The analytic mark helped me to understand what steps I need to take in order to improve my speaking	1	2	3	4

Please add any other comments you would like to make about the test itself and/or your experience of the test in the box below.

Thank you for taking the time to co-operate in this project.

QUESTIONNAIRE 2 – INTERVIEWER

Please fill in the questionnaire about the test you have just done. Put a **circle around the number** of the answer that most accurately corresponds to what you think about the statement. If you make a mistake or change your mind, put a cross through the incorrect answer and circle the one you have chosen.

	Strongly disagree	Disagree	Agree	Strongly agree
1. I was able to manage the interview and give the student a global mark on a scale of 1-10	1	2	3	4
2. I was able to manage the interview and give the student a detailed score at the end of the interview	1	2	3	4
3. I was more focused on managing the interview than on the rating criteria	1	2	3	4
4. I felt comfortable in the dual role of interviewer and rater	1	2	3	4
5. I felt happy about the test procedure	1	2	3	4
6. The student produced a large enough speech sample for assessment	1	2	3	4
7. It was easy to assess how well the candidate was interacting	1	2	3	4
8. I understood what I was assessing in giving the global mark	1	2	3	4
9. I understood what I was assessing in giving the analytic score	1	2	3	4
10. The most important part of my assessment in giving the global mark was grammatical accuracy	1	2	3	4
11. The most important part of my assessment in giving the detailed score was grammatical accuracy	1	2	3	4
12. I think I awarded the student a fair mark in giving the global mark	1	2	3	4
Reason:				
13. I think I awarded the student a fair mark in giving the analytic score	1	2	3	4
Reason:				
14. It was easier to mark a student who expressed an opinion similar to mine in giving the global mark	1	2	3	4
15. It was easier to mark a student who expressed an opinion similar to mine in giving the analytic score	1	2	3	4

+ Please add any other comments you would like to make about the test itself and/or your experience of the test.

Thank you for taking the time to co-operate in this project.

NAME _____

QUESTIONNAIRE 3 – STUDENT

Please fill in the questionnaire about the test you have just done. Put a **circle around the number** of the answer that most accurately corresponds to what you think about each statement. If you make a mistake or change your mind, put a cross through the incorrect answer and circle the one you have chosen.

	Strongly disagree	Disagree	Agree	Strongly agree
1. I felt nervous throughout the whole test	1	2	3	4
2. I think I did well in the test	1	2	3	4
3. I performed to the best of my ability in the test	1	2	3	4
4. I think I spoke enough for the examiner to judge my ability	1	2	3	4
5. I felt comfortable with the procedure of the test ...	1	2	3	4
6. I knew exactly what I had to do	1	2	3	4
7. The test was similar to the kind of task practised in class	1	2	3	4
8. I could answer the questions without difficulty	1	2	3	4
9. I had enough to say about the topic	1	2	3	4
10. I understand what my mark means	1	2	3	4
11. I know what I need to do in order to improve my speaking	1	2	3	4
12. I think that my general self-assessment was a true reflection of my speaking ability in English	1	2	3	4
13. I think that my self-assessment in the group oral test was a true reflection of my speaking ability in English	1	2	3	4
14. I think self assessment can play a useful role in learning generally	1	2	3	4
15. I think my self-assessment should be taken into consideration in my overall grade for the subject Lengua BII	1	2	3	4
16. We should be given the opportunity to use self-assessment more frequently in this subject	1	2	3	4
17. We should be trained in how to assess our language skills in this subject	1	2	3	4

Please add any other comments you would like to make about the test itself and/or your experience of the test in the box below.

Thank you for taking the time to co-operate in this project.

QUESTIONNAIRE 4 – INTERVIEWER

Please fill in the questionnaire about the test you have just done. Put a **circle around the number** of the answer that most accurately corresponds to what you think about the statement. If you make a mistake or change your mind, put a cross through the incorrect answer and circle the one you have chosen.

	Strongly disagree	Disagree	Agree	Strongly agree
1. I was able to manage the interview and give each student a score at the end of the test using the rating scale provided	1	2	3	4
2. I was more focused on managing the interview than on the rating criteria	1	2	3	4
3. I felt comfortable with the test procedure	1	2	3	4
4. The students produced a large enough speech sample for assessment	1	2	3	4
5. It was difficult to manage the test with three students participating	1	2	3	4
6. I felt comfortable in the dual role of interviewer and global rater	1	2	3	4
7. I knew what features to focus on while assessing the candidates	1	2	3	4
8. It was easy to assess how well the candidates were interacting	1	2	3	4
9. It was useful to have a rating scale to refer to when giving the global score	1	2	3	4
10. It was easier to assess students who expressed an opinion similar to mine on the topic	1	2	3	4
11. It was easier to use a scale from 0-5 than one from 1-10	1	2	3	4
12. It was easier to assign meaning to a scale of 0-5 than to one of 1-10	1	2	3	4
13. I think that I awarded the students a fair score	1	2	3	4
Reason:				
14. I think that students can give a true reflection of their general speaking ability using the criteria provided	1	2	3	4
15. I think that students can give a true reflection of their performance in the group oral test using the criteria provided	1	2	3	4
16. Self-assessment is a useful tool for helping students to know how improve their speaking ability in English	1	2	3	4
17. Self-assessment can play a useful role in learning generally	1	2	3	4

18. Self-assessment should be taken into consideration in the students' overall mark for English Language subjects at the ULPGC	1	2	3	4
---	---	---	---	---

Please add any other comments you would like to make about the test itself and/or your experience of the test.

Thank you for taking the time to co-operate in this project.

Appendix 6

Instructions to examiners

INDIVIDUAL ORAL PROFICIENCY INTERVIEW: PROCEDURE

Each interview should last between 5 and 6 minutes and has two parts:

PART 1: Settling in phase (1 minute)

The candidate is invited to give the Interviewer some personal information, according to the suggested questions in the materials pack.

PART 2: Interview on text topic (4-5 minutes)

Immediately before the interview the candidate will have been provided with a short text on a topic of relative controversy and which s/he will have read before entering the interview room. In the second part of the interview, the Interviewer will ask the candidate some questions about the topic of the text, as provided in the materials pack. (It is important to note that reading comprehension is not being tested here, and the candidate should not be required to explain any points of the text itself. Its purpose is principally as a springboard for the topic of discussion).

The interviewer does not need to ask all the questions provided in the pack as long as the candidate is interacting and there is exchange of information. The direction of the conversation can be followed naturally, where this is appropriate, rather than returning to the prescribed questions.

The Interviewer and Rater should exchange roles at intervals throughout the examining session. The Interviewer should introduce the Rater at the start of the test, saying that s/he will just listen to the interview. At no point during an interview should the Rater become involved in the interaction.

At the end of the interview, outside the interview room, the candidates will be provided with a self-assessment sheet on which they should fill in their own perception of how they performed on the test. It is very important that they complete this and hand it in before leaving.

INSTRUCTIONS FOR INTERVIEWER

1. It is important to keep as closely as possible to the time frame, so make sure you check your watch at the start of the interview.
2. Ask only one or two (as necessary) of the questions for the settling-in phase and avoid carrying on a longer conversation.
3. Make a smooth transition to the topic of the text and ask the first question from the materials pack (each numbered pack contains copies of the text and related questions). You may follow the natural course of the conversation and do not need to use all the questions if it is not necessary.
4. Draw the interview to a close within the stipulated time, thank the candidate and say good-bye. Do not give any indication of how well the candidate has performed, and avoid using words like *good* which might imply a judgement on their performance. You can substitute them by *OK, alright, thank you* etc. which are encouraging but neutral. Retrieve the text from the candidate.
5. At the end of the interview, after the candidate has left the room, give a global impression mark of their performance on the University scale of 1-10. The rater will record this on the candidate's mark sheet. **Do not fill in the mark yourself.**
6. Then consider the analytic rating scale and give the candidate a score in each of the categories. The rater will record these scores on the candidate's mark sheet. **Do not discuss the marks with the Rater.**

INSTRUCTIONS FOR RATER

1. At the beginning of each interview, start the tape and say the candidate's name. Stop the tape at the end of the interview.
2. As you listen to the interview, consider the categories and scores on the rating scale, and fill in a score for the candidate in each category.
3. Fill in the number of the test pack used, and the names of the Interviewer and the Rater.
4. Record the global impression mark and the analytic scores given to you by the Interviewer. **Do not discuss these marks.**

GROUP SPEAKING TEST: PROCEDURE

Candidates will be examined in groups of three. Each group oral test should last between 15 and 18 minutes and has two parts:

PART 1: Settling in phase – addressed to individual candidates (1 minute)

Each candidate is invited to give the Interviewer some personal information, according to the suggested questions in the materials pack. The aim here is to create a more relaxed atmosphere and to boost the candidates' confidence.

PART 2: Candidate interaction on text topic (12-15 minutes)

Immediately before the test, the candidates will have been provided with a short text on a topic of relative controversy which they will have read and prepared together immediately before entering the interview room. All candidates are given the same text. (On handing the texts to the candidates, remember to indicate that they should not write on them, roll them up or fold them, or deface them in any other way). In the second part of the test, the Interviewer will give the candidates some written questions about the topic of the text, as provided in the materials pack. (It is important to note that reading comprehension is not being tested here, and the candidates are not required to explain any points of the text itself. Its purpose is principally as a springboard for the topic of discussion).

The Interviewer will invite the candidates to discuss the questions on the sheet amongst themselves without further intervention.

The Interviewer and Rater should exchange roles at intervals throughout the examining session. The Interviewer should introduce the Rater at the start of the test, saying that s/he will just listen to the interview. At no point during an interview should the Rater become involved in the interaction.

At the end of the interview, outside the interview room, the candidates will be provided with a self-assessment sheet on which they should fill in their own perception of how they performed on the test. It is very important that they complete this and hand it in before leaving.

INSTRUCTIONS FOR INTERVIEWER

1. It is important to keep as closely as possible to the time frame, so make sure you check your watch at the start of the interview.
2. Ask each candidate only one or two (as necessary) of the questions for the settling-in phase and avoid carrying on a longer conversation.
3. After all three candidates have had an individual turn, make a smooth transition to the topic of the text. (Suggested rubrics: “You read a text about ... (*topic of text*). Now I’m going to give you some questions about the topic of the text and I would like you to talk **to each other** about them”). Hand the candidates one copy each of the question sheets from the materials pack (each numbered pack contains copies of the text and related questions). The candidates should be given a minute to read the questions and then invited to start if they do not do so spontaneously. At this point, it is often helpful to withdraw eye-contact to avoid the temptation for candidates to address the Interviewer rather than the other candidates in the group.
4. The group should be allowed to follow the natural course of the conversation and it does not matter if they do not use all the questions as long as interaction is taking place.
5. In the event of the interaction coming to a premature end, or one of the candidates failing to produce enough speech for a confident assessment, a further question is provided in the Interviewer version of the test in the materials pack. If this question is used it will signify the end of the group interaction, since the Interviewer will immediately become the focus of attention for all the candidates, no matter who the question is addressed to, so do not use it unless it is absolutely necessary.
6. Draw the interaction to a close within the stipulated time, even if this means interrupting the group’s conversation at a convenient point, thank the candidates and say good-bye. Do not give any indication of how well the test has been carried out, and avoid using words like *good* which might imply a judgement on the candidates’ performance. You can substitute them by *OK*, *alright*, *thank you* etc. which are encouraging but neutral. Retrieve the texts and question sheets from the candidates.
7. At the end of the interview, after the candidates have left the room, give a global impression mark of their individual performance on the Analytic Rating Scale of 0-5 provided. The rater will record this on the candidates’ mark sheets. **Do not fill in the mark yourself. Do not** discuss the marks with the Rater.

INSTRUCTIONS FOR RATER

1. At the beginning of each interview, start the video apparatus. Stop the video at the end of the interview.
2. As you listen to the interaction, consider the categories and scores on the rating scale, and fill in a score for the each candidate in all the categories.
3. Fill in the number of the test pack used, the names of the other candidates in the group, and the names of the Interviewer and the Rater.
4. Record the global impression mark given to you by the Interviewer for each candidate. **Do not** discuss the marks.

Appendix 7

ARELS Marking Key for the Higher Certificate Examination in Spoken English and Comprehension



MARKING KEY
for the
HIGHER CERTIFICATE
EXAMINATION
in
SPOKEN ENGLISH
and
COMPREHENSION

AH69

© OXFORD DELEGACY OF LOCAL EXAMINATIONS 1996

It is an infringement of copyright to reproduce by any method the whole or part of this booklet, or of the tape on which the examination is recorded.

ACKNOWLEDGEMENTS

The UODLE acknowledges the British Broadcasting Corporation for the items used in SECTION FOUR.

MARKING SCHEME: AH69

This KEY is primarily for the use of examiners when marking candidate tapes. A printed text of the examination may be obtained from the Oxford Delegacy of Local Examinations.

At this level every candidate tape is marked at least twice.

NOTES FOR MARKERS

1. The model answers offered in each section are often not definitive. Markers should use their own judgement in scoring each item, bearing in mind the response the setter expected.
2. A foreign accent is not to be penalised unless it interferes with comprehension.
3. Non-British English is acceptable.
4. Please circle or cross through the score for each question. Include half marks in section totals on back page. Ignore any half marks in final (/200) total on back page. Raise any half marks in percentage total to the next whole number.

Add marks very carefully: it is not fair to misgrade candidates just because you can't add.

5. Any relevant comments on the candidate's performance will be read with care and gratitude. If your impression mark is significantly different from the grade indicated by the percentage total, please say why you think this is.
6. The co-ordinator uses the following marks in deciding which grade to award:
 - Fail: under 55% or two failed sections (under 50%)
 - Pass: 55% – 69%
 - Credit: 70% – 79%
 - Distinction: 80% and over

These criteria can sometimes work in an over-arbitrary way, so markers are asked to give an impression mark of grade, independent of percentage total. Please use the general guidelines of exam standard to help your assessment, and mark your impression before adding up your totals.

7. General guidelines of exam standard:
 - Pass: The candidate can communicate and understand effectively. He could manage everyday situations in an English-speaking environment quite well, although with occasional difficulties. More complex situations would still prove troublesome most of the time.
 - Credit: The candidate should manage well in an English-speaking environment. Everyday situations should present no problems. He has the confidence to tackle complex situations, although these might still give trouble.
 - Distinction: Everyday situations should give no trouble at all. The candidate should manage more complex situations with confidence and fair success. He should hold his own easily in an English-speaking environment.

MARKING GUIDE

SECTION ONE

The topics are given to the candidates after the introductory remarks and eight minutes is given for preparation and note making. Candidates should have neither the time nor the space on the Topic Slip to make extensive notes, but if you feel a candidate is reading, please note this on the back page.

The following criteria should be used in assessment:

1. Holding the listener's attention (by the interest and relevance of what the candidate has to say and the skill with which he says it)

0 1 2 3 4 5 6 7 8 9 10 11 12

2. Fluency

0 1 2 3 4 5 6 7 8 9 10 11 12

3. Accuracy of all aspects of the candidate's English

0 1 2 3 4 5 6

PLEASE NOTE NUMBER OF TOPIC CHOSEN ON BACK PAGES

TOTAL	/ 30
-------	------

SECTION TWO

Use these guidelines in assessing candidates' responses. Note that appropriateness of response subsumes appropriateness of register.

- 0: Candidate fails to respond. Candidate's response is likely to be misunderstood / misinterpreted by an average native speaker.
- 1: A response, no matter how inaccurately produced, that makes the candidate understood.
- 2: A response that is comprehensible and reasonably appropriate, although with quite serious faults.
- 3: Appropriate, comprehensible, and unambiguous; there may be faults in several aspects of production, but these will not be serious.
- 4: Appropriate, readily comprehensible, unambiguous; only exceptionally minor faults.

The following suggested responses are only illustrative.

Part One

- | | | | | | | |
|----|--|---|---|---|---|---|
| 1. | <i>Oh, really? Why not? What's happening?</i> | 0 | 1 | 2 | 3 | 4 |
| 2. | <i>Oh, thanks very much. Happy birthday.</i> | 0 | 1 | 2 | 3 | 4 |
| 3. | <i>Yes, of course. Anything special you'd like?</i> | 0 | 1 | 2 | 3 | 4 |
| 4. | <i>Well, you can't leave it. You'd better get a plumber to have a look at it.</i> | 0 | 1 | 2 | 3 | 4 |
| 5. | <i>Oh, no, you should have one if you can. It doesn't do any harm and it could save you from catching the flu.</i> | 0 | 1 | 2 | 3 | 4 |
| 6. | <i>Yes, mine's a bit salty too. Let's change it. / No, mine's all right.</i> | 0 | 1 | 2 | 3 | 4 |

Sub-total	/24
-----------	-----

Part Two

7. *Well, OK, but we'd better not go too far. It looks as if it's going to rain pretty soon.* 0 1 2 3 4
8. *Excuse me, but I've lost a library book. I left it on a bus. What should I do?* 0 1 2 3 4
9. *Oh, you've finished it at last. It looks really great.* 0 1 2 3 4
10. *Excuse me, I'd like to extend my stay for a few more days, if that's possible, or is the room already booked?* 0 1 2 3 4
11. *Excuse me, can you tell me where I can find the butter now?* 0 1 2 3 4
12. *I'm sorry, but Peter White is off sick. Can I help you?* 0 1 2 3 4
13. *Good Heavens! Roger! I hardly recognised you. What's happened? You look so different.* 0 1 2 3 4
14. *Well, I might be. Is the second edition very different? Are there many changes?* 0 1 2 3 4
15. *Hello. Congratulations! You're looking very well. How's the baby? Here's a few flowers for you.* 0 1 2 3 4
16. *That was fantastic! Really delicious. Thanks very much. Let me help you with the dishes. That's the least I can do.* 0 1 2 3 4
17. *Oh, could you give her a message, please. It's Jan and I was supposed to be meeting her for lunch on Saturday, but I have to go to a conference this weekend, so I'm afraid I can't make it. Tell her I'm sorry and I'll ring her next week.* 0 1 2 3 4
18. *Excuse me, could you take a photo of us all with my camera?* 0 1 2 3 4
19. *I got this postcard from a friend of mine who's in the USA on holiday, but I can't read his writing. Do you think you could tell me what he's written?* 0 1 2 3 4
20. *Excuse me, (madam). Would you mind very much if I took a photo of your cottage. It's so beautiful.* 0 1 2 3 4

Sub-total	/56
-----------	-----

TOTAL	/80	Divide by two	TOTAL	/40
-------	-----	------------------	-------	-----

SECTION THREE

Use these guidelines in assessing candidates' responses.

- 0 1 0: Anything that is likely to be misunderstood by an average native speaker. Any word with a faulty stress pattern.
- 1: Clear enough to be readily and unambiguously understood, but may be faulty in incidental respects.
- 0 1 2 0: Anything that is likely to be misunderstood by an average native speaker. Anything that is basically un-English.
- 1: Clear, comprehensible and unambiguous, the candidate makes a reasonable approximation to native patterns.
- 2: The candidate's accent may be faulty, but otherwise there is a very good approach to native patterns.

1.	<i>You were saying the other night</i>	rhythm and catenations	0	1	2
2.	<i>apparently</i>	pronunciation	0	1	
3.	<i>a driving licence</i>	compound noun – one stress on driving	0	1	
4.	<i>because it hasn't got a photo on it</i>	rhythm and catenations	0	1	2
5.	<i>could have borrowed</i>	weak form of "have"	0	1	
6.	<i>I've come across a possible answer</i>	rhythm and catenations	0	1	2
7.	<i>You can get an official card now</i>	rhythm and catenations	0	1	2
8.	<i>and, what's more, it's free.</i>	phrasing	0	1	
9.	<i>Issuing Authority</i>	pronunciation	0	1	2
10.	<i>ADMAIL 173</i>	pronunciation	0	1	
11.	<i>London WIE 2SJ</i>	pronunciation	0	1	
12.	<i>There's no need to write anything down</i>	rhythm and catenations	0	1	
13.	<i>All you have to do then is ..</i>	phrasing	0	1	
14.	<i>1. Fill in the form .. etc</i>	Change of tone for reading list	0	1	
15.	<i>two recent passport-sized photographs</i>	phrasing and pronunciation	0	1	2
16.	<i>Take your birth certificate ...</i>	no fall until the end of the sentence	0	1	2
17.	<i>a doctor, teacher, ... similar standing</i>	list intonation – no fall	0	1	2
18.	<i>MP, JP</i>	pronunciation	0	1	
19.	<i>or a person of similar standing</i>	rhythm and catenations	0	1	2

20. *authorise* pronunciation 0 1
21. *It does, doesn't it?* fall 0 1 2
22. *prayers* pronunciation 0 1
23. *Overall impression* 0 1 2 3 4 5 6 7 8

0: Unable to cope; frequently incomprehensible and too slow to finish; speech lacks natural flow or rhythm.

2: Generally comprehensible, but usually unauthentic in delivery.

4: Almost always comprehensible, but with frequent examples of unauthentic rhythm, catenation, or intonation.

6: Always comprehensible; only occasionally unauthentic in delivery.

8: Except for accent, a very good approach to native patterns of English speech.

(Markers may, of course, give 1, 3, 5, or 7 marks for overall impression.)

TOTAL	/40	Divide by two	TOTAL	/20
-------	-----	------------------	-------	-----

SECTION FOUR

The criterion here is the candidate's understanding of the listening material and questions, not the accuracy of his responses, and marks for each question should be given on this basis.

Use these guidelines in assessing candidates' responses:

- 0: The candidate has clearly understood little or nothing.
- 1: The candidate has understood the main points, but may have missed some details. The main point of the question has been understood.
- 2: (where allocated) The candidate has missed no relevant details of both material and question.

The following answers are only illustrative.

The contents of this section could not be reproduced for copyright reasons.

SECTION FIVE

This is a test of fluency and accuracy in extended speech. Mark for each of the three criteria over the whole section. Definitions of a range of marks are given, but the marker can, of course, give 1, 3, 5, 7, or 9 marks.

The candidate is not required to produce dialogue. Good dialogue should be credited in Category 2, but marks should not be awarded just for the attempt to include it.

1. Pronunciation, stress, rhythm and intonation

	0	1	2	3	4	5	6	7	8	9	10
0	Unintelligible										
2	Poor pronunciation and intonation patterns										
4	Fair control										
6	Very few errors, but hesitant										
8	Accurate control of pronunciation, stress and intonation										
10	Fluent and with natural pace										

2. Appropriate and varied use of vocabulary and dialogue

	0	1	2	3	4	5	6	7	8	9	10
0	Unintelligible										
2	Extremely elementary										
4	Elementary and repetitive										
6	Fair control										
8	Varied and appropriate										
10	Good control; variety in range and style										

3. Appropriate and varied use of structure

	0	1	2	3	4	5	6	7	8	9	10
0	Unintelligible										
2	So foreign as to make it difficult to understand										
4	Inaccurate										
6	Hesitant but generally accurate										
8	Reasonable range and command of structures; very few inaccuracies										
10	Good range and fluent command of structures used; very few inaccuracies										

TOTAL	/30
-------	-----

SECTION SIX

Each question is designed to test accuracy and control of an area of syntax or vocabulary. Please try to ignore mistakes that seem incidental to the points being tested. Markers may award marks for correct responses which the candidate may give, but which are not given below, provided they make sense in context.

- | | | | |
|----|----|--|-------|
| 1. | c. | <i>"This man must be caught," (said the governor.)</i> | 0 1 |
| | d. | <i>Jack was said to be a dangerous criminal.</i> | 0 1 |
| | e. | <i>This was the fourth time (1) Jack had been in prison. (1)</i> | 0 1 2 |
| | f. | <i>The police used a helicopter (1) to (try to) find him. (1)</i> | 0 1 2 |
| | g. | <i>If the van hadn't had a puncture, (1) the (van) driver wouldn't have been out late. (1)</i> | 0 1 2 |
| | h. | <i>Jack was taken back (1) to (the) prison (in the van). (1)</i> | 0 1 2 |
| | i. | <i>Jack wished he hadn't tried to escape.</i> | 0 1 |

Sub total	/11
-----------	-----

- | | | | |
|----|----|-----------------------------|-----|
| 2. | d. | <i>Yes, he has.</i> | 0 1 |
| | e. | <i>No, he couldn't.</i> | 0 1 |
| | f. | <i>Yes, it must.</i> | 0 1 |
| | g. | <i>Yes, we did.</i> | 0 1 |
| | h. | <i>No, you/ we don't.</i> | 0 1 |
| | i. | <i>Yes, we had.</i> | 0 1 |
| | j. | <i>Yes, we shall/ will.</i> | 0 1 |

Sub total	/7
-----------	----

- | | | | |
|----|----|------------------------------------|-----|
| 3. | d. | <i>... looking for Jack.</i> | 0 1 |
| | e. | <i>... to look/ look for Jack.</i> | 0 1 |
| | f. | <i>... looking for Jack.</i> | 0 1 |
| | g. | <i>... look for Jack.</i> | 0 1 |
| | h. | <i>... to look for Jack.</i> | 0 1 |

- i. ... to look for Jack. 0 1
- j. ... to looking for Jack. 0 1

Sub total	/ 7
-----------	-----

- 4. d. Yes, he was rather dirty/ muddy. 0 1
- e. Yes, he was rather tired. 0 1
- f. Yes, he was rather hungry. 0 1
- g. Yes, they were rather loud/ noisy. 0 1
- h. Yes, he was rather pleased/ happy. 0 1
- i. Yes, it is/ was rather funny/ amusing/ comical. 0 1
- j. Yes, he was rather angry/ cross. 0 1

Sub total	/ 7
-----------	-----

5. These answers are only illustrative and other responses may be appropriate in context.

- d. Well, open a window. 0 1
- e. Well, take a pill. 0 1
- f. Well, take it off. 0 1
- g. Well, close it. 0 1
- h. Well, clean them. 0 1
- i. Well, water it. 0 1
- j. Well, feed them. 0 1
- k. Well, make some fresh. 0 1

Sub total	/ 8
-----------	-----

TOTAL	/ 40
-------	------

AH69

Candidate's Number:-	MARKER' IMPRESSION – Complete this section before totalling marks.							
	FAIL		PASS		CREDIT		DISTINCTION	
	Clear		Narrow		Plain		Narrow	
Marker's Number:-	Narrow		Clear		Good		Clear	
			Good					

SECTION	POSSIBLE	MARKS	PASS or FAIL	COMMENTS	
One	30			Topic No.	
Two	40				
Three	20				
Four	40				
Five	30				
Six	40				
TOTAL	200				
%	100			TEAM LEADER'S FINAL ASSESSMENT	
Number of Sections failed (under 50%)				If third hearing required insert THIRD	

**UNIVERSITY OF OXFORD
DELEGACY OF LOCAL EXAMINATIONS**



AH69

CONSOLE OPERATOR'S SCRIPT

**The Text and Editing Instructions for the
ARELS HIGHER EXAMINATION
IN
SPOKEN ENGLISH AND COMPREHENSION**

Schedule and Opex Number: AH69

AH69 CONSOLE OPERATOR'S SCRIPT

KEY

◆ = Cue IN (ie start candidate recording tape)

😊 = Cue OUT (ie stop candidate recording tape)

◆ **OXFORD ARELS EXAMINATION: AH69** 😊 (MUSIC)

This is an OXFORD-ARELS Examination in spoken English at Higher Level, produced by the UNIVERSITY OF OXFORD DELEGACY OF LOCAL EXAMINATIONS.

In this examination your answers are recorded on a tape or cassette, which is then marked just like a written examination paper. It's important, therefore, that your recording is good and clear. Don't speak too loudly, or too close to the microphone. Just relax and speak normally.

Your answer tape usually stops while you're listening to questions or doing examples and starts only when we want to record your answers. Don't worry if sometimes you don't speak for the whole of the time allowed, or if you sometimes can't say all you'd like to. In general, we're testing the quality of your English, not how much you can say.

If something goes wrong during the examination, raise your hand and someone will help you. Before we start, we're going to stop the tape for a moment to check that everything is all right.

STOP MASTER TAPE TO CHECK CANDIDATES ARE READY TO BEGIN THE EXAMINATION.

Good. The first thing we want you to say is your candidate number. It's on the tape box and your admission slip. ◆ Say your candidate number now. (FIVE SECONDS) Good. 😊

SECTION ONE

Look at your TOPIC SLIP. We're going to stop the master tape for eight minutes while you prepare a two-minute talk on one of the five topics on your paper. You may write short notes in the space below the topics to help you in your talk. You will be allowed to look at these notes while you're talking, but don't attempt to write out your talk in full and read it. Note, too, that you'll be marked more on the persuasiveness and interest of your talk than on the accuracy of your grammar and pronunciation. Prepare your talk now.

STOP THE MASTER TAPE FOR EIGHT MINUTES

That is the end of your preparation time. Now you have two minutes to give your talk. Don't read out the topic; just say the number and begin. ♦ Start now. (TWO MINUTES) Thank you. THAT IS THE END OF SECTION ONE.

SECTION TWO 😊

In this section we test your ability to use the everyday language of common situations.

PART ONE

First you'll hear six remarks which might be made to you in various situations when you're using your English. Some are questions and some are comments. After each one, reply in a natural way. Here's an example to help you.

- Sorry to keep you waiting.
- That's all right.

Now, are you ready? Here's the first.

- ♦ 1. Oh, by the way, I shan't be coming in to work on Monday morning.
(SEVEN SECONDS)
- 2. It's my birthday today. Would you like a bit of my cake?
(SEVEN SECONDS)
- 3. You couldn't get me a coffee and a sandwich, could you? I'm not going to have time for lunch.
(SEVEN SECONDS)
- 4. That radiator in the kitchen there keeps dripping all the time. I can't stop it.
(EIGHT SECONDS)
- 5. I hear the company has arranged flu jabs for all the staff. I'm not at all sure I want an anti-flu injection, actually.
(EIGHT SECONDS)
- 6. Gosh, this soup tastes awfully salty. What's yours like?
(SIX SECONDS)

PART TWO 😊

Now you'll hear fourteen situations in which you might find yourself. Say what it seems natural to say in each situation. Ready?

7. One afternoon your friend suggests going for a walk. The sun's shining, but there are a lot of dark rain-clouds building up. ♦ What do you say? (EIGHT SECONDS)
8. 😊 Last week you left your bag on a bus and haven't been able to get it back. Unfortunately, there was a library book in it. You go to the library. ♦ What do you say to the assistant? (EIGHT SECONDS)
9. 😊 Some friends have been redecorating their flat for several weeks. One day you go to see them and it's all been done. ♦ What do you say? (EIGHT SECONDS)
10. 😊 You booked a three-day stay in a hotel. At the end of this time you decide you'd like to stay a few days longer. ♦ What do you say to the receptionist? (EIGHT SECONDS)
11. 😊 In the local supermarket you're trying to find some butter, but it isn't where it used to be. ♦ You find an assistant. What do you say? (SIX SECONDS)
12. 😊 Your colleague, Peter White, is sick, and will be off work until next Monday. You answer the phone and a woman says: "Oh, hello. Could I speak to Peter White, please?" ♦ What do you say? (TEN SECONDS)
13. 😊 You meet an old friend, Roger, on the street. He's always worn very casual clothes and had long, straggly hair and a beard. Now he's wearing a dark suit, he's clean-shaven, and has a short, rather conservative haircut. ♦ What do you say? (EIGHT SECONDS)
14. 😊 In a bookshop you find a copy of Smith's English Grammar at £12. When you take it to the assistant, he says: "This is the second edition, actually. We still have a few copies of the first edition, which we're selling for only £7, if you're interested." ♦ What do you say? (TEN SECONDS)
15. 😊 A friend recently had a baby, so you go to the hospital with a card and some flowers. ♦ What do you say when you see her? (EIGHT SECONDS)
16. 😊 You've just had a very good lunch at a friend's flat. She's obviously made a special effort. ♦ What do you say at the end of the meal? (EIGHT SECONDS)
17. 😊 You've arranged to meet your friend Anna for lunch on Saturday. But today your boss says you have to go to a conference this weekend. You ring Anna's house. Her mother answers the phone and says Anna's out. ♦ What do you say? (TEN SECONDS)
18. 😊 You're in Piccadilly Circus in London with some friends. You have your camera with you and you want somebody to take a photo of the whole group, including yourself. There's a

man standing nearby. ◆ What do you ask him? (EIGHT SECONDS)

19. 😊 You get a postcard from an English friend who's on holiday. Unfortunately, his handwriting is pretty bad and you can't read it. You show the card to an English colleague. ◆

What do you say? (NINE SECONDS)

20. 😊 You're walking through an old English village and see a really beautiful old English cottage with a lovely garden. You'd like to take a photo of it, but there's a woman working in the front garden. ◆ What do you ask her? (TEN SECONDS)

THAT IS THE END OF SECTION TWO.

SECTION THREE 😊

In this section we test your intonation, stress, rhythm, pronunciation and other details of the way you speak. Open your booklet and look at your Reading Passage.

You're ringing your friend Alison, who is 22, but looks about 16. This has caused her problems, but you have found a possible solution for her. You'll hear Alison's voice on the tape, and you must read the part marked CANDIDATE. You have two minutes to study the passage before you start reading. You may write on it if you like. Remember, you'll have to read the part marked CANDIDATE. Study the passage now for two minutes. Do not speak yet. (TWO MINUTES)

Now please be ready to read. Alison speaks first.

◆ Alison: 842341.

CANDIDATE: Hi, Alison. It's me. You were saying the other night that you'd had problems getting served in places because you look so young.

Alison: That's right. It's terrible. I can't get a drink in a bar unless they know me. They insist I'm too young.

CANDIDATE: And apparently they won't accept a driving licence as proof of age because it hasn't got a photo on it. They say you could have borrowed or even stolen someone else's licence.

Alison: That's right. It's a real pain, I can tell you.

CANDIDATE: OK. Well, I've come across a possible answer to this whole problem. You can get an official card now, which proves your age. And it has a photo on it, too.

Alison: What? You mean it's some kind of identity card?

CANDIDATE: Yes, but it's only to show proof of your age. Nothing else. It isn't a credit card

or a bank guarantee card or a passport. But it's official and, what's more, it's free.

Alison: You're kidding! You mean it really costs nothing?

CANDIDATE: That's right. The cards are issued by something called the Portman Group Issuing Authority, and their address is ADMAIL 173, London W1E 2SJ.

Alison: Hold on a minute. Let me get a pen.

CANDIDATE: There's no need to write anything down. I'll call round tomorrow with the leaflet and the application form. All you have to do then is what it says here in the leaflet.

1. Fill in the form with your personal details.
2. Obtain two recent passport-sized photographs (taken full face and without a hat).
3. Take your birth certificate or passport with the form and the photos to a doctor, teacher, lawyer, bank manager, MP, JP or a person of similar standing, and get them to authorise the form and endorse the back of the photos.
4. Post the form and the photos off to the Authority and you get your card.

Alison: It sounds too good to be true.

CANDIDATE: It does, doesn't it? The answer to all your prayers. I'll be round tomorrow with everything you need.

Alison: Marvellous. Thanks very much. See you tomorrow.

THAT IS THE END OF SECTION THREE.

SECTION FOUR 😊

In this section we test your understanding of what you hear.

The contents of this section could not be reproduced for copyright reasons.


SECTION FIVE

In this section we test how well you can speak freely. You'll be marked on the accuracy of your English; the words you use and the way you speak, for example, the intonation and rhythm of your speech.


Please look at your pictures. Jack was sent to Exmoor Prison for ten years. After three years he managed to dig a tunnel under the walls and one cold, wet night he escaped from the prison onto the moors. He thought he was finally safe when he got into the back of a van but, unfortunately for Jack, the van driver was making a delivery to the last place Jack wanted to go.

We want you to tell the story. You'll have two minutes to think about what you're going to say, then two minutes to tell the story in your own words. You must tell the story in the past, and start with the sentence above the pictures: "Three years ago Jack was sent to Exmoor Prison for ten years." Think about what you're going to say now. Do not speak yet.

(TWO MINUTES)

Now please be ready to start. Remember to tell the story in the past and to start with the sentence "Three years ago Jack was sent to Exmoor Prison for ten years."  Start now.

(TWO MINUTES)

Thank you. 

THAT IS THE END OF SECTION FIVE

SECTION SIX

In this section we're interested in how accurately you can speak. We're going to ask you some questions to test your grammar and control of words. Listen carefully to the instructions and examples for each question.

1. Look at your pictures again. You'll hear some sentences about the story. Then you'll hear someone start each sentence in a different way. You must finish it so that it means the same thing. Listen to these examples.

a. Jack's escape took place at 9 p.m..

It was ...

It was 9 p.m. when Jack escaped.

b. It took him six months to dig the tunnel.

Digging ...

Digging the tunnel took him six months.

Now you do the same. Do the examples first for practice. Your answers to these and other practice questions will not be recorded.

a. Jack's escape took place at 9 p.m..

It was ...

(FIVE SECONDS)

It took him six months to dig the tunnel.

Digging ...

(FIVE SECONDS)

◆ c. "We must catch this man," said the governor.

"This man ..."

(SEVEN SECONDS)

d. People said that Jack was a dangerous criminal.

Jack was said ...

(SIX SECONDS)

e. Jack had been in prison three times before.

This was the ...

(SEVEN SECONDS)

f. The police tried to find him using a helicopter.

The police used ...

(SIX SECONDS)

g. The van driver was out late because the van had a puncture.

If the van ...

(EIGHT SECONDS)

h. The van took Jack back to the prison.

Jack ...

(SEVEN SECONDS)

i. Jack was sorry he'd tried to escape.

Jack wished ...

(SIX SECONDS)

2. 😊 The governor of the prison is talking to one of the guards about Jack's escape. Listen to how the guard agrees with everything he says.

a. I hear the man dug a tunnel under the wall.

Yes, he did.

b. We really can't let this sort of thing happen.

No, we can't.

c. It makes us all look stupid.

Yes, it does.

Now you do the same. Agree with everything the governor says in the same way. Do the examples first for practice.

a. I hear the man dug a tunnel under the wall. (THREE SECONDS)

b. We really can't let this sort of thing happen. (THREE SECONDS)

c. It makes us all look stupid. (THREE SECONDS)

◆ d. Apparently, he's been digging for months. (FIVE SECONDS)

e. He couldn't have done it if the guards had been doing their job. (FIVE SECONDS)

f. Our security must be terrible. (FIVE SECONDS)

g. We had a bit of luck getting him back, though. (FIVE SECONDS)

h. You never know what's going to happen. (FIVE SECONDS)

i. We'd better hold a full investigation. (FIVE SECONDS)

j. We'll make sure it doesn't happen again. (FIVE SECONDS)

3. 😊 Listen to this reporter. He is talking to a colleague about Jack's escape, which is now a well-known story. Listen to how the colleague completes his sentences.

a. By 11 o'clock all the police in the county were out ...

... looking for Jack.

b. They even used dogs ...

... to look for Jack.

c. They wasted no time ...

... in looking for Jack.

Now you do the same. Complete the sentences with the correct form of "look for Jack", adding any other words necessary to make the sentence correct. Do the examples first for practice.

a. By 11 o'clock all the police in the county were out... (FOUR SECONDS)

b. They even used dogs. (FOUR SECONDS)

c. They wasted no time ... (FOUR SECONDS)

◆ d. It was midnight before they stopped ... (FIVE SECONDS)

- e. A helicopter was brought in to help ... (FIVE SECONDS)
- f. In that weather it was hardly worth ... (FIVE SECONDS)
- g. The guards would rather stay at home in the warm than... (FIVE SECONDS)
- h. But in the circumstances they couldn't refuse .. (FIVE SECONDS)
- i. As they had let him escape, they felt obliged ... (FIVE SECONDS)
- j. Though none of them was looking forward ... (FIVE SECONDS)

4. 😊 Listen to these people talking about Jack's escape. The man is excited and uses extravagant and exaggerated language. The woman is calmer and uses simpler words.

- a. I heard the man was soaked to the skin.
Yes, he was rather wet.
- b. And he was absolutely freezing to death.
Yes, he was rather cold.
- c. The guards were astonished when Jack got out of the van. (FIVE SECONDS)
Yes, they were rather surprised

Now you do the same. Reply like the woman, using simple words to agree with the man. Do the examples first for practice.

- a. I heard the man was soaked to the skin. (FIVE SECONDS)
- b. And he was absolutely freezing to death. (FIVE SECONDS)
- c. The guards were astonished when Jack got out of the van. (FIVE SECONDS)
- ◆ d. And when they found him he was absolutely filthy. (FIVE SECONDS)
- e. And, what's more, he was totally exhausted. (FIVE SECONDS)
- f. They say he was absolutely famished. (FIVE SECONDS)
- g. The alarm sirens were absolutely deafening. (FIVE SECONDS)
- h. I hear the governor was overjoyed to get him back in the prison. (FIVE SECONDS)
- i. Getting a lift in a van going back into the prison - it's absolutely hilarious.(FIVE SECONDS)
- j. I bet Jack was livid when he saw where he was. (FIVE SECONDS)

5. 😊 Now we want you to talk to your friend. Your friend is rather helpless and you are busy. you have to tell your friend what to do. Listen.

- a. The phone's ringing.

Well, answer it.

- b. The kettle's boiling.

Well, switch it off.

- c. The milk jug's empty.

Well put some more milk in it.

Now you do the same. Tell your helpless friend what to do. Do the examples first for practice.

- | | |
|---|----------------|
| a. The phone's ringing. | (FIVE SECONDS) |
| b. The kettle's boiling. | (FIVE SECONDS) |
| c. The milk jug's empty. | (FIVE SECONDS) |
| ◆ d. It's very hot and stuffy in here. | (FIVE SECONDS) |
| e. I've got a terrible headache. | (FIVE SECONDS) |
| f. I think the soup's burning. | (FIVE SECONDS) |
| g. There's an awful draught from that window. | (FIVE SECONDS) |
| h. My shoes are terribly muddy. | (FIVE SECONDS) |
| i. This plant looks awfully dry. | (FIVE SECONDS) |
| j. These goldfish look really hungry. | (FIVE SECONDS) |
| k. This tea's stone cold. | (FIVE SECONDS) |

THAT IS THE END OF SECTION SIX. 😊

It's also the end of the exam. Please stay in your place until you're asked to leave. Don't take any papers away with you. Thank you. Goodbye.

SECTION THREE: Reading Passage

You're ringing your friend Alison, who is 22, but looks about 16. This has caused her problems, but you have found a possible solution for her. You'll hear Alison's voice on the tape, and you must read the part marked CANDIDATE. You have two minutes to study the passage before you start reading. You may write on it if you like.

Alison: 842341.

CANDIDATE: Hi, Alison. It's me. You were saying the other night that you'd had problems getting served in places because you look so young.

Alison: *That's right. It's terrible. I can't get a drink in a bar unless they know me. They insist I'm too young.*

CANDIDATE: And apparently they won't accept a driving licence as proof of age because it hasn't got a photo on it. They say you could have borrowed or even stolen someone else's licence.

Alison: *That's right. It's a real pain, I can tell you.*

CANDIDATE: OK. Well, I've come across a possible answer to this whole problem. You can get an official card now, which proves your age. And it has a photo on it, too.

Alison: *What? You mean it's some kind of identity card?*

CANDIDATE: Yes, but it's only to show proof of your age. Nothing else. It isn't a credit card or a bank guarantee card or a passport. But it's official and, what's more, it's free.

Alison: *You're kidding! You mean it really costs nothing?*

CANDIDATE: That's right. The cards are issued by something called the Portman Group Issuing Authority, and their address is ADMAIL 173, London W1E 2SJ.

Alison: *Hold on a minute. Let me get a pen.*

CANDIDATE: There's no need to write anything down. I'll call round tomorrow with the leaflet and the application form. All you have to do then is what it says here in the leaflet.

1. Fill in the form with your personal details.
2. Obtain two recent passport-sized photographs (taken full face and without a hat).
3. Take your birth certificate or passport with the form and the photos to a doctor, teacher, lawyer, bank manager, MP, JP or a person of similar standing, and get them to authorise the form and endorse the back of the photos.
4. Post the form and the photos off to the Authority and you get your card.

Alison: *It sounds too good to be true.*

CANDIDATE: It does, doesn't it? The answer to all your prayers. I'll be round tomorrow with everything you need.

Alison: *Marvellous. Thanks very much. See you tomorrow.*

SECTION FIVE: PICTURE STORY A few short hours of freedom.

Three years ago Jack was sent to Exmoor Prison for ten years.



NOTES

Appendix 8

Trinity Grade Examinations in Spoken English for Speakers of Other Languages

Contents

Preface	<i>page</i> 2
Information	4
Initial stage	8
Grade 1	10
Grade 2	12
Grade 3	14
Elementary stage	16
Grade 4	18
Grade 5	20
Grade 6	22
Intermediate stage	24
Grade 7	28
Grade 8	30
Grade 9	32
Advanced stage	34
Grade 10	36
Grade 11	40
Grade 12	44
Regulations	47

Spoken English for speakers of other languages

Information



Aims and objectives

The aim of Trinity's grade examinations in spoken English is to provide a scheme of assessment against which candidates, teachers and parents may measure progress and development, whether towards professional training or as a leisure activity. The grades provide a continuous measure of professional competence for the benefit of candidates, teachers and employers.

The examinations form a series of twelve progressively graded tests, which are designed for speakers of languages other than English and which set realistic objectives in listening to and speaking with other English speakers. They move from a very low level of proficiency (Grade 1) up to an advanced level of proficiency approaching first-language ability (Grade 12). *



How to use this syllabus

The syllabus is presented in four stages. At the beginning of each stage there is an introduction which outlines the candidate profile expected by the end of the stage. These profiles are broadly related to the common reference levels proposed in draft two of the Council of Europe's Common European Framework of Reference (1996). The introduction then describes the format and procedures adopted for the examinations, sets out the assessment criteria and ends by offering guidance to teachers and candidates.

The individual syllabus for each of the three grades in the stage indicates:

- the format for the grade, including the timing
- expected candidate performance (learning outcomes)
- grammatical items
- subject areas for conversation
- assessment criteria
- sample exchanges which might take place between candidate and examiner

Learning outcomes, referred to as candidate performance in the syllabus, are specified for each grade and for the end of each stage.

The syllabus is cumulative and the outcomes for each grade assume mastery of the outcomes of the previous grades.

Trinity examiners take equal account of all internationally accepted standard varieties of English.



The four stages

At Initial stage the conversation is initiated by the examiner. Examinations at Initial stage are based on a broad definition of the first common reference level (A1/A2 Basic User, previously Breakthrough to Waystage) proposed in the Common European Framework.

At Elementary stage the examination is initiated by the candidate who makes a short presentation of a topic of his/her own choice, which naturally leads to conversation with the examiner. Examinations at Elementary stage are based on a broad definition of the second common reference level (B1 Independent User, previously Threshold) proposed in the Common European Framework.

At Intermediate stage the examination is initiated by the candidate whose presentation of a chosen text is included after the initial presentation of the topic, and these lead naturally on to discussion with the examiner. Examinations at Intermediate stage are based on a broad definition of the second common reference level (B1/B2 Independent User, previously Threshold to Vantage) proposed in the Common European Framework.

At Advanced stage the examination is initiated by the candidate who undertakes a listening comprehension task after the initial presentation of the topic and text. Examinations at Advanced stage are based on a broad definition of the third common reference level (C1/C2 Proficient User, previously Effective Operational Efficiency to Mastery) proposed in the Common European Framework.

[Turn over for a chart showing the requirements at each stage.]

<i>Initial</i>	<i>Elementary</i>	<i>Intermediate</i>	<i>Advanced</i>
<i>5-7 minutes</i>	<i>10 minutes</i>	<i>15 minutes</i>	<i>25 minutes</i>
			Conversation
		Conversation	Topic
	Conversation	Topic	Text
	Topic	Text	Listening comprehension
Conversation	Discussion	Discussion	Discussion



Assessment criteria

Assessment criteria for each grade are given under each grade heading.



Written examinations

Accompanying optional examinations which test the ability to communicate in writing will be available at three levels (Elementary, Intermediate and Advanced) from 2000. Details are available in a separate leaflet.

[Turn over for introduction to the Initial stage

Initial stage

Introduction



Candidate profile

By the end of the Initial stage the candidate can

- understand and use familiar everyday expressions and very basic phrases so as to satisfy basic needs of a concrete type relating to family, people known to the candidate, and immediate surroundings
- introduce him/herself and others
- ask and answer questions about personal details and very familiar subject areas and topics, such as where the candidate lives, people the candidate knows and items the candidate possesses
- interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

This profile is based on a broad definition of the first common reference level (A1/A2 Basic User, previously Breakthrough to Waystage) proposed in draft two of the Council of Europe's Common European Framework of Reference (1996).



Format

The conversation consists of four phases in Grades 1 and 2 and three phases in Grade 3:

- greetings and setting at ease (Grades 1, 2 and 3)
- giving instructions (Grades 1 and 2)
- conversation and questions (Grades 1, 2 and 3)
- end of conversation and leave-taking (Grades 1, 2 and 3)



Procedure

In all grades, the examiner begins by greeting the candidate and trying to set him/her at ease. The conversation is then initiated by the examiner to give the candidate the opportunity to demonstrate through both speech and actions the range of language required at this stage.

The candidate may be required to display understanding through gesture and simple actions, such as moving around the room, writing on the board or pointing to specific objects.

In Grades 1 and 2, the examiner gives some simple instructions which the candidate is expected to carry out. This is followed by questions asked by the examiner which relate to the subject areas listed for that particular grade and previous grades.

In Grade 3 it is not normal for many instructions to be given for candidates to carry out. The examiner asks questions and develops the conversation relating to the subject areas listed in Grade 3 and the previous grades.

The examiner selects materials appropriate to the age and maturity of the candidate, including everyday objects and pictures, in order to encourage the conversation to develop. The examiner may also refer to the immediate surroundings of the examination room or centre.

At all grades, the examiner brings the conversation to an end by wishing the candidate goodbye.



Assessment criteria

At each grade, the examiner will apply the following criteria:

- | | |
|----------------------|---|
| Readiness | <ul style="list-style-type: none">• the candidate's understanding of the examiner• satisfying the requirements listed under Candidate Performance for each grade (the examiner allows for hesitation and slowness of response) |
| Pronunciation | <ul style="list-style-type: none">• at all grades, production of individual sounds to form words which are intelligible• additionally at Grade 2, the use of appropriate contracted forms and the beginnings of the use of stress in short answers• additionally at Grade 3, extension of the use of stress and initial use of intonation |
| Usage | <ul style="list-style-type: none">• accuracy of grammatical items used• use of appropriate vocabulary |



Guidance

At this stage questions and answers play a major part in the conversation, but the examiner aims to enable the candidate to participate in a genuine and interesting two-way exchange within the linguistic limits set by the syllabus.

The language used is related to the expected candidate performance, to the grammatical items listed, and to the subject areas specified for each grade and, at Grades 2 and 3, for the preceding grades.

Initial stage

Grade 1



Format

The candidate holds a conversation with the examiner. *Time: 5 minutes.*

There are four phases to the conversation:

- | | |
|---------------------|--|
| Greetings | The examiner says hello and tries to put the candidate at ease. |
| Instructions | The examiner gives a few simple instructions which the candidate carries out. |
| Questions | The examiner asks a few questions related to the subject areas listed opposite. |
| Leave-taking | The examiner brings the conversation to an end by wishing the candidate goodbye. |



Candidate performance

The candidate is expected to

- exchange greetings with the examiner
- understand simple instructions and requests, and show understanding through appropriate actions or the production of appropriate spoken responses or phrases (candidates might occasionally be asked to write briefly or draw something simple on a board or on blank paper)
- give very short, even single-word, answers to simple closed questions and requests for information
- identify and name colours, parts of the body, numbers, items of clothing and objects in the immediate surroundings.



Grammatical items

- Understanding and using the present simple tense of *to be* and other common verbs such as *go, show, point, come, give, sit down, stand up*
- Imperatives (to which candidates should respond)
- Nouns in singular and plural
- Adjectives
- Articles
- Pronouns (including possessives)
- Demonstratives



Subject areas for conversation

- Personal information
- Immediate surroundings
- Clothes
- Parts of the body
- Numbers up to 20
- Colours

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

Maximum use will be made of flash cards, photos, pictures, objects in the room or other objects which the examiner may have brought.



Assessment criteria

These are set out in the introduction to the Initial stage, on page 8.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

Hello! How are you today?
 Stand up please.
 Walk to the window.
 Touch the window.
 Come back to your chair.
 Now sit down, please.
 What colour is that?
 And this one?
 How many pencils are there?
 Give me the blue pencil.
 What's this?
 Show me your ears.
 Look at me. I have ...
 a white shirt
 grey trousers
 black shoes
 blue socks.
 What about you?

Candidate

Hello! I'm fine, thank you.
(candidate stands up)
(candidate walks to the window)
(candidate touches the window)
(candidate comes back to the chair)
(candidate sits down)
 Green
 Brown
 One, two, three ...
(candidate gives examiner the blue pencil)
 (Your) nose/hand/mouth ...
(candidate indicates his/her ears)

 (I have a) red shirt
 green trousers ...

Grade 2



Format

The candidate holds a conversation with the examiner. *Time:* 6 minutes.

There are four phases to the conversation:

- Greetings** The examiner says hello and tries to put the candidate at ease.
Instructions The examiner may give a few instructions for the candidate to carry out.
Questions The examiner asks questions related to the subject areas listed below.
Leave-taking The examiner brings the conversation to an end by wishing the candidate goodbye.



Candidate performance

The candidate is expected to

- understand short questions, requests and statements, and respond with appropriate actions and short answers or statements
- contribute to the conversation using learnt phrases as necessary
- describe people, objects and places very simply
- indicate the positions of people and objects
- talk about current activities
- name the days of the week and the date



New grammatical items

- The simple present tense in questions, statements and negatives
- The use of *there is/are* and *has/have got*
- Question words—*who, what, where, why, how many*
- Possessives should be recognised
- Prepositions of place
- Determiners with countable nouns
- Introduction of the present continuous (questions and answers)

—in addition to items listed for Grade 1



New subject areas for conversation

- Home
- Family members
- Friends
- Animals, pets
- Possessions
- Daily routine and activities
- Days of the week
- Months of the year

—in addition to items listed for Grade 1

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

Maximum use will be made of objects in the examination room, as well as those that can be seen through the window, and any pictures or objects that the examiner may have brought into the room.



Assessment criteria

These are set out in the introduction to the Initial stage, on page 8.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

Hello! Where do you come from?
 Have you got any brothers and sisters?
 What's her name?
 How old is she?
 Please look at this picture.
 How many people are there in the picture?
 Where are they?
 Do you have any pets?
 Tell me about your dog.
 What are these men wearing (in the picture)?
 What's that?
 Put the red pen next to the clock.
 Where is my pencil?
 What day of the week is it?

Candidate

Hello! I come from Granada.
 Yes, I've got one sister.
 (Her name is) Elizabeth.
 She's ...
 There are two people (in the picture).
 They're in a house ...
 Yes, I have a dog and a cat.
 His name is ...
 They are wearing a white shirt and white trousers.
 A clock.
 (*candidate puts the pen next to the clock*)
 (It's) in your pocket.
 It is Monday/Tuesday ...

Grade 3



Format

The candidate holds a conversation with the examiner. *Time: 7 minutes.*

There are three phases to the conversation:

Greetings	The examiner says hello and tries to put the candidate at ease.
Questions and information exchange	The examiner asks questions, requests information and develops the conversation using the subject areas listed below and in the two previous grades.
Leave-taking	The examiner brings the conversation to an end by wishing the candidate goodbye.



Candidate performance

The candidate is expected to

- respond appropriately to simple instructions and requests
- give basic personal information, including information about and description of life and activities at work, school, college or university, at home and during free time
- give basic information about people and places including descriptions of people encountered in daily life at home, work, study and recreation, as well as descriptions of places in the candidate's home town or country
- talk about his/her daily routine, events and weather, and describe what is happening at the moment either in real life or in pictures
- ask for information on the above
- give simple directions
- tell the time



New grammatical items

- Correct use of the present simple and present continuous tenses
- Formation of simple questions using question words as necessary
- Prepositions denoting movement

—in addition to items listed for Grades 1 and 2



New subject areas for conversation

- Work
- School
- College or university
- Home life
- Weather
- Free time
- Places

—in addition to items listed for Grades 1 and 2

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

Use will be made of pictorial material in particular. This is to enable candidates to extend their range of language and to show that they can describe correctly what is currently happening and what happens regularly.



Assessment criteria

These are set out in the introduction to the Initial stage, on page 8.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

Where do you live?
 Is that near here?
 What is the weather like today?
 Is it raining now?
 What do you study at school?
 What do you do when you get home from school?
 What is the name of your best friend?
 Tell me about George.
 Let's look at this picture. What are these people doing?
 What is this man doing?
 And this lady?
 At what time do you go to bed?
 What is the time now?

Candidate

I live in Perugia.
 No, it's half an hour from here.
 It's hot and sticky.
 No, it isn't.
 I study English, maths, science ...
 I listen to some music. I do my homework. And then I have dinner.
 George.
 George is ten years old. He is tall; has brown hair...
 They're sitting in front of a café. They are drinking coffee.
 He's cleaning a window.
 She's carrying ...
 Ten o'clock.
 It's twenty to ...

Elementary stage

Introduction



Candidate profile

By the end of the Elementary stage the candidate can

- understand and use language in familiar situations related to school, work, travel and leisure
- express and ask about personal interests and give opinions on familiar or prepared topics
- talk about past, present and future events
- express hopes and intentions
- maintain the flow of communication with minimal assistance

This profile is based on a broad definition of the second common reference level (B1 Independent User, previously Threshold) proposed in draft two of the Council of Europe's Common European Framework of Reference (1996).



Format

The conversation consists of four phases in all three grades:

- greetings and setting at ease
- presentation and discussion of a prepared topic
- general conversation
- end of conversation and leave-taking



Procedure

The examiner begins by greeting the candidate and trying to set him/her at ease.

The examiner invites the candidate to present a topic that (s)he has prepared. The candidate should bring an object/objects or pictures into the discussion to help illustrate the presentation. The presentation is followed by questions and discussion led by the examiner relating to the presentation. This phase of the conversation should last approximately 5 minutes.

The examiner will then extend the discussion to the subject areas listed for the particular grade and previous grades. The examiner may also explore some of the language items listed for the particular grade and previous grades.

The candidate is encouraged to demonstrate the range of language (s)he commands and which is required at each grade of this stage. This phase of the conversation should last about 4 minutes.

The examiner brings the conversation to an end by saying goodbye to the candidate.



Assessment criteria

At each grade, the examiner will apply the following four criteria:

- | | |
|----------------------|--|
| Readiness | <ul style="list-style-type: none"> • the candidate's understanding of the examiner • maintaining the flow of the conversation through promptness of response, although short pauses will be allowed for candidates to formulate responses at Grades 4 and 5 • satisfying the requirements listed under Candidate Performance for each grade and for all previous grades |
| Pronunciation | <ul style="list-style-type: none"> • production of intelligible individual sounds, including weak forms in connected speech • satisfactory use of stress, rhythm, intonation and linkage features, including unstressed forms, so that speech sounds natural at the sentence level |
| Usage | <ul style="list-style-type: none"> • accuracy of grammatical items used • choice of appropriate vocabulary and grammatical items • range of vocabulary, grammatical items and functions used |
| Focus | <ul style="list-style-type: none"> • communication of sufficient and relevant information required by the tasks set • coherent organisation of information and opinions (Grade 6) • ability to state communicative purpose (Grades 5 and 6) |



Guidance

i Prepared topic The prepared topic may be any topic the candidate is interested in, knowledgeable about and able to talk readily about. The topic need not be chosen from the subject areas listed for the grade. The purpose is to give candidates the opportunity to display the language they know they can use.

Candidates are strongly recommended to bring into the examination one or more pictures, photos, diagrams, models or other suitable objects to illustrate the prepared topic and stimulate the conversation with the examiner. Without appropriate support materials, candidates may not give of their best and could consequently receive lower marks. However, birds, insects, reptiles or other live animals may *not* be brought into the examination room.

Candidates are advised to think carefully about the amount of material necessary for their topic, bearing in mind the time available. They should prepare enough material to sustain a presentation of up to 2 minutes, but not more.

Candidates should not recite presentations they have learned by heart.

In preparing their topic, candidates are advised to introduce a range of relevant vocabulary and to anticipate questions the examiner might ask. They should be prepared to give further examples, explanations and clarifications as requested by the examiner.

A candidate who fails to present a prepared topic will not earn any marks for this section of the examination.

ii General conversation The conversation will include discussion of some of the subject areas listed for the relevant grade. The examiner may sometimes introduce everyday objects or pictures to facilitate conversation. Less reliance is placed on 'question and answer' than at the Initial stage, the overall aim being to set up as genuinely interesting and relaxed an exchange as the candidate's interests and ability permit.

Candidates are encouraged to contribute as much as they can to the conversation and make every effort to show the examiner the range and quality of the language at their command. They should be ready for brief exchanges which could incorporate any of the subject areas or language items listed for the grade.

Elementary stage

Grade 4



Format

The candidate holds a conversation with the examiner. *Total time:* 10 minutes

The two major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- General conversation with the examiner (5 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2 minutes) using objects or pictures to illustrate
- answer questions on the prepared topic, and participate in informal discussion of the topic, during which the examiner might request more information, facts or details
- talk about past events
- talk about future plans
- express likes and dislikes
- express simple comparisons
- contribute to the conversation by making appropriate statements and responses to questions with a minimum of hesitations



New grammatical items

- The simple past tense of regular and common irregular verbs.
- The use of *going to* for future plans.
- Adverbs of manner and frequency.
- Comparatives and superlatives of adjectives.

—in addition to items listed for Grades 1–3



New subject areas for conversation

- Own holidays
- Shops
- Jobs
- Sports
- Hobbies
- Shopping
- Work

—in addition to items listed for Grades 1–3

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

Use may be made of visual or other material to help focus discussion. In particular, candidates are advised to bring material in support of their topic presentations, but this should never include birds, insects, reptiles or other live animals.



Assessment criteria

These are set out in the introduction to the Elementary stage on page 16

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

What are you going to talk to me about?
 What are these people going to do?
 Where did you go for your holiday last year?
 What do you like to do on the beach?
 When did you come to this school?
 How long are you going to stay here?
 How often do you go to the cinema?
 What are some of the differences between
 London and Barcelona?
 Tell me about your hobbies.

Candidate

I'm going to talk about health.
 They're going to get on the plane.
 I went to the beach.
 I like to lie in the sun. I like to swim, but my
 parents don't like swimming very much.
 I came here two years ago.
 I'm going to stay here for nine months.
 About once a month.
 London is bigger than Barcelona The best
 thing about London is Big Ben.

Grade 5



Format

The candidate holds a conversation with the examiner. *Total time:* 10 minutes

The two major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- General conversation with the examiner (5 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2 minutes) using objects or pictures to illustrate it
- answer questions on the prepared topic and participate in informal discussion of the topic, during which the examiner might request more information, facts or details, and give reasons for making particular statements
- give short narrative accounts and descriptions of events
- answer open-ended questions and requests for clarification and further information
- communicate ideas
- give reasons
- express preferences



New grammatical items

- The present perfect tense including use with *for, since, ever, never*
- Connecting clauses using *and, but, because*
- The future tense
- Expressions of preference
- Expressions of quantity
- More expressions relating to past time

—in addition to the items listed for Grades 1–4



New subject areas for conversation

- Festivals
- Travel
- Celebrations and customs
- Entertainment

—in addition to the items listed for Grades 1–4

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

Candidates should refer essentially to the situation and context in their own country. Candidates are not expected to know about these subjects in other countries or in Britain, but are of course free to talk about them if able to do so.

Maximum use should be made of candidates' knowledge of the home environment and their own experiences. It is to be noted that this is the last grade in which the examiner is permitted to make allowances for hesitations in the candidate's communication.



Assessment criteria

These are set out in the introduction to the Elementary stage on page 16.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

How long have you been in this school?

Why did you come to this school?

What is your topic?

Are there any other festivals?

Which do you prefer?

How do you prefer to travel?

Please ask me a question.

What sort of programmes do you like to see on television?

Candidate

I've been here for two years.

Because my parents chose it for me.

I want to talk about festivals in my country.

Yes, we celebrate Easter.

I prefer living in town.

I prefer to travel by car. Of course, travelling by plane is much quicker.

Do you like to travel to other countries?

I like to watch adventure films and news programmes.

Grade 6



Format

The candidate holds a conversation with the examiner. *Total time: 10 minutes*

The two major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- General conversation with the examiner (5 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2 minutes) using objects or pictures to illustrate it
- answer questions on the prepared topic, and participate in informal discussion of the topic, during which the examiner might request more information, facts or details,
- give reasons for making particular statements and be prepared to talk about the purpose or necessity for a course of action or another element of the topic
- express opinions and impressions
- express probability, necessity and purpose
- express intentions
- adapt language in order to manage less predictable elements of the conversation
- maintain the flow of communication



New grammatical items

- The first conditional and the past continuous tense
- Modals appropriate to the types of expression listed above, e.g. *could, must, have to, need to, may, might*
- Infinitive of purpose
- Verbs not normally used in the continuous form
- Further expressions relating to future time

—in addition to items listed for Grades 1–5



New subject areas for conversation

- Environment
- Transport
- Money
- Food
- Health

—in addition to items listed for Grades 1–5

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.



Assessment criteria

These are set out in the introduction to the Elementary stage on page 16.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

What have you decided to talk about?

If I come to your country, can I get health treatment?

What do you intend to do when you have finished school?

Who do you think is responsible for pollution?

What were you doing when you heard the crash?

Tell me what you thought of the play.

What are you doing this evening?

Candidate

I've chosen to talk about the health system in my country.

If you are a foreigner and you come to my country, you can have health treatment free. It means that if you fall ill, you can have treatment in any hospital.

I want to follow my father and become a doctor.

I think industry is responsible for a lot of pollution. ...
If you pollute the atmosphere, you should pay.
That's what I think.

I was talking to my friend on the telephone.

I didn't like it at all. I thought the characters were not real.

I'll probably go to the cinema with my friend.

Intermediate stage

Introduction



Candidate profile

By the end of the Intermediate stage the candidate can

- understand more complex speech used in the discussion of reasonably familiar topics
- converse with fluency and some spontaneity in every day situations, including those related to studies, work or professional activities
- express and support views and opinions in discussion of these and other topics
- contribute with greater independence to conversations relating to the above topics and to a range of topics of social and current interest
- initiate, maintain and influence the direction of conversations on general and more abstract topics

This profile is based on a broad definition of the second common reference level (B1/B2 Independent User, previously Threshold to Vantage) proposed in draft two of the Council of Europe's Common European Framework of Reference (1996).



Format

The conversation consists of five phases in each grade:

- brief introduction and greeting of the candidate and setting him/her at ease
- presentation by the candidate of a prepared topic followed by discussion with the examiner
- presentation by the candidate of a prepared text followed by discussion with the examiner
- general conversation around subject areas selected by the examiner from those listed under the appropriate or previous grades
- end of the conversation signalled by the examiner



Procedure

After the initial greetings and introduction, the examiner invites the candidate to present his/her topic. This should only take 2–3 minutes. Discussion of the topic lasting no more than 3 minutes will follow.

The examiner invites the candidate to present the prepared text. This is followed by questioning and discussion led by the examiner. Presentation and discussion of the text should take up to 5 minutes.

The examiner will then initiate conversation on subject areas selected from the syllabus for a further 4–5 minutes.

The examiner indicates the end of the conversation and says goodbye to the candidate.



Assessment criteria

At each grade the examiner will apply the following criteria:

- | | |
|----------------------|---|
| Readiness | <ul style="list-style-type: none"> • understanding the speech of and points made by the examiner • maintaining the flow of the conversation, displaying promptness of response and avoiding too much repetition • taking the initiative or influencing the direction of the conversation as necessary • satisfying the requirements listed under Candidate Performance for each grade and for all previous grades |
| Pronunciation | <ul style="list-style-type: none"> • production of a combination of individual sounds and the use of stress, rhythm and intonation so as to produce intelligible and natural sounding speech • competent variation of stress and intonation patterns to express attitudes and specific meanings |
| Usage | <ul style="list-style-type: none"> • accuracy of grammatical items used • choice of appropriate vocabulary and grammatical items • range of vocabulary, grammatical items and functions used |
| Focus | <ul style="list-style-type: none"> • communication of sufficient and relevant information required by the tasks set • coherent organisation of information and opinions communicated • ability to state communicative purpose • use of strategies, including rephrasing where necessary, in order to maintain the conversation and to emphasise particular points |



Guidance

i Prepared topic As in the previous grades, candidates are encouraged to prepare any topic they are interested in, knowledgeable about, and able to talk about readily and with confidence. The topic need not be chosen from the subject areas listed for the grade. The purpose is to give candidates the opportunity to display the language they know they can use.

Candidates are strongly recommended to bring into the examination one or more pictures, photos, diagrams, models or other suitable objects to illustrate their prepared topic and stimulate the conversation with the examiner. Without appropriate support materials, candidates may not give of their best and could consequently receive lower marks. However, birds, insects, reptiles or other live animals may *not* be brought into the examination room.

Candidates are advised to think carefully about the amount of material necessary for their topic, bearing in mind the time available. They should prepare enough material to sustain a presentation of up to 3 minutes, but not more.

Candidates should not recite presentations they have learned by heart.

In preparing their topic, candidates are advised to introduce a range of relevant vocabulary and to anticipate questions the examiner might ask. They should be prepared to give further examples, explanation, clarifications and personal opinions as requested by the examiner.

Candidates are reminded that they are judged only on the quality of their communication, as defined by the four criteria of readiness, pronunciation, usage and focus.

A candidate who fails to present a prepared topic will not earn any marks for this section of the examination.



ii Prepared text The theme of the prepared text should be different from the prepared topic. The chief requirement is that the text be material published in English. It may be any book, reader, anthology, magazine, journal, internet text, newspaper article or selection of articles. Simplified and abridged texts are acceptable.

Candidates are advised to select the text themselves where possible. They will be expected to talk about the text knowledgeably and confidently during the time allocated to this phase of the examination, and not simply about the general topic of the text.

The candidate should have read and thought about the text carefully, should be able to give a brief account of the content of the text and should be prepared to explain in more detail any part or feature of it that the examiner may select.

Candidates are reminded that the examiner will be assessing the range and quality of the language used in discussing the text, and not the length or level of difficulty of the chosen text.

The chosen text should be brought into the examination room. Failure to do so is likely to reduce the range and quality of discussion, and consequently the candidate may receive lower marks.

A candidate who fails to prepare and present a text will not earn any marks for this section of the examination.

iii General conversation The conversation will include discussion on one or more of the subject areas listed for the relevant grade. The examiner may sometimes introduce pictures to facilitate conversation. At the Intermediate stage, the candidate will be capable of initiating and sustaining more conversation than at the Elementary stage.

Although candidates' interests and ability may still limit the scope and direction of the conversation, they will be expected to

- take more responsibility for the content
- contribute opinions and ideas as well as information on the subject areas listed for the grade
- maintain the flow of the exchange, and
- demonstrate to the examiner the range and quality of the language at their command.

[Turn over for syllabus for Intermediate stage Grades 7–9

Intermediate stage

Grade 7



Format

The candidate holds a conversation with the examiner. *Total time: 15 minutes*

The three major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- Presentation of a text prepared by the candidate followed by discussion of that text with the examiner (5 minutes).
- General conversation with the examiner (5 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2–3 minutes) and a short presentation of a prepared text (2 minutes)
- answer questions on the prepared topic and on the prepared text
- participate in informal discussion of the topic during which the examiner might request further information, clarifications and further explanations
- participate in informal discussion of the text during which the examiner might request further information and the candidate's opinions
- give advice and opinions
- make suggestions
- express possibility and uncertainty
- talk about the future in the past
- respond appropriately to a change of topic



New grammatical items

- The second conditional
- Simple passive
- Modals of possibility, uncertainty, suggestion (e.g. *should/ought to, may, might*)
- Future perfect tense

—in addition to items listed for Grades 1–6



New subject areas for conversation

- Education
- National customs
- Diet

—in addition to items listed for Grades 1–6

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

The text at this grade may be any book, reader, anthology, magazine, journal, internet text, newspaper article or selection of articles published in English. Simplified and abridged texts are acceptable. Failure to bring the text to the examination room could result in unsatisfactory marks for readiness and focus.



Assessment criteria

These are set out in the introduction to the Intermediate stage on page 25.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

That must be quite expensive. Do you think you are going to buy one?

If you did have a lot of money, which one would you buy?

Could you give me some advice about what to look for?

Have you brought a text to talk about?

Why did you choose this article?

It says here ... Why is that?

What do I have to do to get a driving licence?

Do you think taking exercise is important?

Candidate

I don't think so, because I do not have that much money. But I could ask my father, because he has a business. ...

I'd probably buy a sports car.

I don't know what it's like in England, but you should find a good insurance company ...

Yes, I've brought an article which I took from last Sunday's paper. It's about ...

I was particularly interested in ...

I think it may be that the writer ...

First you have to get an application form and fill it in. When it has been filled in, it must be posted to ...

I think exercise is important for keeping fit, but diet is important too.

Grade 8



Format

The candidate holds a conversation with the examiner. *Total time: 15 minutes*

The three major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- Presentation of a text prepared by the candidate followed by discussion of that text with the examiner (5 minutes).
- General conversation with the examiner (5 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2–3 minutes) and a short presentation of a prepared text (2 minutes)
- answer questions on the prepared topic and on the prepared text
- participate in informal discussion of the topic during which the examiner might request further information, clarifications and further explanations, and ask the candidate for his/her personal opinion and feelings about the topic
- participate in informal discussion of the text during which the examiner might request further information and the candidate's opinions and ask about the author's attitude to the subject
- express feeling and emotion
- express impossibility
- report the conversations of others
- hypothesise
- influence the direction of the conversation
- respond to more complex utterances
- rephrase where necessary in order to maintain conversation



New grammatical items

- The third conditional
- Conditionals with *unless*
- Present perfect continuous tense
- Past perfect tense
- Reported speech

—in addition to items listed for Grades 1–7



New subject areas for conversation

- Social life
- Technology
- The world of work

—in addition to items listed for Grades 1–7

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

The criteria for choice of the prepared text are the same as for Grade 7. Failure to bring the text to the examination room could result in unsatisfactory grades for readiness and focus.



Assessment criteria

These are set out in the introduction to the Intermediate stage on page 25.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner	Candidate
Have you been working in this company for long?	Yes, I have.
How long?	For about two years. I had wanted to work for it since I left school.
How did you find the job?	I saw an advertisement in the paper.
How were you appointed?	I was sent a form to fill in, and then I went for an interview with the director. He asked me some questions and then I was offered the job.
What do you think of this project?	I think it is a good idea, but I think it is a long process and very complicated.
What does the director of the project say?	Last week he said that he was sure there would be enough money for the project.
When will the project be completed?	They might finish it by the end of next year.
Did you enjoy reading the book?	Yes I did, but you will not find betrayal and passion in this book. They fell in love because of their loneliness.
What about the place of women at work?	I think in England it is about equal between men and women, but in Japan it is very traditional.
Would you encourage young people to go to museums?	I found many school children in the museum, drawing pictures. Teachers should take the children there and explain what is there.
Which were the important subjects at school?	There was science and history. But we were also encouraged to do art.

Grade 9



Format

The candidate holds a conversation with the examiner. *Total time:* 15 minutes

The three major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- Presentation of a text prepared by the candidate followed by discussion of that text with the examiner (5 minutes).
- General conversation with the examiner (5 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2–3 minutes) and a short presentation of a prepared text (2 minutes)
- answer questions on the prepared topic and on the prepared text
- participate in informal discussion of the topic during which the examiner might request further information, clarifications and further explanations, ask the candidate for his/her personal opinion and feelings about the topic, find out the candidate's feeling on the topic and discuss any conclusions the candidate might draw
- participate in informal discussion of the text during which the examiner might request further information and the candidate's opinions, ask about the author's attitude to the subject, and find out what the candidate's attitude is towards the text
- express abstract ideas
- express regrets, wishes and hopes
- show ability to emphasise main points



New grammatical items

- The habitual past using *used to*
- Verbs followed by gerund and/or infinitive according to meaning
- Relative clauses with and without relative pronouns
- More complex forms of the passive

—in addition to items listed for Grades 1–8



Subject areas for conversation

All subject areas listed for Grades 1–8:

- education
- national customs
- social life
- technology
- the world of work
- diet

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.

The criteria for choice of prepared text are the same as for Grade 7. Failure to bring the text to the examination room could result in unsatisfactory grades for readiness and focus.



Assessment criteria

These are set out in the introduction to the Intermediate stage on page 25.

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

Did you just go with your friends, or did members of your family go along?

Would you say there is much skill involved?

Are you going again next year?

What does the writer write about in this article?

Do you think that this is true?

How do you think university education could be improved in your country?

Candidate

This is a picture that was taken in the mountains. I had the photos developed last week. I used to go to the mountains when I was very young to spend my Christmas holidays I enjoy going on holiday with my friends ... I wish my brothers could have come with me, but ...

When I was a child, I used to take part in some ski races.

I would say so, although there is a lot of technique necessary ...

I had hoped to, but I'm afraid I'm unlikely to be given enough time off.

The writer talks about a particular seasonal disorder which affects people in winter.

They say in this article that as winter approached, she didn't carry on with her job, she felt depressed ...

I think that it is true. But I think that physical illness has a relationship with something in the brain, because this woman, for example, lost her job ...

I would increase the research facilities and staff pay—if I had the power!

Advanced stage

Introduction



Candidate Profile

By the end of the Advanced stage the candidate can

- understand the main points, arguments, inferences, changes in register and emphasis in extended, complex and sometimes unstructured speech
- contribute and respond confidently and appropriately in interaction in all social and professional contexts, except on matters outside their cultural range
- control the direction of the conversation and maintain its flow with ease, relating skilfully to the contributions of the examiner
- exhibit a high degree of control over the range and accuracy of the vocabulary and grammar used as well as over their pronunciation

This profile is based on a broad definition of the third common reference level (C1/C2 Proficient User, previously Operational Efficiency to Mastery) proposed in draft two of the Council of Europe's Common European Framework of Reference (1996).



Format

The conversation consists of six phases:

- brief introduction and greeting of the candidate and setting him/her at ease
- presentation by the candidate of a prepared topic followed by discussion with the examiner
- presentation by the candidate of a prepared text followed by discussion with the examiner
- listening to a short text read aloud once followed by questions
- general conversation around subject areas selected by the examiner from those listed under the appropriate or previous grades
- end of the conversation signalled by the examiner



Procedure

After the initial greetings and introduction, the examiner invites the candidate to present his/her topic. This should only take 2–3 minutes. Discussion of the topic lasting no more than 3 minutes will follow.

The examiner invites the candidate to present the prepared text. This is followed by questioning and discussion led by the examiner. Presentation and discussion of the text should take up to 7 minutes.

The candidate is then invited to listen to a short text which the examiner will read aloud once only. The candidate may take notes during the reading. This is followed by a few general and specific questions to check the candidate's understanding. This activity should take about 6 minutes.

The examiner will then initiate conversation on subject areas selected from the syllabus for a further 4–5 minutes.



Assessment criteria

The examiner will apply the criteria of readiness, pronunciation, usage and focus that are set out with the detailed syllabus for each grade.



Guidance

Examinations at this stage demand a much higher level of language proficiency and conversational ability than that asked for at the Intermediate stage. Candidates will be well motivated and have particular reasons for wanting to be fluent in English. Candidates will be mature and experienced enough to handle abstract concepts and to contribute to discussion of matters of major importance in today's world.

i Prepared topic Candidates are encouraged to prepare any topic which they are particularly interested in or knowledgeable about. The topic need not be chosen from the subject areas listed for the grade. The purpose is to give them the opportunity to display an extended command of the language when presenting and discussing a topic that is well known to them. The guidance given in the Introduction to the Intermediate stage on page 00 is still relevant at the Advanced stage.

A candidate who fails to present a prepared topic will not earn any marks for this section of the examination.

ii Prepared text Only one text should be prepared and presented at these grades. Texts which have been simplified for learners of English are not acceptable at this stage. Candidates will be expected to talk about the content of the text, the opinions and attitudes of the author, and their own opinions and attitude to the text. The discussion will be in greater depth and more detailed than at the Intermediate Stage.

The guidance given in the introduction to the Intermediate Stage on page 00 is still relevant at the Advanced stage.

A candidate who fails to prepare and present a text will not earn any marks for this section of the examination.

iii Listening comprehension To assist understanding, the examiner focuses the candidate's attention on the key idea of a passage which is then read aloud once only. While it is being read, candidates may make notes for use during the discussion.

The examiner will ask some questions to ascertain the extent of the candidate's understanding of general and specific points and then extend discussion on the theme of the passage, time permitting. The passages increase in difficulty from Grade 10 to Grade 12.

iv General conversation The conversation will include more detailed discussion around one or more of the given subject areas. Examiner and candidate are responsible in approximately equal measure for content, coherence and direction of the conversation.

Advanced stage

Grade 10



Format

The candidate holds a conversation with the examiner. *Total time: 25 minutes*

The four major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- Presentation of a text prepared by the candidate followed by discussion of that text with the examiner (7 minutes).
- Reading aloud of a short text by the examiner, followed by questions to check the candidate's understanding of the text (6 minutes).
- General conversation with the examiner (6 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2–3 minutes) and a slightly longer presentation of one prepared text (3–4 minutes)
- answer questions and enter into discussion following the presentations of the topic and text and be prepared to account for statements made and to explain his/her attitudes to the topic and text
- understand the gist and main points of a spoken text
- summarise information, ideas and arguments
- develop an argument and defend a point of view
- sustain discussion at all points in the conversation
- maintain the flow of the conversation without breakdown



New grammatical items

- Past perfect continuous tense
- *Must* plus present perfect
- *Could have* plus participle

—in addition to items listed for Grades 1–9



New subject areas for conversation

- International events
- Social issues
- The economy
- Ambitions
- Equal opportunities
- Science/technology

—in addition to items listed for Grades 1–9

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.



Assessment criteria

- | | |
|----------------------|---|
| Readiness | <ul style="list-style-type: none">• understanding of all the main points and arguments made by the examiner• responding appropriately and without hesitation• taking the initiative where appropriate• maintaining the flow of the conversation with a minimum of self-correction |
| Pronunciation | <ul style="list-style-type: none">• production of individual sounds so as to be intelligible to the listener, with only minimal transfer of sounds from the mother tongue• use of stress and intonation patterns that are recognizably specific to English, but with occasional lapses of intelligibility• use of appropriate stress and intonation patterns to emphasise meaning |
| Usage | <ul style="list-style-type: none">• range of vocabulary and grammar appropriate to the subjects under discussion• grammatical accuracy that denotes full control over all items specified for Grades 7 and below and only occasional lapses in control of items listed for Grades 8 and above |
| Focus | <ul style="list-style-type: none">• entirely appropriate content of all contributions to the conversation• achievement of communicative purpose of all contributions to the conversation• adequate organisation of content of contributions to the conversation• evidence of strategies to initiate and control the conversation from time to time |

Turn over

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

So how was this situation resolved?

Now moving to the text:

Is that true? How realistic is the text?

Do you not think this title is misleading?

What do you understand about equal opportunities?

Do you think this person influenced you?

If you hadn't have gone into the theatre, what do you think you would have done?

To what extent do you think the media influences everyday life?

Candidate

The first problem we had was how to choose a cast for the play ...

So what happened was that they were allowed to have children and wives in jail if the husband was arrested ...

This text is about Rio, which has been elected the most exciting city.

It's quite realistic, but there are a number of misconceptions held by the public ...

Yes, I think so, because ...

I believe I saw the film when I was a child. I mean when I really was a child.

No prejudice at all, no prejudice because of sex, because of ...

Until a few years ago, a woman could be doing exactly the same work as a man and earning less money.

I think she could have influenced me, although I didn't realise it at the time. I must have been quite naive then ...


I'd have probably been a lawyer ...

It rather depends on what you mean by media, but I have little doubt that ...

Advanced stage

[Turn over for Grade 11


Grade 11

 ***Format***

The candidate holds a conversation with the examiner. *Total time: 25 minutes*

The four major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- Presentation of a text prepared by the candidate followed by discussion of that text with the examiner (7 minutes).
- Reading aloud of a short text by the examiner, followed by questions to check the candidate's understanding of the text (6 minutes).
- General conversation with the examiner (6 minutes).

 ***Candidate performance***

The candidate is expected to

- make a short presentation of a prepared topic (2–3 minutes) and a slightly longer presentation of one prepared text (3–4 minutes)
- answer questions and enter into discussion following the presentations of the topic and text and be prepared to account for statements made and to explain his/her attitudes to the topic and text
- understand the main points, inferences and changes of register in both the spoken text and in the examiner's natural speech
- justify points made during his/her own contributions to the conversation
- evaluate and challenge statements and arguments made by the writer of the prepared text and by the examiner at any time during the whole conversation
- maintain the flow of conversation with ease, changing the direction as necessary

 ***New grammatical items***

- Full range of conditionals
- Complex adverbial, noun-phrase and sentence structures

—in addition to items listed for Grades 1–10



New subject areas for conversation

- The media
- Advertising
- Lifestyles
- The arts

—in addition to items listed for Grades 1–10

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.



Assessment criteria

- | | |
|----------------------|---|
| Readiness | <ul style="list-style-type: none">• understanding of all the main points, arguments, inferences and changes in register made by the examiner• responding appropriately with confidence and without hesitation• taking the initiative and changing direction of conversation where appropriate• maintaining the flow of the conversation with ease |
| Pronunciation | <ul style="list-style-type: none">• production of individual sounds so as to be intelligible to the listener, with only occasional sounds that deviate from an internationally intelligible model• use of stress and intonation patterns that are recognizably specific to English, but with only an occasional lapse of intelligibility• use of appropriate stress and intonation patterns to emphasise meaning and attitude |
| Usage | <ul style="list-style-type: none">• range of vocabulary and grammar appropriate to the subjects under discussion• grammatical accuracy that denotes full control over all items specified for Grades 9 and below and only occasional lapses in control of items listed for Grades 10 and 11 |
| Focus | <ul style="list-style-type: none">• entirely appropriate content for all contributions to the conversation• achievement of communicative purpose in all contributions to the conversation• adequate organisation of content in contributions to the conversation• evidence of strategies to initiate and control the conversation when desired |

Turn over

Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

Candidate

... When the Health Secretary announced that there was a link between BSE and CJD, it led to a crisis in exports. It also led to a crisis of confidence at home. Overnight, Britain lost nearly £500,000,000, which was devastating ... Rather than to protect health, it was done to punish Britain for giving misleading information about it.

Do you think this is an acceptable explanation?

Well, I'm not a politician ...

That's a play on words. Can you explain it?

Well, I think it denotes a change of some kind ...

What would you say is the thesis of this article?

Now the ban is lifted, but the warning still exists. We can't say that beef is safe yet ...

What is the view of comedy in your country?

Many critics say we are laughing at ourselves when ...

How would you characterise the difference between humour in Britain and in your country?

Maybe British jokes are more sophisticated and humour in our country is more direct.

How powerful would you say the media are in your country?

I would say the influence of the television is quite strong. The state channels try to influence us politically but they don't succeed. And the independent channels take a more objective view ...

(Listening comprehension)

In what you read out just now, there are a number of ideas about laughter. For example, some say laughter has some magic power. There's another story about the fairy princess for whom laughter ...

Advanced stage

[Turn over for Grade 12

Grade 12



Format

The candidate holds a conversation with the examiner. *Total time: 25 minutes*

The four major phases of the conversation are:

- Presentation of a topic prepared by the candidate followed by discussion of that topic with the examiner (5 minutes).
- Presentation of a text prepared by the candidate followed by discussion of that text with the examiner (7 minutes).
- Reading aloud of a short text by the examiner, followed by questions to check the candidate's understanding of the text (6 minutes).
- General conversation with the examiner (6 minutes).



Candidate performance

The candidate is expected to

- make a short presentation of a prepared topic (2–3 minutes) and a slightly longer presentation of one prepared text (3–4 minutes)
- answer questions and enter into discussion following the presentations of the topic and text and be prepared to account for statements made and to explain and justify his/her attitudes to the topic and text
- understand the main points, inferences, and changes of register and emphasis in both the spoken text and in the examiner's natural speech
- defend and justify points made during his/her own contributions to the conversation
- evaluate and challenge statements and arguments made by the writer of the prepared text and by the examiner at any time during the whole conversation
- control and sustain the discussion at all times
- maintain the flow of conversation with ease



Grammatical items

There are no new grammatical items at this grade. Candidates are expected to have mastered all the items listed for all the previous grades.



Subject areas for conversation

- The media
- Advertising
- Lifestyles
- The arts
- International events
- Social issues
- The economy
- Science/technology
- Equal opportunities

There are no new subject areas for this grade. The examiner may discuss two or more of the above.

Candidates should be able to make use of a range of vocabulary items relating to the above subject areas.



Assessment criteria

- | | |
|----------------------|---|
| Readiness | <ul style="list-style-type: none">• understanding of all the main points, arguments, inferences and changes in register and emphasis made by the examiner• responding appropriately with confidence and ease at all times• taking the initiative and changing the direction of conversation where appropriate |
| Pronunciation | <ul style="list-style-type: none">• controlling and maintaining the flow of the conversation in a natural way• production of individual sounds so as to be fully intelligible to the listener, with only the rare sound that deviates from an internationally intelligible model• use of stress and intonation patterns that are recognizably specific to English, and without any lapse of intelligibility• use of appropriate stress and intonation patterns to emphasise meaning and attitude |
| Usage | <ul style="list-style-type: none">• full range of vocabulary and grammar appropriate to the subjects under discussion• grammatical accuracy that denotes full control over all items specified for all grades |
| Focus | <ul style="list-style-type: none">• entirely appropriate content of all contributions to the conversation• achievement of communicative purpose in all phases of the conversation• competent organisation of content of contributions to the conversation• evidence of strategies to initiate and control the conversation |

Turn over



Examiner and candidate language

The sample exchanges below show some ways in which examiners and candidates might express themselves during the conversation. It is stressed that these are only examples.

Examiner

What conclusion did the writers come to?

When do you see the problem being resolved?

Would you say that Charlie Chaplin was an unhappy man?

Candidate

No, the writers didn't come to any conclusion. They discussed the problem ...

It's difficult to answer. But I hope it'll be resolved very soon ...

No, I can't say that he was miserable, no. But in most comedies he portrayed the life of poor people ...

(Extract from a candidate's presentation of the topic)

... I don't know, but I know that the estimated reserves were 20 billion barrels. ... As I said, the most important problem, and that is my pet subject because I am an international lawyer and I am very interested in this problem, is the international status of First I want to refer to the past, when two agreements were signed ...

(Candidate comments on lifestyle)

Yes, there are a lot of differences. For example, when you enter a pub: people don't think that it's the wrong thing for a woman to come into a pub. But back home, it's not so. For women to enter a pub is certainly disapproved of.

Further examples would not be very informative at this grade, since candidates will be expected to have a control of language approaching that of a native speaker and be able to discuss a wide range of subject matter as in normal conversation and debate.

Appendix 9

Example of the Cambridge ESOL Speaking Examination

PAPER 5 SPEAKING

GENERAL DESCRIPTION

Paper format	The Speaking test contains four parts.
Timing	15 minutes.
No. of parts	4.
Interaction pattern	Two candidates and two examiners. One examiner acts as both interlocutor and assessor and manages the interaction either by asking questions or by providing cues for candidates. The other acts as assessor and does not join in the conversation.
Task types	Short exchanges with the interlocutor and with the other candidate; a one-minute 'long turn'; a collaborative task involving the two candidates; a three-way discussion.
Task focus	Exchanging personal and factual information, expressing and finding out about attitudes and opinions.
Marks	Candidates are assessed on their performance throughout the test.

STRUCTURE AND TASKS

PART 1

Task type and format	Conversation between the candidates and the interlocutor. The candidates are asked to respond to one another's questions about themselves and to respond to the interlocutor's questions.
Focus	General interactional and social language.
Timing	3 minutes.

PART 2

Task type and format	Individual 'long turns' with brief responses from the second candidate. Each candidate in turn is given visual prompts. They talk about the prompts for about one minute; the second candidate responds as specified.
Focus	Organising a larger unit of discourse by describing, comparing and contrasting, and speculating.
Timing	One-minute 'long turn' for each candidate.

PART 3

Task type and format	Two-way conversation between the candidates. The candidates are given visual and spoken prompts, which are used in a decision-making task. At the end of this part, candidates are asked to report on the outcome of their discussion.
Focus	Negotiating and collaborating, discussing, evaluating, speculating, expressing and justifying opinions, agreeing and/or disagreeing, decision-making and/or selecting.
Timing	4 minutes.

PART 4

Task type and format	Discussion on topics related to the collaborative task. The interlocutor leads a discussion to explore further the topics or issues of the collaborative task.
Focus	Exchanging information, expressing and justifying opinions, agreeing and/or disagreeing.
Timing	4 minutes.

5

The four parts of the Speaking test

Format

The paired format of the CAE Speaking test (two examiners and two candidates) offers candidates the opportunity to demonstrate, in a controlled but friendly environment, their ability to use their spoken language skills effectively in a range of contexts. The test takes 15 minutes. One examiner, the interlocutor, conducts the test and gives a global assessment of each candidate's performance. The other, the assessor, does not take any part in the interaction but focuses solely on listening to, and making an assessment of, the candidates' oral proficiency.

At the end of the Speaking test, candidates are thanked for attending, but are given no indication of the level of their achievement.

The standard format is two examiners and two candidates, and, wherever possible, this will be the form which the Speaking test will take. In cases where there is an uneven number of candidates at a centre, the last Speaking test of the session will be taken by three candidates together instead of two. The test format, test materials and procedure will remain unchanged but the timing will be longer: 23 minutes instead of 15. A 1:1 test format will only be allowed in exceptional circumstances and emergencies.

The Speaking test consists of four parts, each of which is assessed. Each part of the test focuses on a different type of interaction: between the interlocutor and each candidate, between the two candidates, and among all three. The patterns of discourse vary within each part of the test.

■ PART 1 – INTERVIEW

This part tests the candidate's ability to use general interactional and social language.



Sample task and assessment criteria: pages 60 and 63.

This part of the test gives candidates the opportunity to show their ability to use general interactional and social language and talk about their interests, studies, careers, etc. Candidates are expected to respond to the interlocutor's and their partner's questions, and to listen to what their partner has to say.

In this part of the test, the interlocutor asks candidates for some information about themselves. Candidates then ask each other questions using prompts given by the interlocutor. The interlocutor then asks the candidates to offer their opinion on certain topics.

■ PART 2 – LONG TURN

This part tests the candidates' ability to produce an extended piece of discourse.



Sample task and assessment criteria: pages 60–61 and 63.

In this part of the test, candidates are given the opportunity to speak for one minute without interruption. Each candidate is asked to comment on and react to a different set of pictures or photographs. Candidates may be asked to describe, compare, contrast, comment, identify, eliminate and hypothesise or speculate. Tasks may be completely different for each candidate, or they may be 'shared', e.g. when there is a group of three candidates. Shared tasks set candidates the same task but each candidate receives different visual stimuli.

Candidates can show their ability to organise their thoughts and ideas, and express themselves coherently in appropriate language. Candidates should pay attention while their partner is speaking, as they are asked to comment briefly (for about 20 seconds) after their partner has spoken. Candidates should be made aware, however, that they should not speak during their partner's long turn.

Candidates will always be asked to speculate about something which relates directly to the focus of the visuals. They will never be asked merely to describe the visuals.

■ PART 3 – COLLABORATIVE TASK

This part tests the candidates' ability to engage in a discussion and to work towards a negotiated outcome of the task set.



Sample task and assessment criteria: pages 62 and 63.

The candidates are given oral instructions and provided with a visual stimulus, e.g. several photographs, artwork or computer graphics, to form the basis for a task which they carry out together. Candidates are expected to work towards a negotiated completion of the task and are assessed on their ability to negotiate and collaborate with each other while doing this. At the end of this part of the test, candidates are asked to report on the outcome of their discussion.

The task gives candidates the opportunity to show their range of language and their ability to invite the opinions and ideas of their partner. There is no right or wrong answer to the task and candidates can agree to differ.

■ PART 4 – DISCUSSION

This part tests the candidates' ability to engage in a discussion based on the topics or issues raised in the collaborative task in Part 3.



Sample task and assessment criteria: pages 62 and 63.

In this part of the test, the interlocutor directs the interaction by asking questions which encourage the candidates to widen the scope of the topics or issues introduced in Part 3. The questions often focus on more abstract issues as the discussion develops.

This part of the test gives candidates an opportunity to show that they are capable of discussing topics and certain issues in more depth than in the previous parts of the test.

Preparation

General

- It is essential that students are able to participate in pair and group activities effectively, showing sensitivity to turn-taking and responding appropriately to their partners. Pair and group activities should, therefore, be a regular feature of classroom learning.
- Students should be given extensive practice in listening carefully to instructions and remembering what they are asked to do.
- Students should be encouraged to react to pictures, photographs and graphics, etc. rather than merely describe them.
- Students should know exactly what to expect in each part of the test and they should be equipped with the right kind of language for each part, e.g. giving personal information, exchanging information/opinions, giving reasons, speculating, agreeing and disagreeing politely, justifying and negotiating.
- Students should be encouraged to speak clearly so that they can be heard and understood, and paraphrase effectively when they do not know or cannot remember a word. Students should be made aware that different varieties of standard English accents in the UK and elsewhere in the world are acceptable.
- It is useful to give students a 'mock' Speaking test before the examination so that they have an idea of how long each part of the test will be, and how they can maximise the time available to show the examiners what they can do.
- Students should be aware that if they are uncertain about what they have to do, they can ask for the instructions to be repeated but to do this too often will leave them less time to concentrate on the task itself.

- Students should be advised not to wait too long before they begin to speak. A short pause to gather their thoughts is acceptable, but anything longer than this will give them less time to produce a sample of language.

- Students should realise that producing a one-word answer will not give them the opportunity to show their range of language, etc. so they should expand on their answers and responses wherever possible.

N.B. In some centres candidates from the same school are paired together. However, where candidates from a number of different schools are entered at the same centre, some candidates may find that they are paired with a candidate from another school. Students should check with the centre through which they are entering for the local procedure.

By part

■ PART 1

- In this part of the test, examiners will ask candidates a range of questions about their everyday life, for example sports they enjoy, travel and holidays, work experience and so on. Encourage students to respond promptly with answers which are complete and spontaneous. Rehearsed speeches should be avoided as these might be inappropriate for the question asked.
- Encourage your students to look for opportunities to socialise with English speakers. In class, they could role-play social occasions in which they meet new people, e.g. parties, long train journeys, joining a new class, starting a new job.
- Students could be put into small groups to brainstorm questions from the categories above. The different groups could then answer each other's questions.
- The questions asked in Part 1 may relate to past experiences, present activities, or future plans. Make your students aware of the different structures required to respond to these questions appropriately.
- Students should be made aware that they are expected to react naturally to their partners and not rehearse speeches for this part of the test. They should show sensitivity to each other's contributions, invite their partners to participate, and not dominate the interaction. It is essential to demonstrate in class what is required in this part of the test.
- Encourage students to reformulate the interlocutor's prompt in the second section of Part 1. For example:

Interlocutor: Now I'd like you to ask **each other** something about your reasons for learning English.

X Candidate 1: Anna, what are your reasons for learning English?

✓ Candidate 1: Anna, why did you decide to start studying English?

5

■ Train students to 'think on their feet' and answer a question quickly even if they have never thought about that particular subject before. For example:

Interlocutor: What will you be doing in 10 years' time?

X Candidate 1: Oh, er. I've never ... I don't know.

✓ Candidate 1: I will probably be working for a very large international company and hopefully earning a lot of money, or I might be married with children of my own.

■ Encourage students to practise Part 1 in groups of three. One student could be the interlocutor and the other two the candidates, and they could then reverse roles. Materials from past papers can be used for this activity.

■ Advise students to try and use a variety of tenses, language and structures in this part of the test. This will create a good impression and give them confidence to tackle the other parts of the test.

■ PART 2

■ Give students practice in talking for one minute on a set subject, or 'holding the floor' in a classroom situation so that they can organise their thoughts and ideas quickly during this long turn.

■ Students need to be clear about what is considered an adequate response, e.g. their responses need to go beyond the level of pure description and contain a speculative element. For example:

X Candidate 1: In the first picture, the scene looks modern, in the other, it looks old-fashioned.

✓ Candidate 1: Both pictures of the building portray a calm and peaceful setting, but the older scene suggests that there was more traffic on the river at the time, whereas the more modern image ...

■ Read out some tasks from past papers, then hand over the corresponding sets of visuals and see if students can remember what they have to do. Tell them to listen for the introductory rubric, e.g. 'You will each have the same set of pictures to look at. They show people doing different jobs.' Students should then listen for a further three aspects: the first is always 'describe' or 'compare and contrast', the second is introduced by the word 'saying', and the third by the word 'and', e.g. 'I'd like you to **compare and contrast** two or three of these pictures, **saying** what the people might be thinking about **and** how difficult it might be for them to do these jobs.'

■ Give students practice using tasks with differing numbers of visuals. Some CAE tasks have five visuals, others four, three or two. When there are four or more, candidates will be asked to compare and contrast two or three. When there are three or fewer, candidates will be asked to talk about them all.

■ Tell students not to waste precious time saying, 'I'm going to talk about the picture in the top left-hand corner and the one in the bottom right-hand corner.' This is not necessary

and most students tend to make grammatical mistakes when trying to describe where the pictures are. In addition, if they run out of time, they have restricted their choice of pictures and may feel they cannot talk about the others. Tell students simply to start talking about their chosen pictures. Comparing and contrasting them will be enough to identify which pictures they are talking about.

■ Tell students not to adopt 'closure' techniques such as, 'That's it! I've finished!' They should keep talking until the interlocutor says, 'Thank you.' In this way, they will maximise the time available for their one-minute long turn.

■ Build up a bank of pictures which you can use for practice in the classroom. Encourage students to react to the pictures without giving them a specific task. This will help them to look at pictures in more depth, and train them to think of something to say if they run out of ideas during the examination itself.

■ Make sure that students have plenty of practice in organising their ideas coherently. Useful phrases to link ideas and compare and contrast pictures will help them. They can build up their own lists of suitable phrases throughout the course, thus ensuring that they have a range of language and structures to draw upon when necessary.

■ PART 3

■ Encourage students to make use of conversation 'fillers', e.g. 'Well, now, let me see ...', which they can call upon (sparingly) to give themselves time to think, and to make use of strategies which invite their partner to contribute to the discussion, e.g. 'Would you agree ...?'

■ Each time you do a Part 3 task in class, read the task aloud to students and see if they can remember what they have to do. It will help students to know that there is always a 'set up' rubric, e.g. 'Here are some pictures of ...' or 'I'd like you to imagine that ...' After the visual stimulus is handed over, the interlocutor will outline the task, which has two distinct prongs, e.g. 'Talk to each other about how these things are threatening the world we live in, and then decide which two are the biggest threat.' Although the completion of the task is not essential, as the interlocutor will ask them to summarise what they have decided at the end of the task, it is advisable for students to attempt to reach the specified outcome within the time allotted. If they do not listen carefully to the task, or remember what they have to do, they may be depriving themselves of the opportunity to demonstrate their command of a wide range of linguistic resources and communication strategies.

■ Warn students not to reach their decision in the first minute or so of the test. If students begin by saying, 'Well, I think we should choose this one and this one', they leave themselves with nothing to talk about for the remainder of the time. Train them to discuss each piece of visual stimulus in detail before reaching a decision. The core of the task is in the evaluation of the visuals, not in simply saying 'We have

chosen these two.' 4 minutes is a long time and students need strategies for making the most of the time available.

■ Doing timed tasks in class will help students make the best use of the 4 minutes available for Part 3.

■ Students should be encouraged to react to as great a variety of visual stimuli as possible and express ideas and opinions of their own. Simply agreeing or disagreeing with their partner, or echoing what their partner has said, will not enable them to show what they can do. They should always expand on what they say, e.g. instead of saying, 'Yes, I agree', a better response might be, 'Yes, I certainly do agree that this is a very serious problem, as you so rightly pointed out. But it is one we could do something about. It might be easier to solve than some of the others, don't you think?'

■ Conduct 'mock' Part 3 practice giving each student a different role card, e.g. Student A could constantly interrupt Student B, or Student B could be instructed to say almost nothing at all, or give one-word responses, thus forcing Student A to keep talking, or be constantly inviting their partner to speak. This will provide invaluable training in sensitivity to turn-taking and in managing and developing the interaction.

■ PART 4

■ Encourage students to talk about topical issues and issues of general interest and express an opinion about them so that they can participate fully in the last part of the test. They are asked questions by the interlocutor and they are expected to develop the discussion, rather than simply give one-word answers.

■ Tell students that they are not being assessed on their ideas, but examiners can only assess candidates on the language they produce, and those candidates who fail to make a contribution will not do well. Reading an English newspaper, or listening to or watching the international news on a regular basis will help give candidates ideas they may be able to use in Part 4 of the test.

■ Set up a regular debating lesson in class. Students A and B could be given a short time to argue either for or against an issue. The other members of the class could then be invited to express their own ideas. This will encourage students to have the confidence to express their ideas in public, and comment on issues they may never have thought about before.

■ After doing a Part 3 task, ask students what kinds of questions they think they may be asked. In groups, they could produce three or four and then compare them with those produced by other students. This will help them to be prepared for what they might be asked in Part 4 of the test.

■ At this stage of the test, the worst thing that can happen is a long silence. Train students to react almost immediately to what they are asked to talk about or to give themselves a little time by 'thinking aloud', e.g. 'Well, that is something I've never actually thought about but, on reflection, I would say that ...'

■ Students may be getting tired by this stage in the test. It is important that they are given practice sessions of 15 minutes so that they know exactly what it is like to do a Speaking test for this length of time. The impression they make at the end of the test is equally as important as the one they have made throughout the rest of the test. Regular participation in a complete test will train students not to lose their level of concentration as the end of the test approaches.

For Oral Examiners' Use Only

CAE PAPER 5

PART 1 (3 minutes / 3 minutes 30 seconds)

Interlocutor Good morning (afternoon/evening). My name is and this is my colleague,
And your names are?
Can I have your mark sheets, please?
Thank you.

First of all, we'd like to know a little about you.
(Select one or two questions and ask candidates in turn, as appropriate.)

- Where do you both/all live?
- How long have you been studying English?
- Have you been studying English together?
- What countries have you visited?

Now I'd like you to ask each other something about
(Select one or two prompts in any order, as appropriate.)

- things you particularly like about living in this country.
- entertainment and leisure facilities in this area.
- your reasons for studying English.
- a change you would like to make to your life in the future.

(Ask candidates one or more further questions in any order, as appropriate.)

- How important do you think English is in this country?
- How would you feel about going to live abroad permanently?
- What interesting events have happened in your life recently?
- Who do you think has had the greatest influence on your life so far? (Why?)
- What are your earliest memories of school?

Thank you.

For Oral Examiners' Use Only

CAE PAPER 5

PART 2 (4 minutes)

1. **AVIATION** (Compare, contrast and speculate)

Test Material 1

Interlocutor In this part of the test, I'm going to give each of you the chance to talk for about a minute, and to comment briefly after your partner has spoken.

First, you will each have the same set of pictures to look at. They show people making different kinds of visits.

Hand over the same set of pictures to each candidate.

(Candidate A), it's your turn first. I'd like you to compare and contrast two or three of these situations, saying why the people might be making these visits, and how important the visits might be for the people involved.

Don't forget, you have about one minute for this.

All right? So, (Candidate A), would you start now, please?

Candidate A ⌚ Approximately one minute.

Interlocutor Thank you.

Now, (Candidate B), can you tell us which visit you think would be the most memorable?

Candidate B ⌚ Approximately twenty seconds.

Retrieve pictures.

Interlocutor Thank you.

2. **Approaches to learning** (Compare, contrast and speculate)

Test Material 2

Interlocutor Now, I'm going to give each of you another set of pictures to look at. They show people learning in different situations.

Hand over the same set of pictures to each candidate.

Now, (Candidate B), it's your turn. I'd like you to compare and contrast two or three of these pictures, saying how the atmosphere is different in each situation, and what the benefits of each method of learning might be.

Don't forget, you have about one minute for this.

All right? So, (Candidate B), would you start now, please?

Candidate B ⌚ Approximately one minute.

Interlocutor Thank you.

Now, (Candidate A), can you tell us which method of learning you think is the most effective?

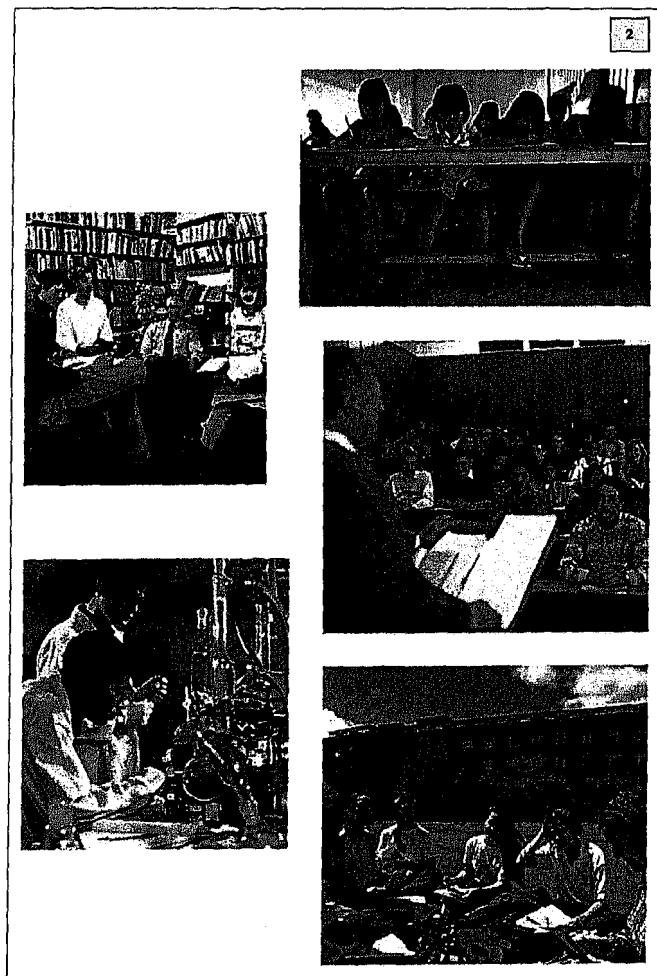
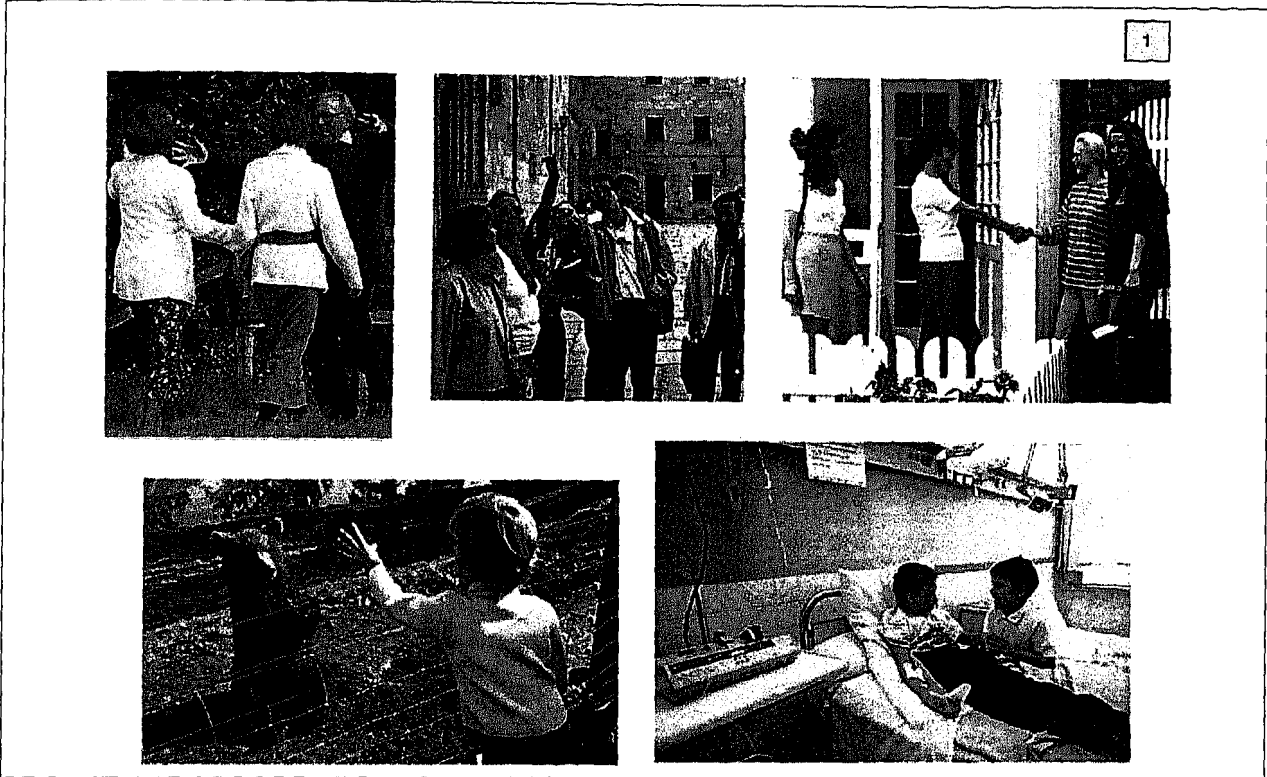
Candidate A ⌚ Approximately twenty seconds.

Retrieve pictures.

Interlocutor Thank you.

PAPER 5: SPEAKING

Part 2



For Oral Examiners' Use Only

CAE PAPER 5 **PARTS 3 and 4** (6 minutes for parts 3 and 4)

21. **International English Dictionary** (Discuss, evaluate and select) **Test Material 21**

PART 3

Interlocutor Now, I'd like you to discuss something between/among yourselves, but please speak so that we can hear you.

I'd like you to imagine you are choosing a cover for a new international English dictionary. Here are some suggestions for the cover picture.

Place picture sheet 21 in front of the candidates.

Talk to each other about what message these pictures communicate about the dictionary, and then decide which picture would be most successful in appealing to people worldwide.

You have about four minutes for this. (Six minutes for groups of three.)

Candidates Ⓞ *Approximately four minutes. (Six minutes for groups of three.)*

Interlocutor Thank you.

So, which picture have you chosen?

Retrieve picture sheet 21.

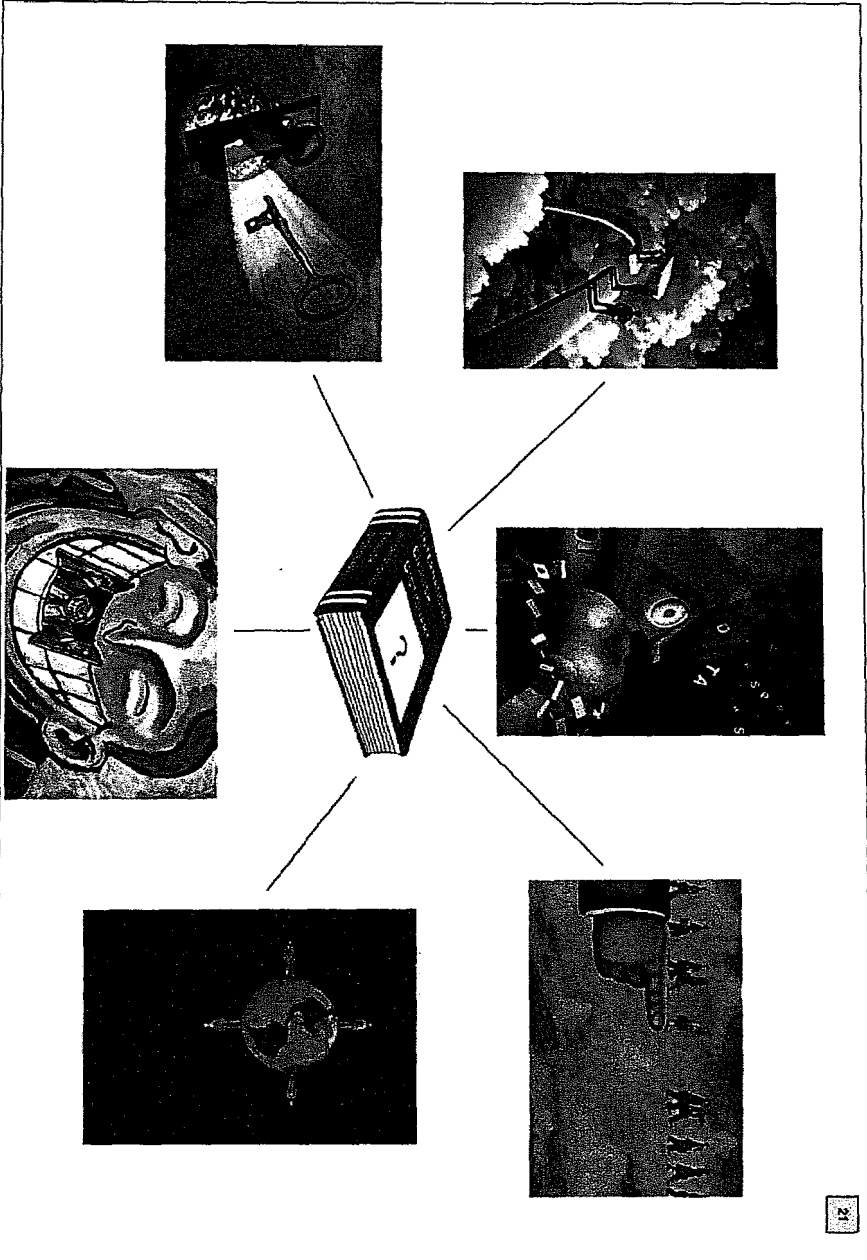
PART 4

Interlocutor *Select any of the following questions as appropriate:*

- How important are reference books such as dictionaries for students? (Why?)
- Some people say we read less nowadays than we used to. What's your opinion?
- How do you think we can encourage young people to read more?
- In the future, do you think we will all speak the same language? (Why (not)?)
- How has modern technology helped people to communicate with each other?

Thank you. That is the end of the test.

Check that all test materials have been replaced in the file.



21

Assessment

Throughout the test candidates are assessed on their own individual performance and not in relation to each other. Both examiners assess the candidates according to criteria which are interpreted at CAE level. The assessor awards marks according to four analytical criteria:

- Grammar and Vocabulary
- Discourse Management
- Pronunciation
- Interactive Communication.

The interlocutor awards a Global Achievement mark, which is based on the analytical scales.

These criteria should be interpreted within the overall context of the Cambridge Common Scale for Speaking on page 64, where CAE is at Level C1.

■ Grammar and Vocabulary

This refers to the accurate and appropriate use of grammatical forms and vocabulary. It also includes the range of both grammatical forms and vocabulary. Performance is viewed in terms of the overall effectiveness of the language used.

RANGE: The active use of a range of grammatical forms and vocabulary.

ACCURACY: The accurate use of grammatical forms and syntax.

APPROPRIACY: The appropriate use of vocabulary to deal with the tasks.

■ Discourse Management

This refers to the candidate's ability to link utterances together to form coherent monologue and contributions to dialogue. The utterances should be relevant to the tasks and to preceding utterances in the discourse. The discourse produced should be at a level of complexity appropriate to CAE level and the utterances should be arranged logically to develop the themes or arguments required by the tasks. The extent of the contributions should be appropriate, i.e. long or short as required at a particular point in the dynamic development of the discourse in order to achieve the task.

COHERENCE: The logical arrangement of utterances to form spoken discourse and to develop arguments or themes.

EXTENT: The appropriate length of individual contributions (long or short) to develop the discourse and deal with the tasks.

RELEVANCE: The relevance of contributions to the tasks and to preceding contributions in the discourse.

■ Pronunciation

This refers to the candidate's ability to produce comprehensible utterances to fulfil the task requirements. This includes stress, rhythm and intonation as well as individual sounds. Examiners put themselves in the position of the non-ESOL specialist and assess the overall impact of the pronunciation and the degree of effort required to understand the candidate.

STRESS AND RHYTHM: The appropriate use of strong and weak syllables in words and connected speech, the linking of words, and the effective highlighting of information-bearing words in utterances.

INTONATION: The use of a sufficiently wide pitch range and the appropriate use of intonation to convey intended meanings.

INDIVIDUAL SOUNDS: The effective articulation of individual sounds to facilitate understanding.

Different varieties of English, e.g. British, North American, Australian, etc. are acceptable, provided they are used consistently throughout the test.

■ Interactive Communication

This refers to the candidate's ability to take an active part in the development of the discourse, showing sensitivity to turn-taking and without undue hesitation. It requires the ability to participate in the range of interactive situations in the test and to develop discussions on a range of topics by initiating and responding appropriately. It also refers to the deployment of strategies to maintain and repair interaction at an appropriate level throughout the test so that the tasks can be fulfilled.

INITIATING AND RESPONDING: The ability to participate in a range of situations and to develop the interaction by initiating and responding appropriately.

HESITATION: The ability to participate in the development of the interaction without undue hesitation.

TURN-TAKING: The sensitivity to listen, speak, and allow others to speak, as appropriate.

■ Global Achievement Scale

This scale refers to the candidate's overall effectiveness in dealing with the tasks in the four separate parts of the CAE Speaking test. The global mark is an independent, impression mark which reflects the assessment of the candidate's performance from the interlocutor's perspective.

■ Typical minimum adequate performance

Develops the interaction with contributions which are mostly coherent and extended when dealing with the CAE level tasks. Grammar is mostly accurate and vocabulary appropriate. Utterances are understood with very little strain on the listener.

5

Marking

Assessment is based on performance in the whole test, and is not related to performance in particular parts of the test.

In many countries, Oral Examiners are assigned to teams, each of which is led by a Team Leader who may be responsible for approximately 15 Oral Examiners. Team Leaders give advice and support to Oral Examiners, as required.

The Team Leaders are responsible to a Senior Team Leader, who is the professional representative of Cambridge ESOL for the Speaking tests. Senior Team Leaders are appointed by Cambridge ESOL and attend an annual co-ordination and development session in the UK. Team Leaders are appointed by the Senior Team Leader in consultation with the local administration.

After initial training of examiners, standardisation of marking is maintained by both annual examiner co-ordination sessions and by monitoring visits to centres by Team Leaders. During co-ordination sessions, examiners watch and discuss sample Speaking tests recorded on video and then conduct practice tests with volunteer candidates in order to establish a common standard of assessment.

The sample tests on video are selected to demonstrate a range of nationalities and different levels of competence, and are pre-marked by a team of experienced assessors.

Cambridge ESOL Common Scale for Speaking

The Cambridge ESOL Common Scale for Speaking has been developed to help users to:

- interpret levels of performance in the Cambridge Tests from beginner to advanced
- identify typical performance qualities at particular levels
- locate performance in one examination against performance in another.

The Common Scale is designed to be useful to test candidates and other test users (e.g. admissions officers or employers). The description at each level of the Common Scale aims to provide a brief, general description of the nature of spoken language ability at a particular level in real-world contexts. In this way the wording offers an easily understandable description of performance which can be used, for example, in specifying requirements to language trainers, formulating job descriptions and specifying language requirements for new posts.

LEVEL MASTERY C2 CERTIFICATE OF PROFICIENCY IN ENGLISH:

Fully operational command of the spoken language

- Able to handle communication in most situations, including unfamiliar or unexpected ones.
- Able to use accurate and appropriate linguistic resources to express complex ideas and concepts and produce extended discourse that is coherent and always easy to follow.
- Rarely produces inaccuracies and inappropriacies.
- Pronunciation is easily understood and prosodic features are used effectively; many features, including pausing and hesitation, are 'native-like'.

LEVEL EFFECTIVE OPERATIONAL PROFICIENCY C1 CERTIFICATE IN ADVANCED ENGLISH:

Good operational command of the spoken language

- Able to handle communication in most situations.
- Able to use accurate and appropriate linguistic resources to express ideas and produce discourse that is generally coherent.
- Occasionally produces inaccuracies and inappropriacies.
- Maintains a flow of language with only natural hesitation resulting from considerations of appropriacy or expression.
- L1 accent may be evident but does not affect the clarity of the message.

LEVEL VANTAGE B2 FIRST CERTIFICATE IN ENGLISH:

Generally effective command of the spoken language

- Able to handle communication in familiar situations.
- Able to organise extended discourse but occasionally produces utterances that lack coherence and some inaccuracies and inappropriate usage occur.
- Maintains a flow of language, although hesitation may occur whilst searching for language resources.
- Although pronunciation is easily understood, L1 features may be intrusive.
- Does not require major assistance or prompting by an interlocutor.

LEVEL THRESHOLD B1 PRELIMINARY ENGLISH TEST:

Limited but effective command of the spoken language

- Able to handle communication in most familiar situations.
- Able to construct longer utterances but is not able to use complex language except in well-rehearsed utterances.
- Has problems searching for language resources to express ideas and concepts resulting in pauses and hesitation.
- Pronunciation is generally intelligible, but L1 features may put a strain on the listener.
- Has some ability to compensate for communication difficulties using repair strategies but may require prompting and assistance by an interlocutor.

LEVEL WAYSTAGE A2 KEY ENGLISH TEST:

Basic command of the spoken language

- Able to convey basic meaning in very familiar or highly predictable situations.
- Produces utterances which tend to be very short – words or phrases – with frequent hesitations and pauses.
- Dependent on rehearsed or formulaic phrases with limited generative capacity.
- Only able to produce limited extended discourse.
- Pronunciation is heavily influenced by L1 features and may at times be difficult to understand.
- Requires prompting and assistance by an interlocutor to prevent communication from breaking down.

Appendix 10

Mark Sheets

INDIVIDUAL ORAL INTERVIEW ASSESSMENT SHEET -
LENGUA BII

GROUP _____

NAME _____

RATER

INTERVIEWER

Grammar and vocabulary		
Pronunciation		
Discourse structure		
Interaction		
Global impression /10		

Observations:

Test Pack used _____

Interviewer _____

Rater _____

GROUP SPEAKING TEST ASSESSMENT SHEET –
LENGUA BII

NAME _____

Other candidates interviewed _____

SCORE

Grammar and vocabulary	
Pronunciation	
Discourse structure	
Interaction	
Global (Interviewer)	
Average	

Observations:

Test Pack used _____

Interviewer _____

Rater _____

Appendix 11

Student Self-assessment Mark Sheet

SELF-ASSESSMENT SHEET: SPEAKING - LENGUA BII

GROUP _____

NAME _____

The purpose of this self-assessment sheet is to gather information on **your own perception** of your speaking skills in English. This self-assessment will be contrasted with the mark you receive in the speaking test at the end of the course in June.

With reference to the statements on the Assessment Criteria Sheet, fill in a mark between 1 and 5 in each box. Please be honest in your assessment.

MARK	
Grammar and vocabulary	
Pronunciation	
Discourse structure	
Interaction	

SELF-ASSESSMENT SHEET: SPEAKING - LENGUA BII (INTERVIEW)

GROUP _____

NAME _____

The purpose of this self-assessment sheet is to gather information on **your own perception** of how you think you performed on the interview. This evaluation sheet will be contrasted with the mark given to you by the interviewer. It is purely for research purposes and will not count towards your final mark in the subject Lengua BII.

Using the statements on the Assessment Criteria Sheet, fill in a mark between 1 and 5 in each box. Please be as honest as possible in your assessment.

MARK	
Grammar and Vocabulary	
Pronunciation	
Discourse structure	
Interaction	

SELF-ASSESSMENT SHEET: SPEAKING - LENGUA BII (GROUP
SPEAKING TEST)

GROUP _____

NAME _____

The purpose of this self-assessment sheet is to gather information on **your own perception** of how you performed in the group oral test. This sheet will be contrasted with your mark on the test and the self-assessment sheet you filled in about your general speaking skills in English.

Using the statements on the Assessment Criteria Sheet (Group Oral), fill in a mark between 1 and 5 in each box. Please be as honest as possible in your assessment.

MARK	
Grammar and Vocabulary	
Pronunciation	
Discourse structure and Development of ideas	
Interaction	

Appendix 12

Excel Spreadsheet – Data for the ‘Individual Oral Proficiency Interview’

Oral Interview. Test 1

Student no.	Male/Female	R Grammar	R Pronunt'n	R Discourse	R Interact'n	R Mean	I Grammar	I Pronunciati'n	I Discourse	I Interact'n	I Mean	I Global	SI Grammar	SI Pronunciati'	SI Discourse	SI Interact'n	SI Mean	Test Pack	Quest-1	Quest-2	Quest-3	Quest-4	Quest-5	Quest-6	Quest-7	Quest-8	Quest-9	Quest-10	Quest-11	Quest-12	Quest-13	Quest-14	Quest-15
03	F	2,5	2,5	2,0	2,0	2,3	2,5	2,5	2,5	1,5	2,3	4,5	3,0	2,0	1,0	2,0	2,0	4	4	2	2	2	2	1	3	1	2	3	2	1	3	2	3
04	F	2,0	2,5	2,5	1,5	2,1	2,0	2,5	2,0	1,5	2,0	4,0	2,0	2,0	1,0	3,0	2,0	1	3	2	2	2	3	2	3	2	3	3	3	3	3	3	3
05	M	3,5	3,0	3,0	4,0	3,4	3,5	2,5	4,0	3,0	3,3	6,0	2,0	3,0	2,0	3,0	2,5	9	1	3	2	2	4	0	2	3	4	4	4	4	3	3	4
07	M	4,5	2,5	5,0	5,0	4,3	4,0	2,5	4,0	4,0	3,6	7,0	3,0	2,0	3,0	4,0	3,0	1	3	3	2	3	3	2	3	2	3	3	4	4	4	3	3
13	F	3,5	3,0	3,0	2,5	3,0	2,0	2,0	2,5	3,0	2,4	5,0	3,0	3,0	3,5	3,0	3,1	10	3	3	3	3	3	2	3	3	2	2	3	3	2	3	3
14	F	4,0	4,0	4,0	3,5	3,9	3,5	4,0	3,5	3,5	3,6	7,0	2,0	3,0	3,0	3,0	2,8	3	3	3	2	4	3	2	3	2	3	3	3	3	2	3	3
17	M	2,5	2,5	2,0	1,5	2,1	2,0	3,0	2,0	2,0	2,3	4,5	2,0	2,0	2,0	3,0	2,3	6	4	1	1	2	3	0	2	1	4	4	4	4	3	2	2
18	F	3,0	3,0	3,0	3,0	3,0	2,5	3,5	3,5	3,0	3,1	6,0	3,0	3,0	3,0	4,0	3,3	2	3	3	2	3	3	2	2	2	3	3	3	3	2	2	3
27	F	3,0	4,0	3,0	2,0	3,0	3,0	3,0	2,0	2,5	2,6	5,0	1,0	2,0	2,0	1,0	1,5	9	4	2	2	2	2	3	2	2	3	3	4	4	2	2	3
30	F	2,0	2,5	2,0	1,5	2,0	2,5	3,0	2,5	2,0	2,5	4,5	2,0	2,0	2,0	3,0	2,3	2	4	3	2	2	2	2	2	3	3	3	2	3	2	2	4
32	M	2,5	3,0	3,0	2,5	2,8	2,0	2,5	1,5	2,0	2,0	4,5	2,0	2,0	1,0	4,0	2,3	7	3	2	2	3	3	3	3	4	4	3	4	4	2	4	4
37	F	3,0	3,5	2,5	2,5	2,9	3,0	3,0	4,0	4,0	3,5	6,5	2,0	2,0	2,0	3,0	2,3	5	3	3	2	3	3	2	3	3	3	3	4	4	4	2	3
39	M	4,0	5,0	4,0	5,0	4,5	4,0	4,0	4,0	5,0	4,3	7,5	3,0	3,0	2,0	4,0	3,0	3	2	3	1	2	3	1	2	4	4	4	4	4	2	3	4
40	F	4,5	3,7	4,0	4,5	4,2	3,0	3,0	2,5	2,0	2,6	6,0	3,0	4,0	3,0	4,0	3,5	3	3	2	2	2	2	1	2	1	3	2	3	3	2	3	3
41	F	2,2	2,5	2,5	3,0	2,6	2,5	3,0	2,0	3,0	2,6	3,5	3,0	2,0	3,0	3,0	2,8	2	4	1	1	2	4	1	2	2	3	3	4	4	3	4	4
55	F	3,0	3,5	2,5	3,0	3,0	2,5	3,0	2,5	2,0	2,5	5,5	3,0	4,0	3,0	4,0	3,5	3	3	3	2	2	3	0	3	2	3	3	3	3	2	3	3
59	F	4,0	4,0	4,0	5,0	4,3	3,0	3,0	3,0	5,0	3,5	6,5	2,0	3,0	1,0	2,0	2,0	5	3	3	2	1	1	3	2	3	4	4	3	3	4	3	3
60	F	3,0	3,0	2,0	4,0	3,0	4,0	4,0	3,0	4,0	3,8	7,5	2,0	2,0	1,0	2,0	1,8	3	4	3	2	2	3	2	2	1	2	3	3	3	2	3	3
66	M	3,5	4,0	3,5	3,0	3,5	4,0	4,0	3,5	3,0	3,6	7,5	3,0	4,0	3,0	4,0	3,5	1	1	3	3	2	4	2	4	4	4	4	4	4	2	4	4
69	F	2,0	2,0	1,5	1,5	1,8	2,5	2,5	2,5	3,0	2,6	4,5	2,0	3,0	2,0	2,0	2,3	4	4	1	2	2	3	3	2	2	3	3	3	3	3	2	4
72	M	3,0	3,5	3,0	3,5	3,3	3,0	3,5	3,0	4,0	3,4	6,5	4,0	4,0	4,0	4,0	4,0	10	3	3	2	2	3	3	3	3	3	3	3	3	3	3	2
79	F	2,5	3,0	2,5	2,5	2,6	2,5	3,0	2,5	2,0	2,5	5,0	2,0	3,0	3,0	3,0	2,8	4	3	3	2	2	3	3	3	2	3	3	0	3	2	3	3
82	F	2,5	3,5	3,0	3,5	3,1	2,0	2,5	2,0	1,5	2,0	4,0	2,0	3,0	2,0	3,0	2,5	5	4	2	1	2	3	1	2	3	2	2	3	3	2	2	3
84	F	4,0	4,0	3,0	3,0	3,5	3,0	4,0	4,0	4,0	3,8	6,0	2,0	3,0	3,0	3,0	2,8	4	4	3	2	2	0	3	2	2	3	3	3	3	3	3	3
86	F	2,0	3,0	2,0	3,0	2,5	3,0	3,0	3,0	4,0	3,3	6,0	2,0	3,0	2,0	3,0	2,5	1	3	3	2	3	3	3	2	2	3	3	2	2	3	3	3
89	M	1,5	3,0	1,5	1,5	1,9	1,5	2,0	1,0	2,0	1,6	4,0	5,0	4,0	3,0	4,0	4,0	7	4	3	2	2	3	3	2	2	3	2	3	2	3	4	3
94	F	4,0	4,0	4,0	5,0	4,3	4,0	3,0	4,0	5,0	4,0	7,0	3,0	3,0	3,0	4,0	3,3	9	3	3	3	3	3	2	3	2	3	3	3	3	2	4	4
96	M	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	8,5	5,0	5,0	4,0	4,0	4,5	10	2	4	3	3	4	2	4	4	3	4	2	2	2	2	2
99	F	4,0	5,0	4,0	4,0	4,3	4,0	5,0	4,0	5,0	4,5	8,0	3,0	4,0	3,0	3,0	3,3	2	3	3	2	3	4	1	3	4	4	3	3	3	3	4	4
100	F	3,0	4,0	3,0	2,0	3,0	3,0	4,0	2,0	2,0	2,8	6,0	2,0	1,0	2,0	1,0	1,5	1	3	2	2	2	3	3	3	2	4	4	3	3	2	3	3
103	F	1,0	2,0	1,0	2,0	1,5	1,0	1,0	1,0	1,0	1,0	2,0	3,0	4,0	2,0	3,0	3,0	4	4	3	2	3	3	1	3	2	1	1	3	4	2	2	3
108	M	2,0	3,0	2,0	3,0	2,5	3,0	3,0	3,0	3,0	3,0	6,0	2,0	3,0	2,0	3,0	2,5	7	4	3	2	2	2	3	3	3	3	3	3	3	3	3	3

Oral Interview. Test 1

Student no.	Male/Female	R Grammar	R Pronunt'n	R Discourse	R Interact'n	R Mean	I Grammar	I Pronunciat'n	I Discourse	I Interact'n	I Mean	I Global	SI Grammar	SI Pronunciat'	SI Discourse	SI Interact'n	SI Mean	Test/Pack	Quest-1	Quest-2	Quest-3	Quest-4	Quest-5	Quest-6	Quest-7	Quest-8	Quest-9	Quest-10	Quest-11	Quest-12	Quest-13	Quest-14	Quest-15	
110	F	2,0	3,0	4,0	4,0	3,3	3,0	4,0	3,0	4,0	3,5	6,0	3,0	3,0	3,0	3,0	3,0	6	3	3	2	3	3	2	3	3	3	3	4	4	2	3	3	
111	F	2,0	2,0	3,0	4,0	2,8	2,0	3,0	2,0	3,0	2,5	5,5	3,0	3,0	3,0	3,0	3,0	5	3	3	3	2	3	2	3	3	3	3	3	3	3	3	3	
113	F	1,0	2,0	1,0	2,0	1,5	1,0	1,0	1,0	2,0	1,3	3,0	2,0	2,0	2,0	3,0	2,3	9	3	2	2	3	2	3	2	2	3	3	4	4	2	2	4	
114	F	3,0	4,0	3,0	5,0	3,8	2,0	3,0	3,0	4,0	3,0	5,5	4,0	4,0	3,0	3,0	3,5	9	3	3	2	2	3	2	2	3	3	3	3	3	2	2	3	
125	F	4,0	5,0	5,0	4,0	4,5	4,0	5,0	4,0	4,0	4,3	8,0	3,0	4,0	2,0	3,0	3,0	2	2	3	2	3	3	3	2	3	3	3	4	4	2	2	4	
126	F	2,0	3,0	3,0	3,0	2,8	2,0	4,0	2,0	3,0	2,8	6,0	3,0	2,0	3,0	3,0	2,8	7	2	3	2	3	3	2	3	3	3	3	4	4	4	4	4	
127	F	2,0	2,0	2,0	3,0	2,3	2,0	2,0	2,0	3,0	2,3	4,5	2,0	3,0	2,0	1,0	2,0	9	3	3	3	2	3	2	2	3	2	3	3	3	2	2	3	
128	F	1,0	1,0	1,0	2,0	1,3	2,0	2,0	2,0	3,0	2,3	4,0	2,0	3,0	2,0	2,0	2,3	5	3	2	2	3	3	3	3	2	3	3	2	3	2	2	3	
129	F	4,0	5,0	4,0	3,0	4,0	4,0	4,0	3,0	4,0	3,8	7,0	2,0	3,0	2,0	2,0	2,3	8	2	3	3	2	4	2	3	2	3	3	3	3	3	2	4	4
130	M	3,0	3,0	3,0	4,0	3,3	3,0	3,0	3,0	3,0	3,0	6,0	3,0	3,0	3,0	3,0	3,0	9	2	3	2	3	3	3	3	3	3	3	3	3	3	3	3	
134	F	2,0	2,5	2,0	1,5	2,0	2,0	2,0	2,0	2,0	2,0	4,0	3,0	3,0	3,0	3,0	3,0	9	3	2	2	2	2	3	2	3	3	3	3	3	3	3	3	
137	M	3,0	3,0	2,0	2,0	2,5	2,0	3,0	2,0	2,0	2,3	5,0	1,0	3,0	2,0	2,0	2,0	6	4	2	1	3	2	3	2	2	3	3	3	3	3	3	3	
139	F	1,0	2,0	1,0	2,0	1,5	1,0	2,0	2,0	2,0	1,8	4,0	2,0	4,0	3,0	3,0	3,0	8	3	3	4	3	3	2	2	2	2	2	3	2	3	3	3	
141	F	2,0	3,0	2,0	2,0	2,3	2,0	3,0	2,0	1,0	2,0	4,0	1,0	1,0	1,0	2,0	1,3	10	4	3	2	3	1	1	2	2	3	2	4	4	3	2	3	
142	F	4,0	4,0	4,0	4,0	4,0	4,0	4,0	4,0	4,0	4,0	7,0	1,0	2,0	1,0	3,0	1,8	7	4	0	2	3	3	3	3	3	2	2	3	3	2	2	4	
143	F	4,0	5,0	3,0	3,0	3,8	3,0	4,0	4,0	4,0	3,8	7,0	3,0	4,0	3,0	3,0	3,3	10	3	3	2	2	3	2	3	3	2	3	3	3	2	3	3	
144	F	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0	4,0	2,0	2,0	1,0	2,0	1,8	3	3	1	1	3	3	2	2	2	3	3	3	3	3	3	3	
146	F	2,0	2,0	2,0	3,0	2,3	2,0	2,0	2,0	4,0	2,5	5,0	1,0	2,0	2,0	2,0	1,8	2	3	3	1	2	2	2	2	2	3	3	3	3	2	3	3	
151	F	3,0	4,0	3,0	4,0	3,5	4,0	4,0	3,0	3,0	3,5	6,5	2,0	3,0	2,0	2,0	2,3	7	2	3	3	3	4	2	3	2	3	3	3	2	3	3		

Appendix 13

Mean Comparisons for the ‘Individual Oral Proficiency Interview’

1. Comparación de la Medias Globales

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Mean	51	1,25	5,00	2,9750	,91434
I Mean	51	1,00	5,00	2,8946	,86218
SI Mean	51	1,25	4,50	2,6446	,69659
N válido (según lista)	51				

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Mean - I Mean	,0804	,47763	,06688	-,0539	,2147	1,202	50	,235

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Mean - SI Mean	,3304	,92406	,12939	,0705	,5903	2,553	50	,014

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	I Mean - SI Mean	,2500	,94108	,13178	-,0147	,5147	1,897	50	,064

2. Comparación de la Medias de Grammar and Vocabulary

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. tıp.
R Grammar	51	1,00	5,00	2,8176	1,00692
I Grammar	51	1,00	5,00	2,7451	,91855
SI Grammar	51	1,00	5,00	2,4706	,87984
N válido (según lista)	51				

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	R Grammar - I Grammar	,0725	,59702	,08360	-,0954	,2405	,868	50	,390

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Grammar - SI Grammar	,3471	1,20206	,16832	,0090	,6851	2,062	50	,044

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 I Grammar - SI Grammar	,2745	1,18031	,16528	-,0575	,6065	1,661	50	,103

3. Comparación de la Medias de Pronunciation

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. tıp.
R Pronunt'n	51	1,00	5,00	3,2196	,96727
I Pronunt'n	51	1,00	5,00	3,0686	,91662
SI Pronunt'n	51	1,00	5,00	2,8824	,86364
N válido (según lista)	51				

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	R Pronunt'n - I Pronunt'n	,1510	,61364	,08593	-,0216	,3236	1,757	50	,085

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	R Pronunt'n - SI Pronunt'n	,3373	1,04918	,14691	-,0422	,6323	2,296	50	,026

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	I Pronunt'n - SI Pronunt'n	,1863	1,12677	,15778	-,1306	,5032	1,181	50	,243

4. Comparación de la Medias de Discourse Structure

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Discourse	51	1,00	5,00	2,8039	1,02994
I Discourse	51	1,00	5,00	2,7353	,92926
SI Discourse	51	1,00	4,00	2,3431	,80306
N válido (según lista)	51				

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Discourse - I Discourse	,0686	,74175	,10387	-,1400	,2772	,661	50	,512

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Discourse - SI Discourse	,4608	1,11733	,15646	,1465	,7750	2,945	50	,005

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 I Discourse - SI Discourse	,3922	1,05505	,14774	,0954	,6889	2,654	50	,011

5. Comparación de la Medias de Interaccion

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. tıp.
R Interact'n	51	1,50	5,00	3,0588	1,09839
I Interact'n	51	1,00	5,00	3,0294	1,10640
SI Interact'n	51	1,00	4,00	2,8824	,81602
N válido (según lista)	51				

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Interact'n - I Interact'n	,0294	,81493	,11411	-,1998	,2586	,258	50	,798

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación típ.	Error tít. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	R Interact'n - SI Interact'n	,1765	1,13059	,15831	-,1415	,4945	1,115	50	,270

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación típ.	Error tít. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	I Interact'n - SI Interact'n	,1471	1,25815	,17618	-,2068	,5009	,835	50	,408



Appendix 14

Pearson Correlation Test for the 'Individual Oral Proficiency Interview'

Tablas de Correlaciones (Pearson)

1. De las notas del Rater (R) y del Interviewer (I) en el Oral Interview

Correlaciones

		R Grammar	R Pronunt'n	R Discourse	R Interact'n	R Mean
R Grammar	Correlación de Pearson	1	,801**	,864**	,674**	,933**
	Sig. (bilateral)	,	,000	,000	,000	,000
	N	51	51	51	51	51
R Pronunt'n	Correlación de Pearson	,801**	1	,741**	,530**	,853**
	Sig. (bilateral)	,000	,	,000	,000	,000
	N	51	51	51	51	51
R Discourse	Correlación de Pearson	,864**	,741**	1	,757**	,943**
	Sig. (bilateral)	,000	,000	,	,000	,000
	N	51	51	51	51	51
R Interact'n	Correlación de Pearson	,674**	,530**	,757**	1	,839**
	Sig. (bilateral)	,000	,000	,000	,	,000
	N	51	51	51	51	51
R Mean	Correlación de Pearson	,933**	,853**	,943**	,839**	1
	Sig. (bilateral)	,000	,000	,000	,000	,
	N	51	51	51	51	51
I Grammar	Correlación de Pearson	,812**	,717**	,776**	,654**	,828**
	Sig. (bilateral)	,000	,000	,000	,000	,000
	N	51	51	51	51	51
I Pronunt'n	Correlación de Pearson	,669**	,789**	,703**	,488**	,737**
	Sig. (bilateral)	,000	,000	,000	,000	,000
	N	51	51	51	51	51
I Discourse	Correlación de Pearson	,798**	,665**	,718**	,667**	,798**
	Sig. (bilateral)	,000	,000	,000	,000	,000
	N	51	51	51	51	51
I Interact'n	Correlación de Pearson	,606**	,546**	,615**	,727**	,703**
	Sig. (bilateral)	,000	,000	,000	,000	,000
	N	51	51	51	51	51
I Mean	Correlación de Pearson	,804**	,755**	,784**	,717**	,857**
	Sig. (bilateral)	,000	,000	,000	,000	,000
	N	51	51	51	51	51

** La correlación es significativa al nivel 0,01 (bilateral).

2. De las notas del Rater (R) y de los Alumnos (SI) en el Oral Interview

Correlaciones

		R Grammar	R Pronunt'n	R Discourse	R Interact'n	R Mean
R Grammar	Correlación de Pearson	1	,801**	,864**	,674**	,933**
	Sig. (bilateral)	,	,000	,000	,000	,000
	N	51	51	51	51	51
R Pronunt'n	Correlación de Pearson	,801**	1	,741**	,530**	,853**
	Sig. (bilateral)	,000	,	,000	,000	,000
	N	51	51	51	51	51
R Discourse	Correlación de Pearson	,864**	,741**	1	,757**	,943**
	Sig. (bilateral)	,000	,000	,	,000	,000
	N	51	51	51	51	51
R Interact'n	Correlación de Pearson	,674**	,530**	,757**	1	,839**
	Sig. (bilateral)	,000	,000	,000	,	,000
	N	51	51	51	51	51
R Mean	Correlación de Pearson	,933**	,853**	,943**	,839**	1
	Sig. (bilateral)	,000	,000	,000	,000	,
	N	51	51	51	51	51
SI Grammar	Correlación de Pearson	,194	,233	,280*	,323*	,291*
	Sig. (bilateral)	,173	,099	,046	,021	,038
	N	51	51	51	51	51
SI Pronunt'n	Correlación de Pearson	,258	,348*	,221	,345*	,329*
	Sig. (bilateral)	,068	,012	,119	,013	,019
	N	51	51	51	51	51
SI Discourse	Correlación de Pearson	,238	,202	,276*	,266	,277*
	Sig. (bilateral)	,092	,155	,050	,059	,049
	N	51	51	51	51	51
SI Interact'n	Correlación de Pearson	,265	,203	,329*	,331*	,319*
	Sig. (bilateral)	,060	,153	,018	,018	,023
	N	51	51	51	51	51
SI Mean	Correlación de Pearson	,287*	,299*	,333*	,382**	,367**
	Sig. (bilateral)	,041	,033	,017	,006	,008
	N	51	51	51	51	51

** . La correlación es significativa al nivel 0,01 (bilateral).

* . La correlación es significante al nivel 0,05 (bilateral).

Appendix 15

Excel Spreadsheet – Data for the ‘Group Speaking Test’

Group Oral Test

Student no.	Male/Female	R Grammar	R Pronunciat'	R Discourse	R Interact'n	R Mean	S Grammar	S Pronunciat'	S Discourse	S Interact'n	S Mean	I Global	Test Pack	Quest-III 1	Quest-III 2	Quest-III 3	Quest-III 4	Quest-III 5	Quest-III 6	Quest-III 7	Quest-III 8	Quest-III 9	Quest-III 10	Quest-III 11	Quest-III 12	Quest-III 13	Quest-III 14	Quest-III 15	Quest-III 16	Quest-III 17
02	F	2,5	4,0	2,0	3,5	3,0	3,0	4,0	3,0	4,0	3,5	1,5	1	3	3	2	2	2	3	3	4	3	3	2	2	3	3	3	3	3
03	F	2,5	3,0	3,0	2,5	2,8	2,0	3,0	2,0	3,0	2,5	3,0	11	4	3	2	3	3	3	2	3	2	3	2	3	3	3	3	3	4
04	F	2,5	3,0	2,0	2,0	2,4	1,5	2,0	2,0	2,0	1,9	2,5	11	3	3	2	2	3	3	2	3	2	3	3	3	4	3	3	4	3
05	M	2,5	2,5	2,0	2,0	2,3	2,0	3,0	3,0	3,0	2,8	2,5	2	3	2	1	2	2	3	3	2	3	3	3	1	1	3	3	3	3
07	M	3,5	2,5	4,0	5,0	3,8	4,0	3,0	4,0	5,0	4,0	3,5	11	3	3	2	4	3	3	4	3	2	4	3	3	3	3	4	3	3
12	F	2,5	3,0	3,0	2,0	2,6	3,0	5,0	4,0	5,0	4,3	2,0	1	3	3	3	4	4	3	3	3	2	3	3	3	3	3	3	3	3
13	F	2,5	2,5	2,5	2,0	2,4	3,0	3,0	3,0	4,0	3,3	2,5	11	4	3		2	3	3	3	2	3	4	4	3	3	3	3	3	4
14	F	4,0	3,5	4,0	4,0	3,9	3,0	4,0	3,0	4,0	3,5	4,0	1	4	3	2	3	3	3	3	3	2	3	2	3	3	3	3	3	4
16	F	2,5	3,0	3,0	2,0	2,6	3,0	4,0	3,0	2,0	3,0	2,5	8	4	2	3	1	2	3	3	2	2	4	2	3	3	3	3	3	3
17	M	3,0	3,0	3,0	2,5	2,9	2,0	3,0	2,0	4,0	2,8	3,5	2	3	3	1	2	4	3	3	3	1	3	4	3	3	3	2	3	4
18	F	4,0	4,0	4,0	4,0	4,0	2,0	3,0	3,0	3,0	2,8	3,5	1	3	3	2	3	3	3	4	3	3	4	4	4	4	3	3	3	4
20	F	2,5	3,0	3,0	3,0	2,9	2,0	2,0	1,0	3,0	2,0	2,5	1	4	2	1	2	2	3	4	2	3	4	4	1	1	3	2	3	4
22	F	2,0	3,0	1,5	2,0	2,1	3,0	4,0	4,0	4,0	3,8	2,8	10	3	2	2	2	3	3	3	3	3	4	4	2	2	3	2	3	3
25	F	2,0	2,0	1,5	1,5	1,8	3,0	5,0	3,0	3,0	3,5	1,0	1	3	3	2	2	2	3	3	3	3	3	3	2	2	3	3	3	3
26	F	2,0	3,0	2,0	2,0	2,3	3,0	3,0	2,0	2,0	2,5	2,5	5	3	2	3	2	4	2		2	2	2	2	2	2	3	3	3	4
27	F	3,0	3,0	2,5	3,0	2,9	2,0	2,0	2,0	1,0	1,8	3,0	1	4	2	3	2	3	3	3	2	3	3	3	3	3	3	3	3	3
30	F	2,5	2,5	2,5	2,0	2,4	2,0	3,0	2,0	3,0	2,5	3,0	11	4	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3
31	F	1,8	1,8	1,5	2,0	1,8	3,0	4,0	3,0	3,0	3,3	1,5	8	4	3	2	2	2	4	3	2	3	2	3	2	2	3	3	3	3
33	F	2,0	2,5	2,0	1,5	2,0	2,0	3,0	2,0	1,5	2,1	2,5	10	4	3	2	1	3	2	3	2	3	4	3	3	3	3	3	3	3
37	F	4,0	4,0	4,0	5,0	4,3	3,0	4,0	4,0	5,0	4,0	3,0	11	1	3	3	3	3	4	3	3	3	4	3	3	3	3	3	3	3
38	F	4,0	3,0	3,0	3,5	3,4	2,0	2,0	2,0	4,0	2,5	3,5	1	4	2	2	2	3	3	4	2	4	3	4	3	3	3	2	3	3
39	M	2,5	3,0	3,5	4,0	3,3	3,0	3,0	4,0	5,0	3,8	3,0	9	2	3	2	1	3	4	2	4	3	3	3	2	2	3	2	4	4
40	F	4,0	4,0	4,0	4,0	4,0	3,0	4,0	3,0	4,0	3,5	4,0	6	3	4	3	4	4	4	3	4	4	4	4	4	4	4	4	3	4
42	M	4,0	3,5	4,0	5,0	4,1	3,0	4,0	3,0	3,0	3,3	4,5	8	2	4	3	4	4	4	4	4	3	4	4	3	3	4	4	4	4
45	F	1,5	1,0	1,5	2,5	1,6	3,0	3,0	3,0	5,0	3,5	1,5	7	4	2	2	2	3	3	3	3	3	3	3	3	3	3	2	2	3
46	M	2,5	3,0	3,0	2,0	2,6	3,0	4,0	3,0	3,0	3,3	2,5	6	2	3	2	2	3	3	3	2	2	3	3	2	2	3	3	3	3
52	F	1,0	2,0	2,0	2,0	1,8	2,0	4,0	3,0	3,0	3,0	1,5	2	3	3	2	2	2	3	3	3	3	3	3	3	2	3	3	3	3
53	F	2,5	3,0	2,5	2,0	2,5	2,0	4,0	3,0	5,0	3,5	2,0	7	4	3	3	3	3	3		3	3	2	3	3	3	3	3	3	3
55	F	3,0	3,0	3,0	4,0	3,3	3,0	4,0	3,0	4,0	3,5	3,0	6	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
56	F	3,0	3,0	3,0	2,5	2,9	4,0	3,0	3,0	5,0	3,8	2,0	7	1	3	3	2	3	3	2	3	2	3	3	3	3	4	4	4	3
58	F	2,5	2,5	3,0	3,0	2,8	3,0	4,0	3,0	4,0	3,5	3,0	9	1	3	2	2	3	2	3	3	3	3	3	2	2	4	3	3	3
61	F	3,0	4,0	3,0	4,0	3,5	3,0	4,0	3,0	3,0	3,3	2,0	1	3	3	2	2	3	4	4	3	3	4	3	3	3	2	2	4	4

Group Oral Test

Student no.	Male/Female	R Grammar	R Pronunclat'	R Discourse	R Interact'n	R Mean	S Grammar	S Pronunclat'	S Discourse	S Interact'n	S Mean	I Global	Test Pack	Quest-III 1	Quest-III 2	Quest-III 3	Quest-III 4	Quest-III 5	Quest-III 6	Quest-III 7	Quest-III 8	Quest-III 9	Quest-III 10	Quest-III 11	Quest-III 12	Quest-III 13	Quest-III 14	Quest-III 15	Quest-III 16	Quest-III 17	
64	F	2,5	2,5	3,0	3,0	2,8	3,0	4,0	3,0	3,0	3,3	2,0	5	3	3	2	3	4	3	3	3	3	3	3	3	3	2	3	2	3	
66	M	3,5	3,5	3,0	3,0	3,3	3,0	3,0	3,0	3,0	3,0	3,5	2	4	3	2	2	2	4	4	3	3	4	3	3	3	3	4	3	4	
67	F	1,5	1,5	2,0	2,0	1,8	3,0	3,0	4,0	3,0	3,3	1,8	7	3	2	3	2	3	3	2	3	3	3	3	3	2	2	3	3	3	
69	F	3,0	3,0	2,5	2,0	2,6	2,0	3,0	2,0	2,0	2,3	2,5	10	3	2	1	1	3	3	3	2	3	3	3	4	4	3	2	3	3	
70	F	3,0	3,5	3,0	2,5	3,0	3,0	4,0	3,0	4,0	3,5	2,5	9	2	3	3		4	3	3	3	2	3	2	3	3	3	3	3	3	
71	F	1,0	2,5	1,5	1,5	1,6	3,0	4,0	3,0	3,0	3,3	1,8	6	3	3	2	3	3	3		2	3	3	3	2	2	3	3	4	4	
72	M	2,0	4,0	3,0	5,0	3,5	4,0	4,0	4,0	4,0	4,0	2,0	9	2	3	2	1	3	2	2	3	2	3	3	3	3	3	3	3	4	4
74	F	5,0	5,0	5,0	4,5	4,9	5,0	5,0	4,0	5,0	4,8	4,5	1	1	3	4	3		4		4	4	4	4	3	3	4	3	4	4	
78	F	2,0	3,0	2,0	4,0	2,8	3,0	3,0	3,0	4,0	3,3	2,5	2	1	3	3	3	4	4	3	3	3	3	3	3	3	3	3	3	3	3
79	F	3,0	3,0	3,0	2,0	2,8	3,0	3,0	3,0	3,0	3,0	4,0	3	3	3	2	3	3	3	3	3	2	3	3	3	3	3	2	3	3	
81	M	3,0	4,0	3,0	3,0	3,3	3,0	5,0	3,0	3,0	3,5	2,0	1	2	3	2	3	4	3	3	4	3	4	1	4	4	2	3	3	4	
82	F	2,5	3,0	3,0	4,0	3,1	3,0	4,0	3,0	4,0	3,5	3,0	6	3	3	2	2	3	3	4	3	3	3	3	3	3	4	3	3	3	
83	M	2,0	3,0	3,0	4,0	3,0	3,0	2,0	3,0	4,0	3,0	2,0	1	3	2	3	3	3	2	2	2		3	4	3	3	4	3	3	3	
93	F	1,5	2,0	1,0	1,5	1,5	1,0	3,0	3,0	3,0	2,5	1,5	3	3	2	2	2	2	3	3	2	2	4	4		3	3	2	3	3	
94	F	3,0	3,0	3,0	4,0	3,3	3,0	3,0	3,0	4,0	3,3	4,0	4	4	3	2	4	3	4	4	3	3	3	4	3	3	3	4	3	4	
96	M	5,0	5,0	5,0	5,0	5,0	5,0	5,0	4,0	4,0	4,5	5,0	4	2	4	2	4	3	4	4	4	3	4	4	4	4	4	4	1	2	2
100	F	3,0	3,0	2,5	4,0	3,1	2,0	2,0	2,0	5,0	2,8	2,0	1	4	2	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3	3
104	F	2,5	3,0	2,5	3,5	2,9	3,0	3,0	2,0	4,0	3,0	3,0	4	3	3	2	3	3	4	3	2	2	3	3	3	3	3	3	3	3	2
106	F	3,0	4,0	3,0	2,0	3,0	3,0	3,0	2,0	3,0	2,8	2,5	4	3	2	3	2	3	2	3	2	3	3	2	3	3	3	2	3	3	
107	M	2,5	3,0	2,0	2,0	2,4	3,0	3,0	3,0	4,0	3,3	2,5	10	1	3	3	2	3	4	4		2	3	4	3	3	4	3	3	4	
108	M	2,0	3,0	3,0	2,5	2,6	3,0	3,0	3,0	3,0	3,0	2,5	4	3	3	2	2	3	3	4	3	3	4	3	3	3	3	3	3	3	4
110	F	2,0	2,0	2,0	2,0	2,0	2,0	3,0	2,0	3,0	2,5	2,5	3	3	2	2	2	3		4	2	2	3	3	3	3	2	3	3	3	
111	F	3,0	3,0	3,0	3,5	3,1	3,0	4,0	4,0	4,0	3,8	3,0	7	2	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	4	
113	F	2,0	2,0	2,0	1,5	1,9	2,0	2,0	3,0	3,0	2,5	1,5	3	4	3	2	3	3	3	3	3	2	3	4	3	3	3	3	3	3	
114	F	3,0	3,0	2,5	2,0	2,6	2,0	3,0	2,0	3,0	2,5	3,0	6	2	3	2	3	3	3	3	2	3	4	3	3	2	2	3	2	3	
116	F	2,0	2,5	2,0	1,5	2,0	3,0	4,0	3,0	3,0	3,3	1,8	6	3	3	2	3	3	3	3	3	2	3	3	2	2	3	3	3	3	
117	F	3,0	3,0	3,5	3,0	3,1	3,0	3,0	2,5	3,0	2,9	2,0	4	2	3	2	3	3	4	4	2	2	3	3	3	3	2	3	3	3	
120	M	2,0	2,5	2,5	3,0	2,5	2,0	3,0	3,0	5,0	3,3	2,5	1	2	3	1	3	4	4	4	2	3	3	2	4	4	4	4	4	4	
122	F	2,0	2,5	2,0	3,0	2,4	3,0	4,0	4,0	2,0	3,3	2,5	5	4	3	2	3	3	3	3	3	2	3	3	3	2	3	4	3	4	
123	F	2,5	3,0	3,0	3,0	2,9	2,0	3,0	3,0	3,0	2,8	2,0	3	3	2	3	2	3	4	3	2	3	4	4	3	3	3	3	3	3	
125	F	3,0	4,0	3,5	3,0	3,4	4,0	4,0	3,0	3,0	3,5	4,0	3	3	3	3	3	3	3	3	3	2	3	4	3	3	4	3	3	4	
126	F	2,5	2,5	2,5	3,0	2,6	3,0	4,0	3,0	3,0	3,3	2,0	5	2	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	

Group Oral Test

Student no.	Male/Female	R Grammar	R Pronunciat'	R Discourse	R Interact'n	R Mean	S Grammar	S Pronunciat'	S Discourse	S Interact'n	S Mean	I Global	Test Pack	Quest-II 1	Quest-II 2	Quest-II 3	Quest-II 4	Quest-II 5	Quest-II 6	Quest-II 7	Quest-II 8	Quest-II 9	Quest-II 10	Quest-II 11	Quest-II 12	Quest-II 13	Quest-II 14	Quest-II 15	Quest-II 16	Quest-II 17
127	F	2,0	3,0	2,0	1,0	2,0	3,0	4,0	2,0	4,0	3,3	2,0	5	3	3	2	2	3	4	3	2	1	3	3	3	2	3	3	3	3
128	F	1,0	1,5	1,5	2,0	1,5	2,0	3,0	2,0	3,0	2,5	2,0	8	4	2	3	3	3	4	3	2	2	1	3	3	3	3	3	3	3
129	F	2,0	3,0	3,0	2,0	2,5	2,0	3,0	2,0	4,0	2,8	3,0	5	3	2	2	2	2	3	3	2	3	2	2	2	3	3	3	3	3
130	M	2,5	3,0	2,5	2,0	2,5	3,0	3,0	4,0	4,0	3,5	2,5	5	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4
133	F	2,0	3,0	2,5	2,5	2,5	4,0	3,0	3,0	3,0	3,3	2,5	2	3	2	3	2	3	3	3	2	2	3	3	3	3	3	3	3	3
135	F	2,5	3,0	3,0	3,0	2,9	3,0	4,0	4,0	5,0	4,0	2,0	10	3	2	2	2	3	3	3	2	2	3	4	3	3	4	4	4	3
136	M	2,0	3,0	2,0	3,0	2,5	3,0	4,0	2,0	2,0	2,8	1,8	8	2	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3
137	M	2,0	2,0	2,5	2,5	2,3	2,0	2,0	2,0	3,0	2,3	2,0	8	3	2	2	3	2	3	3	3	2	3	3	3	2	3	3	3	3
138	F	2,0	2,5	2,0	2,5	2,3	2,0	3,0	3,0	3,0	2,8	3,0	4	3	2	2	2	3	3	3	2	2	3	3	3	3	3	3	3	3
139	F	2,0	2,5	2,0	2,5	2,3	3,0	4,0	3,0	4,0	3,5	2,5	2	4	3	4	4	2	4	3	3	3	2	4	3	3	3	4	3	3
140	F	2,5	4,0	2,5	4,0	3,3	3,0	4,0	3,0	4,0	3,5	2,0	1	1	2	3	2	2	4	3	3	3	2	2	2	2	2	2	2	3
143	F	2,5	3,0	3,0	3,0	2,9	2,0	4,0	3,0	2,0	2,8	3,0	2	4	2	2	2	2	3	3	3	2	3	3	3	2	3	3	3	3
144	F	1,0	1,5	1,0	1,5	1,3	2,0	2,0	2,0	3,0	2,3	1,0	8	3	2	2	3	2	3	3	2	2	3	3	3	3	3	3	3	3
151	F	3,0	3,5	3,0	3,0	3,1	2,0	3,0	2,0	3,0	2,5	2,5	10	2	3	3	3	3	3	3	4	3	3	3	3	3	3	3	3	3

Appendix 16

Mean Comparisons for the ‘Group Speaking Test’

1. Comparación de la Medias Globales

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
R Mean	78	1,3	5,0	2,765	,7325
S Mean	78	1,8	4,8	3,141	,5845
I Global	78	1,0	5,0	2,577	,8085
N válido (según lista)	78				

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Mean - I Global	,188	,5782	,0655	,058	,319	2,879	77	,005

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Mean - S Mean	-,376	,7022	,0795	-,534	-,217	-4,725	77	,000

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 S Mean - I Global	,564	,8947	,1013	,362	,766	5,568	77	,000

2. Comparación de la Medias de Grammar and Vocabulary

Estadísticos de muestras relacionadas

	Media	N	Desviación típ.	Error típ. de la media
Par 1 R Grammar	2,562	78	,8013	,0907
S Grammar	2,750	78	,7151	,0810

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 R Grammar - S Grammar	-,188	,8560	,0969	-,381	,005	-1,944	77	,055

3. Comparación de la Medias de Pronunciation

Estadísticos de muestras relacionadas

		Media	N	Desviación tıp.	Error tıp. de la media
Par 1	R Pronunt'n	2,946	78	,7272	,0823
	S Pronunt'n	3,397	78	,7786	,0882

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	R Pronunt'n - S Pronunt'n	-,451	,8486	,0961	-,643	-,260	-4,696	77	,000

4. Comparación de la Medias de Discourse Structure

Estadísticos de muestras relacionadas

		Media	N	Desviación tıp.	Error tıp. de la media
Par 1	R Discourse	2,673	78	,7929	,0898
	S Discourse	2,865	78	,6917	,0783

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	R Discourse - S Discourse	-,192	,8945	,1013	-,394	,009	-1,899	77	,061

5. Comparación de la Medias de Interaction

Estadísticos de muestras relacionadas

		Media	N	Desviación tıp.	Error tıp. de la media
Par 1	R Interact'n	2,821	78	,9999	,1132
	S Interact'n	3,455	78	,9194	,1041

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación tıp.	Error tıp. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	R Interact'n - S Interact'n	-,635	1,0681	,1209	-,875	-,394	-5,248	77	,000

Appendix 17

Pearson Correlation Test – ‘Group Speaking Test’

Tablas de Correlaciones (Pearson)

1. De las notas del Rater (R) y del Student (I) en el Group Oral Test

Correlaciones

		R Grammar	R Pronunt'n	R Discourse	R Interact'n	R Mean
R Grammar	Correlación de Pearson	1	,754**	,856**	,629**	,906**
	Sig. (bilateral)		,000	,000	,000	,000
	N	78	78	78	78	78
R Pronunt'n	Correlación de Pearson	,754**	1	,730**	,584**	,847**
	Sig. (bilateral)	,000		,000	,000	,000
	N	78	78	78	78	78
R Discourse	Correlación de Pearson	,856**	,730**	1	,703**	,925**
	Sig. (bilateral)	,000	,000		,000	,000
	N	78	78	78	78	78
R Interact'n	Correlación de Pearson	,629**	,584**	,703**	1	,851**
	Sig. (bilateral)	,000	,000	,000		,000
	N	78	78	78	78	78
R Mean	Correlación de Pearson	,906**	,847**	,925**	,851**	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	78	78	78	78	78
S Grammar	Correlación de Pearson	,367**	,461**	,472**	,436**	,492**
	Sig. (bilateral)	,001	,000	,000	,000	,000
	N	78	78	78	78	78
S Pronunt'n	Correlación de Pearson	,206	,366**	,255*	,176	,273*
	Sig. (bilateral)	,071	,001	,024	,123	,016
	N	78	78	78	78	78
S Discourse	Correlación de Pearson	,173	,192	,280*	,350**	,292**
	Sig. (bilateral)	,129	,092	,013	,002	,010
	N	78	78	78	78	78
S Interact'n	Correlación de Pearson	,239*	,168	,305**	,383**	,321**
	Sig. (bilateral)	,035	,141	,007	,001	,004
	N	78	78	78	78	78
S Mean	Correlación de Pearson	,323**	,383**	,429**	,443**	,450**
	Sig. (bilateral)	,004	,001	,000	,000	,000
	N	78	78	78	78	78

** . La correlación es significativa al nivel 0,01 (bilateral).

* . La correlación es significante al nivel 0,05 (bilateral).

Resumen en lengua castellana de la tesis doctoral

Resumen en lengua castellana de la tesis doctoral

1. Planteamiento teórico

La evaluación ocupa un lugar fundamental, incluso constituye a menudo el aspecto más relevante, en la gran mayoría de planes docentes de lenguas extranjeras. Por lo tanto, su análisis y validación es un asunto de suma importancia y de gran interés en la investigación de la Lingüística Aplicada a la Adquisición de Segundas Lenguas. Adquirir la capacidad de justificar de manera objetiva las pruebas evaluables y la validez de sus calificaciones forma parte de nuestra tarea como diseñadores y receptores de exámenes, y, en consecuencia, marcarnos el reto de definir los constructos de la lengua y de desarrollar argumentos de validez que se puedan aplicar a la práctica de la evaluación es responsabilidad del docente. Los métodos de evaluación que empleemos deben ser apropiados en función del contexto en el que se usan y deben facilitar calificaciones que reflejen de la manera más precisa posible la actuación lingüística y la habilidad subyacente del discente. Es más, deben reflejar las directrices del enfoque y del contenido del programa de enseñanza y aprendizaje.

Esta preocupación por establecer afirmaciones acertadas y significativas respecto a la actuación lingüística y la habilidad subyacente nos lleva precisamente a considerar el diseño del baremo de evaluación como un aspecto fundamental. Como parte de nuestra tarea docente, solemos dedicar tiempo y esfuerzo a la elaboración de pruebas de evaluación para los cursos que hemos impartido y, por lo general, nuestra atención se centra en la selección de las tareas y los temas del examen, y en las preguntas a las cuales queremos que nuestros alumnos respondan. En raras ocasiones prestamos atención al baremo de evaluación que se pretende emplear en una prueba de competencia lingüística: es algo que ya existe, que se nos

presenta previamente diseñado. Colocamos a nuestros alumnos en una escala del 0 al 10 según estimaciones que son poco o nada fiables, porque no corresponden a unos criterios descritos con anterioridad ni a una definición del constructo de la variable que intentamos medir. Esto nos lleva a comparar seguramente la actuación de los alumnos, de tal manera que son las actuaciones por sí solas, aquellas que se observan durante una determinada sesión de examen, las que nos guían en la aplicación de unos criterios de evaluación asumidos. Parece evidente que este procedimiento sólo nos proporciona información sobre la actuación lingüística observada durante la prueba, y que dicha información no sirve como generalización de la habilidad subyacente, ya que el baremo no contempla una definición fija y estable de lo que se pretende medir con él (definición del constructo), ni tampoco señala cuáles de sus características se han demostrado a través de la actuación.

En el caso de las pruebas orales, esta debilidad se hace todavía más patente, debido a la naturaleza de la destreza que queremos medir. Poco se conoce sobre el rápido y complejo procesamiento cognitivo que tiene lugar en la producción del lenguaje hablado, lo cual limita necesariamente nuestra habilidad de describir los elementos que componen realmente la competencia del habla. A esto hay que añadir la naturaleza efímera del lenguaje hablado: si no se graba, el discurso permanece tan sólo como una idea o recuerdo en la mente tanto del hablante como del oyente. Al contrario de lo que ocurre con la palabra escrita, que permanece estática e invariable, al repetir oralmente una frase de más de quince o veinte palabras pronunciada de manera espontánea, nuestra tendencia primordial de prestar atención al contenido antes que a la forma del mensaje seguramente nos

impide reproducir de manera exacta, palabra por palabra, lo que acabamos de escuchar.

Por lo tanto, y con objeto de poder medir la competencia del habla, es necesario elaborar en primer lugar una descripción de los distintos elementos que componen el acto oral, es decir, proporcionar una definición del constructo que nos posibilite el diseño de un baremo de evaluación óptimo, y diseñar pruebas que nos permitan establecer generalizaciones sobre la habilidad lingüística subyacente más allá de la actuación. Con ello, estaremos construyendo al mismo tiempo un argumento de validez que refuerza el vínculo entre la puntuación asignada en una prueba y su significado. Asimismo, es nuestro deber intentar definir los constructos para las pruebas orales de una manera relevante y comprensible para los alumnos a los que va destinado el diseño de la prueba. De esta manera, nuestra definición de constructo se verá también reforzada por el propósito de la prueba.

Esta definición debería basarse en teorías sobre la competencia lingüística, y debería describir los componentes relacionados, y a la vez diferenciados, que constituyen el constructo del habla. Dos de las pilares más importantes para esta descripción guardan relación con la *competencia estratégica* y la *competencia interaccional*. La competencia estratégica (Bachman, 1990; Bachman y Palmer, 1996; Canale y Swaine, 1980; Canale, 1983) hace referencia a la capacidad cognitiva de manejar la conversación, pero, para poder incluirla en una definición de constructo, sería necesario tener en cuenta los elementos del habla observables que proporcionan evidencia de su uso. Estos elementos se tendrían que describir dentro del marco del baremo de evaluación, al objeto de que los evaluadores y usuarios de la prueba los puedan tener en cuenta como factores diferenciados y cuantificables del constructo.

Aún no existe un consenso generalizado sobre la definición de la competencia interaccional. Algunos investigadores, generalmente los que se interesan por definir la competencia comunicativa, incluyen la competencia interaccional como parte de la anterior, y la consideran una habilidad interna del individuo (Bachman, 1990). Esta postura contrasta con la opinión de McNamara (1997:447), quien la considera un constructo social, relacionada con el comportamiento y con la manera en que los participantes co-construyen el habla durante la interacción conversacional

La investigación actual tiende a alejarse de teorías que abarcan todo tipo de representaciones globales de constructos y que intentan reflejar verdades universales para múltiples situaciones, basándose en el argumento de que tales teorías no pueden proporcionar representaciones coherentes y significativas de constructos que median en la interacción social. En esta línea, Chaloub-Deville (2003) propone una ampliación del concepto de competencia interaccional al incluir el contexto en que ocurre la interacción como parte de la definición del constructo. Este enfoque presupone que si no desarrollamos una teoría que contemple los contextos, aunque la mayoría de las habilidades esenciales de los usuarios del lenguaje sean internas, estables o invariables, no habrá evidencia suficiente para hacer afirmaciones generales sobre las habilidades y actuaciones. La autora argumenta la necesidad de comprender mejor la interacción compleja entre los sistemas estables y los sistemas interaccionales más variables, para poder desarrollar informes de estimaciones sobre la habilidad lingüística subyacente.

Como hemos podido observar, la definición de un constructo es una tarea compleja y en continuo desarrollo, y está estrechamente relacionada con las teorías de la validación. A pesar de que el camino hacia la consecución de una definición

amplia y global del habla aún queda lejos, es de vital importancia ser conscientes de la relevancia de su inclusión en los procedimientos de diseño de nuestras pruebas orales. Estas pruebas deberían basarse por la tanto en argumentos válidos, cuyas fundamentaciones guarden relación con teorías sobre la descripción del lenguaje, la adquisición de una segunda lengua y la medición de destrezas lingüísticas.

Dado que constituyen instrumentos para la recogida de datos sobre la competencia oral, nuestras pruebas orales deberían facilitarnos información de una manera sistemática, mediante tareas u otras técnicas experimentales que se puedan aplicar a otros candidatos y en diferentes sesiones del mismo examen. El resumen de la evidencia (la puntuación) nos debe informar sobre el constructo tal y como lo hemos definido y también debería permitirnos hacer deducciones sobre la actuación del discente en contextos diferentes al de la prueba. Esto nos enfrenta a la vez con el reto de diseñar las tareas y materiales adecuados para nuestras pruebas con la finalidad de obtener una muestra de las características y tamaño apropiados para la evaluación en una situación concreta.

Por lo tanto, queda patente la importancia de establecer nuestro procedimiento evaluador dentro de un marco teórico, si queremos ser sistemáticos, coherentes y teleológicos en la tarea de desarrollar pruebas sobre la competencia lingüística de nuestros alumnos. Este tema constituye desde hace tiempo una preocupación personal, debido a las consecuentes repercusiones en las vidas de los usuarios primarios respecto a la evaluación de sus competencias lingüísticas: los estudiantes que se someten a nuestros exámenes.

El presente estudio se ocupa de la evaluación de alumnos del segundo curso del programa de grado 'Traducción e Interpretación' en la Universidad de Las

Palmas de Gran Canaria (ULPGC), España, donde el alumnado recibe instrucción en inglés como lengua extranjera en la asignatura *Lengua BII*. Las notas y calificaciones que obtienen nuestros estudiantes repercuten directamente en aspectos como la continuidad de sus estudios, la adjudicación de becas estatales, el acceso a otros programas de aprendizaje del mismo nivel o de un nivel superior, tales como títulos de Máster o programas de doctorado. Asimismo, repercuten en su participación en programas de intercambio nacionales, europeos e internacionales y en programas de prácticas de la carrera realizada. Además, es muy probable que estas calificaciones influyan en su estado emocional, su desarrollo personal, su autoestima y su visión general de la vida. La nota media de su carrera determina sin duda la elección de sus futuras vidas profesionales y la incorporación al mundo del trabajo. Debido a este entramado humano y social que impregna tanto los procedimientos académicos como el significado e interpretación de nuestros baremos, surge en definitiva nuestro interés y responsabilidad en todo lo que concierne a la evaluación del alumno.

2. Objetivos de la investigación

Con el presente estudio trataremos de examinar con profundidad algunos aspectos de las cuestiones anteriormente señaladas para proponer posteriormente posibles cambios de nuestros procedimientos actuales de evaluación basados en la constatación empírica. Nuestros planteamientos en la presente investigación se centran en las tres principales áreas de interés relacionadas con el campo de la evaluación de las destrezas del lenguaje oral: 1) **el formato de la prueba**, 2) **la evaluación y el baremo evaluador**, y 3) **el papel de la autoevaluación en la enseñanza y el aprendizaje**.

2.1 Formato de la prueba

En cuanto al formato de la prueba, nuestro enfoque consistirá en contrastar el uso de una **entrevista oral individual** entre un candidato y un entrevistador (y que incluye la presencia de un evaluador independiente como variable de control), con un **examen oral en grupo** en que los estudiantes se presentan en grupos de tres e interactúan entre ellos durante la prueba. Para este último formato, contaremos con la presencia, por un lado, de un interlocutor responsable de iniciar y de conducir la prueba, y, por otro, de un evaluador objetivo que no participará en la interacción y cuya única responsabilidad es la de evaluar la producción oral de cada estudiante. Intentaremos descubrir si el formato de la prueba en grupo causa un estado de menor ansiedad en los alumnos, ya que se encuentran acompañados y apoyados por sus compañeros, y también si, desde la perspectiva de un observador objetivo (evaluador), es más fácil evaluar la interacción gestionada por el entrevistador. Fulcher (2004:186) cita un estudio inédito del *University of Cambridge Local Examinations Syndicate*, que demostraba que, en una prueba oral llevada a cabo en pareja, los turnos de los candidatos a examen se incrementaron, y el tiempo dedicado por el interlocutor a hablar se redujo sustancialmente cuando se contrastaba con el formato de la entrevista individual. Partimos de la suposición de que este hecho incide en el formato de la prueba en grupo y que además se le brindará la oportunidad a los alumnos de utilizar un abanico más amplio de estructuras y funciones lingüísticas en esta situación. Esto último es mucho más probable que se produzca en una prueba oral en grupo que en una entrevista individual, dada la negociación de significado característica de la interacción entre un grupo de personas, lo que a su vez impulsa la adquisición de la segunda lengua. Swain (2001:274) afirma que “los diálogos construyen procesos cognitivos y

estratégicos que a su vez construyen la actuación del alumno, información que puede ser de gran relevancia para dar validez al significado que atribuimos a las puntuaciones derivadas de las pruebas”. Esto quiere decir que el formato de la prueba oral en pareja o en grupo es capaz de generar actuaciones lingüísticas que nos permiten evaluar constructos mucho más complejos que los producidos en el formato de la entrevista individual.

2.2 Evaluación

Nuestra segunda área de interés trata de la **evaluación** y del problema que supone la definición de un constructo que reconozca la construcción cooperativa del discurso y de su sentido, y, a su vez, que la muestra que se ha de evaluar se ha producido a través de la interacción de participantes, posiblemente con la colaboración y la ayuda de todos. Con estas consideraciones, diseñaremos un **baremo de evaluación** basado en una definición del constructo que propone la interacción como uno de sus componentes y, al mismo tiempo, trataremos de recoger la descripción de otras características del habla que permitan la evaluación de una actuación individual dentro de la situación del grupo. Al implementar este baremo, nos interesa ver qué hacen los evaluadores al aplicar los criterios descritos y al adjudicar una puntuación. Observaremos si estos evaluadores se mantienen objetivos en su evaluación o si, por el contrario, interiorizan el baremo y utilizan su interpretación individual y personalizada para puntuar la actuación del alumno.

2.3 Auto-evaluación

Nuestra tercera área de interés tiene que ver con la utilidad que encuentran nuestros alumnos en las puntuaciones que reciben al final del proceso de

evaluación de acuerdo con el sistema actual (la escala universal de 0 a 10) y si es posible obtener una repercusión sobre su motivación y aprendizaje al incluirles en un proceso de **autoevaluación** en el que utilizan la misma escala descriptiva de evaluación empleada por los evaluadores. En este último caso intentaremos averiguar si, al proveer a los alumnos del mismo baremo que se usará para evaluar sus destrezas al hablar inglés, estos sujetos consideran que la autoevaluación puede ser una herramienta útil a la hora de aprender y mejorar su competencia lingüística. También nos aproximaremos a la cuestión de la objetividad y certeza de su autoevaluación, y al hecho de si estos dos aspectos deberían formar parte de su nota final para la asignatura que cursan, Lengua BII. Asimismo recogeremos las opiniones de los profesores/evaluadores sobre estos dos aspectos de la autoevaluación del alumno para determinar su grado de coincidencia o divergencia, con el objetivo final de comparar las puntuaciones del alumnado con las del evaluador en cada sesión de prueba realizada. Nos interesa averiguar en este sentido si el nivel de correlación demostrado entre las puntuaciones apoya o no la inclusión de la autoevaluación en nuestro programa de estudio actual, y, en caso afirmativo, nos interesa saber cuáles son los pasos preliminares que habría que tomar para comenzar a implementarlo en futuros programas de aprendizaje y evaluación.

3. Planteamiento metodológico

Tal y como señalamos anteriormente, el origen y motivo de la presente investigación tiene sus raíces en la detección de ciertas carencias en los procedimientos que se están llevando a cabo para evaluar la competencia oral en la asignatura de Lengua BII del segundo año de la carrera de Traducción e

Interpretación de la Facultad de Traducción e Interpretación de la Universidad de Las Palmas de Gran Canaria, en la que el inglés constituye la primera lengua extranjera de los estudiantes. Estas carencias guardan relación con los conceptos de validez y fiabilidad, dos aspectos relacionados entre sí y esenciales para la evaluación, y que representan la compatibilidad del examen con el programa de la asignatura impartida durante el año académico. En este sentido, cabe plantearse algunas cuestiones relacionadas con la validez y fiabilidad de la entrevista oral como procedimiento adecuado para la evaluación de la competencia oral de los estudiantes. La más evidente consiste en la falta de una definición apropiada del constructo del habla así como de una descripción más precisa del nivel de competencia requerido para aprobar el examen. Sin estas directrices, nunca podremos estar seguros de que todos los candidatos a examen se están evaluando de la misma manera, ya que confiamos la evaluación únicamente a las creencias subjetivas e interiorizadas del examinador, a las impresiones que suscita la oralidad y a una idea subjetiva de cómo actúan los candidatos en la entrevista. Se trata, más bien, de la implicación de aspectos socioafectivos - como la empatía con el punto de vista de un candidato o el grado de conocimiento que tengamos de él- que de factores relacionados con el uso de la lengua o el nivel de competencia.

Otro hecho que cabe señalar sobre el procedimiento de la entrevista es la falta de una estructura formal o estandarizada de la prueba: a cada candidato se le interroga de manera arbitraria, se emplean textos no estandarizados y el examen se desarrolla según las respuestas dadas. Podríamos sostener que este modo de proceder es, en algunos aspectos, reflejo de conversaciones auténticas (si bien existen muchos tipos de “conversación” que se definen por características como la situación, el conocimiento del tema, las estructuras sociales y de poder, etc.), pero una

conversación, por definición, no constituye un examen. Para considerarla un examen como tal, hace falta que tenga ciertas características que conduzcan a su evaluación de una manera estandarizada (una definición del constructo que hay que medir y unos descriptores), que sean aplicables a otro perfil de candidatos que realizan el mismo examen (las actividades orales pueden variar en contenido, pero no deben cambiar el procedimiento) y cuyos resultados puedan generalizarse a la competencia global implicada (una descripción de la manera en que la definición del constructo aluda al uso auténtico de la lengua y a la habilidad subyacente). Sin estas características, el examen no puede ser validado. Sin embargo, esto no quiere decir que un examen no pueda reproducir algunas de las características del uso lingüístico real, pero su función primaria siempre será la de una herramienta de medición y, como tal, presentará ciertas limitaciones en cuanto a su autenticidad como vehículo para comprobar el uso de la lengua. El hecho de que el examen sea o no sea auténtico, es decir, que sea coherente con el programa de estudios impartido y con la ponderación idónea del aprendizaje y progreso gradual, es otra cuestión, tal y como vimos anteriormente.

Otro factor importante es el sistema de puntuación empleado para evaluar la actuación de los estudiantes. La ULPGC, así como prácticamente todas las instituciones universitarias españolas, emplea el baremo de puntuación de 0-10 como única vía posible de calificación dentro del sistema educativo oficial, de manera que todos los estudiantes –independientemente de que estudien Filología Moderna, Medicina, Derecho o Ingeniería Naval- obtienen una puntuación según el baremo siguiente: 0 – 4.9 – suspenso; 5 – 6.9 – aprobado; 7 – 8.9 – notable; 9 – 9.9 – sobresaliente; 10 – Matrícula de Honor (lo que supone matrícula gratuita en una asignatura del año académico siguiente).

Con ello se aprecia que la distribución entre las puntuaciones es desigual y desequilibrada, puesto que la primera (con 50 puntos posibles) indica únicamente que no se ha conseguido el nivel requerido. Al *aprobado* y al *notable* se les adjudican meramente 20 puntos en comparación, mientras que el *sobresaliente* obtiene solo 10 puntos posibles y la matrícula de honor únicamente uno. Estas son las únicas referencias que estima la propia universidad, y todos aquellos responsables de establecer dichos baremos en las evaluaciones se ven en la obligación de idear sus propios mecanismos de aplicabilidad, los cuales se basan necesariamente en el criterio e interpretación personal de sus significados.

La interpretación del 10 como puntuación resulta reveladora: aunque la puntuación de 10 se pueda conseguir en los exámenes de Matemáticas, ciencia supuestamente racional y objetiva, es cuestionable el concepto de perfección, el cual constituye una meta idealizada e inalcanzable dentro del sistema educativo universitario. Es más, guarda relación con el progreso, el descubrimiento, el análisis crítico y el planteamiento de nuevas ideas (estén o no estén de acuerdo los profesores/examinadores con la opinión de los estudiantes), y con la indagación de lo que ya se sabe en relación con su posible repercusión en el estudio futuro; cuanto más sabemos, mayor entendimiento obtenemos sobre cuánto nos queda por saber. Con tantos factores sin resolver, ¿cómo es posible obtener un 10 absoluto?

Otros, sin embargo, consideran que la reproducción fiel de lo que han impartido durante el curso como *input* para el estudiante representa una entidad cuantificable, y, por tanto, no les supone un obstáculo adjudicar una puntuación numérica que indique el porcentaje de información correcta retenida por el estudiante en los trabajos, exámenes y pruebas. En este caso, el 10 es una puntuación perfectamente aceptable. Otros son de la convicción de que los

candidatos a examen necesitan ser juzgados según lo que se espera de ellos en una situación dada y en un cierto nivel previamente definido. En este último caso también el 10 es posible, aunque constituye una puntuación inusual y excepcional. La cuestión que se plantean estos examinadores es si hay algo más que puede esperarse del estudiante en cuanto al nivel y a las circunstancias en las que se le está examinando.

También es interesante destacar el modo en que los profesores emplean las puntuaciones en toda la universidad. En *Lengua BII* usamos únicamente las puntuaciones de números enteros y el decimal (.5), mientras que en *Lengua A*, *Lengua BI* y *Lengua BIII* se utiliza toda la escala gradual de puntuación. La puntuación más baja dada en *Lengua BII* es de 3.5 (siendo una puntuación que se adjudica en muy raras ocasiones), pero es el 4 la nota prototípica en los casos en que el candidato no ha alcanzado el nivel requerido, puesto que se considera una señal adecuada de insuficiencia y tiene el mismo efecto que las puntuaciones más bajas. Otros profesores de otras asignaturas de la Facultad de Traducción e Interpretación adjudican sin embargo puntuaciones oficiales de 1.8 o 2.3, las cuales indican que hace falta mucho más esfuerzo y mejora para alcanzar el aprobado. Para estos profesores/examinadores, las puntuaciones muy bajas representan la necesidad de que el estudiante estudie y aprenda mucho más si quiere aprobar el examen de una asignatura.

Como resumen de estas observaciones, podemos señalar que el profesorado de la Universidad de Las Palmas aplica los baremos oficiales según sus propios criterios, y que estos baremos no contienen significados objetivos de evaluación. Otra paradoja es la puntuación media que obtiene el estudiante de todos los años cursados de carrera, la cual comprende entre 30 y 40 asignaturas en total, y es fruto

del conjunto de puntuaciones obtenidas de 30 o 40 profesores diferentes, de los que cada uno aporta su propia interpretación personal a dicho baremo de puntuación. La validez de la media basada en puntuaciones que constituyen sólo en su apariencia elementos pertenecientes a un sistema único es cuestionable, pero la forma de conseguir unificar el sistema para aportarle más autenticidad y fiabilidad requiere de una extensa investigación que excede los límites de la presente tesis doctoral. El cometido principal en este caso es la necesidad de cuestionar y reflexionar sobre las acciones que llevamos a cabo al adjudicar las puntuaciones que repercuten en la vida de los demás y que pueden tener gran relevancia en este sentido. Por ello, es nuestra responsabilidad acometer apropiadamente dichas acciones y disipar nuestro exceso de confianza en un sistema tradicional ampliamente aceptado por todos, lo que no significa necesariamente que sea el más válido y fiable.

4. Diseño del estudio

Por ese motivo, nuestro presente estudio trata de comparar el método tradicional de evaluación de la competencia oral, la 'Entrevista individual de competencia oral', que se viene aplicando en la Facultad de Traducción e Interpretación de la ULPGC hasta la fecha, con un nuevo tipo de examen oral, en el que los estudiantes son evaluados en grupo con otros candidatos que realizan el mismo examen con el mismo nivel, es decir, la 'Prueba oral de grupo'. Al objeto de cotejar la validez y fiabilidad de ambos tipos de examen oral, la 'Entrevista individual de competencia oral' y la 'Prueba oral en grupo', y sus correspondientes implicaciones socioafectivas, iniciamos un estudio basado en la evaluación de la competencia oral de los estudiantes mediante la aplicación de ambos tipos de

pruebas durante un intervalo de tiempo relativamente corto (aproximadamente 6 semanas) para comprobar si se mostraban cambios en la manifestación externa de su habilidad oral.

Tras la realización de cada tipo de examen, a los candidatos se les pedía cumplimentar una hoja de autoevaluación en la que expusieran sus propias percepciones sobre sus actuaciones en la prueba, y, una vez obtenidos sus resultados, debían responder un cuestionario sobre el examen mismo y su experiencia con dicho examen.

Previamente a esta tarea, los estudiantes debían reflexionar sobre sus opiniones acerca de su propia competencia oral en inglés en circunstancias exteriores a las del examen en la misma hoja de autoevaluación. Esta autoevaluación se llevó a cabo tras una actividad oral realizada en el aula que además pretendía familiarizar al alumnado con los criterios que posteriormente se emplearían para autoevaluar su actuación en el examen.

Una semana antes de la realización de cada prueba oral, a los entrevistadores se les entregaba una carpeta que contenía una breve descripción de las cuatro categorías de la competencia oral que iban a ser evaluadas, los criterios de calificación y planillas de puntuación, las instrucciones sobre el procedimiento del examen, una selección de preguntas para la fase introductoria de la prueba y una serie de materiales que se usarían para los exámenes individuales. De los examinadores se requería que se familiarizaran con las instrucciones y el material antes de comenzar cada sesión de evaluación.

Las carpetas de materiales entregadas contenían fotocopias del texto para los candidatos a examen (una en el caso de la entrevista y tres en el caso del examen oral en grupo) y una copia extra como referencia para el examinador, junto

con fotocopias con preguntas relacionadas con el tema del texto. La copia del entrevistador de la hoja de preguntas contenía una pregunta de apoyo adicional, que podía utilizarse a discreción del entrevistador, y que se aplicaría en el caso de que el examen finalizara antes de lo previsto debido a la incapacidad de los candidatos de continuar la interacción, o si un estudiante producía una muestra considerablemente más pequeña a la de sus compañeros en la exposición oral.

Cada carpeta de materiales fue diseñada para incitar a los candidatos, a través de las preguntas, a expresar sus opiniones sobre un tema controvertido y a centrar su atención en las distintas perspectivas en que podría ser concebido. El tipo de texto del tema era un artículo periodístico auténtico de publicación reciente y de corta extensión, y se esperaba que los estudiantes estuviesen familiarizados hasta cierto punto con todos los temas elegidos, ya sea por su conocimiento cultural general o por su identificación personal con el contenido, al ser similar con alguna situación de su ámbito de experiencia. En ambos formatos de examen se informaba a los candidatos, antes de la realización del examen, de que el texto solo podía emplearse como comodín para la discusión, y que no era necesaria una comprensión profunda y minuciosa de dicho texto.

Dado que los examinadores iban a adoptar los papeles tanto de interlocutor como de evaluador, fue necesario instruirlos a todos por igual en ambos roles antes de que las distintas sesiones del examen tuviesen lugar. De esta manera, nos reunimos con los examinadores y discutimos las categorías objeto de evaluación y los baremos de evaluación, los dos procedimientos de examen, tal y como se exponían en las instrucciones del examinador, y el nivel del examen (definido como “nivel avanzado” o como el nivel C1 del Marco de Referencia Común Europeo).

Los examinadores tenían que realizar un simulacro del examen con otros estudiantes que no formaban parte del examen, para controlar cómo llevaban a cabo la ejecución y evaluación de los exámenes. De esta manera, el grupo de examinadores discutiría y analizaría las puntuaciones, al objeto de establecer un baremo estándar común para las sesiones de las pruebas.

4.1 Entrevista individual de competencia oral

En función del enfoque evaluador común que se realiza en el examen tipo “entrevista individual”, en el que el entrevistador formula preguntas y los estudiantes tienen que contestarlas, a los estudiantes se les preparaba de antemano para la entrevista oral individual solamente en el sentido de que se les informaba de que el examen se basaría en un texto que deberían leer antes de entrar en el aula del examen y que únicamente se emplearía como apoyo orientativo del tema de la discusión. Asimismo, no era necesario realizar una demostración práctica de la entrevista para saber lo que ocurriría durante el examen, aunque sí se les aclaraba que la comprensión de lectura no formaba parte de los objetivos del examen. Dado que los estudiantes realizarían el examen de manera voluntaria, se les aclaraba que simplemente se trataba de un simulacro de prueba previo al examen final de *Lengua BII*.

4.2 Prueba oral de grupo

En contraste con la entrevista, y de cara a la prueba oral de grupo, los estudiantes realizaron una sesión previa en el aula con el fin de prepararse para el examen de la asignatura *Lengua BII*. Se distribuyeron fotocopias del mismo texto, con preguntas adicionales, y la clase pudo comprobar una demostración del

examen en grupo formado por tres alumnos voluntarios y un profesor como interlocutor (no había presente un evaluador ni se adjudicaban puntuaciones). La idea consistía en entablar una discusión en clase sobre el examen, con el profesor presente, en la que los estudiantes formulaban preguntas relacionadas con posibles dudas sobre la elaboración de la prueba. El profesor distribuía a continuación una selección de carpetas de materiales y todos los estudiantes ponían en práctica como mínimo dos exámenes orales diferentes en grupos de tres. La finalidad del ejercicio consistía en averiguar si la familiarización con el formato del examen, combinada con el apoyo de los compañeros, reduciría la ansiedad de los estudiantes y, por consiguiente, repercutiría en su actuación. El hecho de brindar al estudiantado la posibilidad de observar lo que ocurre exactamente en el examen y proveerles de una práctica de realización de uno o más exámenes, pretendía incrementar la autoconfianza en la ejecución de la prueba y una mejor comprensión de lo que se espera de ellos durante la interacción con los otros compañeros. La familiarización gradual con los criterios de evaluación extraídos de las dos autoevaluaciones ya realizadas debería consolidar el conocimiento de lo que sería evaluado y la manera en que los examinadores analizaban sus actuaciones. Por otro lado, este procedimiento guarda un estrecho vínculo con las otras pruebas escritas finales para la asignatura Lengua BII, en las que se realizan exámenes piloto durante el segundo cuatrimestre con la finalidad de familiarizar al estudiantado con el formato de examen y animarlos a utilizar estrategias apropiadas en cada parte del examen.

4.3 Baremos de evaluación

Al objeto de saber cómo puntuar la competencia oral de los candidatos, se consultaron diferentes baremos de evaluación de un nivel similar (ARELS, Trinity,

y Cambridge ESOL). La finalidad de dicha consulta era informarnos sobre los diferentes enfoques existentes para evaluar el constructo del habla. Estos enfoques vienen resumidos a continuación.

4.3.1 ARELS Higher Certificate Examination in Spoken English and Comprehension

Es importante destacar que el primero, el *ARELS Higher Certificate Examination in Spoken English and Comprehension*, no se emplea para una situación individual y no se puntúa en tiempo real, sino que consiste en una grabación en laboratorio con las respuestas recogidas en una cinta para la posterior evaluación que llevan a cabo dos evaluadores independientes como mínimo. Este modelo resulta inadecuado en su aplicación a nuestra situación evaluadora, dado que tenemos que examinar a 120 candidatos en la asignatura de Lengua BII, y nuestro laboratorio solo dispone de 19 plazas. Asimismo, disponemos de un número muy escaso de personal destinado a la gestión y corrección de los exámenes. Sin embargo, no hemos descartado estos modelos, puesto que resultan de interés para identificar, en sus baremos, los puntos de coincidencia útiles y susceptibles de aplicación a nuestra situación evaluadora, y para extraer la base teórica apropiada para el procedimiento de evaluación.

Es relevante destacar que las calificaciones finales – Suspenso (*Fail*), *Aprobado (Pass)*, Mérito (*Credit*) y Distinción (*Distinction*) – se alcanzan en función de un porcentaje calculado a partir de las planillas de puntuación de los candidatos. Sin embargo, se señala lo siguiente: “estos criterios pueden funcionar en ocasiones de una forma muy arbitraria, de manera que los evaluadores deben dar además una evaluación global, independientemente del total del porcentaje”. Esto se debe hacer antes de sumar el total de la puntuación para cada parte de la prueba.

Esta constatación parece cuestionar la objetividad de los criterios de evaluación o parece destacar que podría haber deficiencias subyacentes en la planilla de puntuación (*marking key*). El hecho de que el evaluador pueda obtener una impresión general de la actuación de un candidato que difiere, con respecto a la puntuación, de la calificación final conseguida según las notas otorgadas en el examen, sugiere la hipótesis de que la aplicación del criterio ha sido incorrecta o de que los mismos criterios no se pueden interpretar con facilidad.

Cualquiera de estas circunstancias es posible, pero un análisis preciso de los criterios tiende por lo general hacia el último caso. Cada sección del examen tiene un método diferente de puntuación, se centra además en distintos aspectos del constructo y usa un baremo diferente; algunas partes del examen se puntúan del 0-1 y otras del 0-12, intercalándose además una escala diseñada a partir de otros procedimientos de evaluación. De esta manera se crea un complejo sistema de puntuación que será el punto de referencia constante de los evaluadores cuando escuchen y corrijan las verbalizaciones del estudiantado mediante las cintas. El hecho de que haya descriptores para algunas de las puntuaciones y no para otras, resulta en un principio confuso, especialmente cuando se aplica el baremo 0-12 en la primera sección, ya que contiene un número de puntos inusual para cualquier baremo, por lo que es probable que la mayoría de los evaluadores la conviertan en un porcentaje para poder usarla con precisión. Esto significaría, a pesar del previo aprendizaje de estandarización de los baremos, que los evaluadores utilizarían una versión del baremo según su idiosincrasia y subjetividad, a la que cada individuo le asignaría su propia interpretación. En este sentido sería comparable al baremo de 0-10 de la ULPGC, en la que no existen, previamente a su aplicación, otros significados externos a los de suspenso o aprobado.

El procedimiento de evaluación es loable en su amplio intento de aislar y evaluar de manera detallada muchos aspectos del constructo del habla, pero la complejidad de su estructura y la necesidad de hacer una constante y detallada referencia a las instrucciones conllevan un gasto considerable de tiempo para ser aplicado con aceptable precisión, y además resulta impracticable en nuestra situación debido a las circunstancias administrativas poco propicias, dadas por un elevado número de estudiantes y un número muy pequeño de examinadores.

4.3.2 Trinity Grade Examinations in Spoken English for Speakers of Other Languages

Este es un tipo de examen independiente, y no forma parte de un programa de estudios que combine destrezas escritas y orales. La serie de pruebas emplea un mismo formato y se compone de una serie de doce exámenes de grado progresivo divididos en cuatro amplios estadios (*Initial, Elementary, Intermediate* y *Advanced*) que oscilan desde un nivel bajo de competencia (Grade 1) hacia un nivel avanzado de competencia que se aproxima a la habilidad en la primera lengua (Grade 12).

No obstante, este concepto de “*first language ability*” (habilidad en la lengua materna) no está claramente definido en el programa, aunque el criterio de evaluación parece indicar que el concepto impreciso y ampliamente aceptado de “hablante nativo de la lengua culta” (*educated native speaker*) subyace en los descriptores conceptuales de la habilidad. En definitiva, son constataciones de los criterios de evaluación observados en el grado 12, tales como “responder apropiadamente con seguridad y soltura en todo momento”, “contenido totalmente apropiado a todas las contribuciones conversacionales” o “evidencia de estrategias de iniciación y control de la conversación” u “organización competente del

contenido de las contribuciones conversacionales”, que nos indican que se trata de una clase de hablante nativo experto.

Cada grado se evalúa según cuatro áreas del constructo de habla definidas por los propios descriptores: “soltura” (*readiness*), “pronunciación” (*pronunciation*), “uso” (*usage*) y “focalización” (*focus*). Los doce niveles poseen diferentes descripciones sobre la capacidad que se espera de los candidatos en estas cuatro categorías para pasar de grado. La “soltura” incluye la propiedad de comprender y responder adecuadamente, controlando la fluidez de la conversación y tomando la iniciativa, un aspecto que se incluye en los estadios más elevados. La “pronunciación” se vincula a la producción de los sonidos individuales, así como a los patrones de entonación y acento. El “uso” incluye la precisión gramatical y léxica y la “focalización” considera la conveniencia y organización del contenido del discurso de los candidatos. Por lo tanto, estos criterios parecen cubrir todas las áreas que se señalan en los baremos para otros exámenes y pruebas, si bien existe una ligera diferencia en las terminologías empleadas en los enunciados de las categorías.

Por desgracia, no se indican detalles específicos sobre el desglose de las puntuaciones, y solo se sabe que los candidatos obtienen un informe de evaluación y una puntuación sobre 100, donde el 85+ equivale a una calificación de “Pass with Distinction”, el 75-84 a “Pass with Merit” y el 65-74 a “Pass”. En este baremo observamos que no se ha tenido en cuenta la puntuación de aprobado tradicional de 50%, y que los candidatos deben obtener como mínimo un 65% de los objetivos requeridos para conseguir el certificado. Esto podría tener su causa en el hecho de que los estadios se basan en los criterios evaluadores procedentes del Marco Común de Referencia del Consejo de Europa, que emplea el concepto “*can do*” en

sus constataciones para definir los distintos niveles. Obviamente, si un candidato solo puede realizar la mitad de los objetivos requeridos en la definición de un nivel, no puede considerarse realmente que haya alcanzado dicho nivel.

El planteamiento de los doce niveles de competencia oral, que significa el progreso de un grado hacia el siguiente y para el que se va añadiendo nuevo material y que incluye al mismo tiempo lo que se ha hecho previamente, es un sistema complejo, y, si retomamos las definiciones de los criterios de evaluación, encontraremos que es difícil en ocasiones establecer diferencias de puntuación de un grado a otro. A modo de ejemplo, la diferencia hecha por el examinador entre el grado 11 y el grado 12 respecto a la “soltura”, consiste en que su definición viene marcada por distintos matices, tales como “comprender cambios de registro” (Grade 11) y “comprender cambios de registro y acentuación” (Grade 12). Incluso en la categoría de “soltura” observamos dos matices a la vez: “controlar y mantener la fluidez de la conversación con facilidad” (Grade 11), y “controlar y mantener la fluidez de la conversación de una manera natural” (Grade 12). En el caso de la “pronunciación”, encontramos que los “sonidos ocasionales” han sido reemplazados por los “sonidos poco frecuentes” que “se desvían de una pronunciación globalmente inteligible”, y en la “focalización” apreciamos un cambio de “adecuado” a “competente” en la organización del contenido en las contribuciones conversacionales. Aparte de esto, no se señalan otras diferencias en los criterios de evaluación para estos dos niveles. Ante estas sutiles diferencias, resulta cuestionable la justificación de que un comité de examen fomente la realización gradual de los exámenes en función de los respectivos niveles, con el consecuente pago de los estudiantes cada vez que realizan una prueba. No obstante, del estudio final se deduce que escribir criterios de evaluación que diferencien

claramente entre niveles y grados de conocimiento constituye una tarea ardua y extremadamente compleja, la cual requiere de un extenso estudio de los baremos existentes y de una consideración muy reflexiva. Para que un baremo sea significativo, hace falta elaborarlo cuidadosamente y ponerlo a prueba las veces que sea necesario antes de que sea completamente operativo en cualquier sector de la educación.

Asimismo es cuestionable que un entrevistador/evaluador pueda prestar atención al uso que muestra el candidato de “la amplia gama de oraciones condicionales” (Grade 11) en contraposición a las oraciones condicionales de segundo y tercer tipo, a las condicionales con *unless* y *could have* + participio (introducidas a partir del nivel 7 en adelante) en los estadios más altos de competencia, donde el candidato debe realizar una actuación cercana a la habilidad de la primera lengua. Resulta extremadamente difícil procesar el contenido de lo que se está diciendo al mismo tiempo que se escucha detenidamente la serie de estructuras gramaticales que se están usando al hablar, y no resulta nada fácil tanto para los nativos como para los casi nativos del inglés usar durante una conversación natural una serie de estructuras gramaticales con la finalidad de demostrar que *se sabe*. La única manera de paliar este problema es que el examinador elabore preguntas que motiven el uso de estos diferentes tiempos verbales, aunque parece que, en los niveles más altos, los candidatos tienen que llevar el control de la conversación, caso en el que el entrevistador/evaluador debe prestar más atención al contenido y direccionalidad de la discusión, puesto que esto es bastante más imprevisible. Estas dificultades parecen consolidar la presencia de un evaluador independiente de alguna manera, ya sea mediante la grabación del examen en cintas (lo cual requiere que cada examen “se lleve a cabo” dos veces

como mínimo) o la presencia física de otro examinador en el aula, cuya atención no se ve mermada al no tener que conducir la entrevista (lo cual conduce a una estructura de poder aún más desequilibrada). Estas cuestiones vienen también señaladas en el diseño de nuestro propio procedimiento de examen.

La idea de un informe de evaluación en el examen tipo “Trinity” es particularmente interesante e innovadora, y podría servir de apoyo para proveer a los candidatos con las directrices necesarias para identificar sus propios puntos fuertes y débiles, así como permitirles determinar con precisión algunas áreas en las que necesitan mejorar sus destrezas orales. Hemos incorporado este aspecto en nuestro modelo evaluador mediante la combinación del procedimiento de autoevaluación y la familiarización con los baremos de evaluación, lo cual proporciona al estudiante una idea razonable de lo que hace bien y de lo que necesita mejorar. El hecho de que los estudiantes dispongan de los mismos criterios que los que emplean los examinadores significa que los profesores dejan de implicarse en la intensa realización de informes individuales de más de 100 estudiantes, mientras que, al mismo tiempo, los propios alumnos obtienen información sobre la manera en que han sido evaluados, lo cual va más allá de las numeraciones empleadas en el baremo *Pass/Fail* (aprobado/suspenso).

4.3.3 Cambridge ESOL Examinations

La prueba oral para el *Cambridge ESOL* forma parte de un examen global de las cuatro destrezas que incluye una prueba de comprensión oral. Las franjas y puntuaciones empleadas para el examen oral en toda la serie de pruebas siguen el mismo diseño básico, si bien el texto se adapta en función del nivel evaluado. Existen varios niveles de realización, los cuales corresponden a los niveles del

Marco Común de Referencia Europeo: Aprendizaje, Enseñanza, Evaluación. El Consejo de Europa los define como “Usuario Básico” (basic user): A1 y A2; “Usuario Independiente” (independent user): B1 y B2, y “Usuario Competente” (proficient user): C1 y C2 (para una definición más precisa, consúltese http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/levels.html). El nivel que se describe en nuestra investigación, y que corresponde al estadio de aprendizaje de nuestro segundo año académico, se vincula al C1 y, en consecuencia, el baremo de Cambridge que hemos utilizado como punto de partida es el del *Certificate in Advanced English*.

La prueba tiene lugar en tiempo real, en el que los candidatos se examinan por parejas, con dos examinadores presentes, y cuya duración es de 15 minutos. Se pretende evaluar “la interacción conversacional en inglés en una serie de contextos”, y, por medio de tareas, los alumnos se centran en “el intercambio de información personal y basada en hechos, expresando y comprendiendo las actitudes y opiniones” (*CAE Handbook*)¹. Se divide en cuatro partes: una sección de entrevista, un turno individual en un largo periodo de tiempo, una tarea colaborativa y una discusión a tres bandas (dos candidatos y el interlocutor).

El examen oral se evalúa en cuatro áreas: gramática y vocabulario, manejo del discurso, pronunciación y comunicación interactiva. Mientras los candidatos tienen acceso a las características de dichas categorías que se encuentran en cualquiera de los libros del curso que los prepara para el examen, los baremos y descriptores de evaluación no están disponibles al dominio público y por lo tanto no se han reproducido en este trabajo.

¹ <http://www.cambridgeesol.org/exams/cae.htm>

4.3.4 Diseño del baremo de evaluación para *Lengua BII*

Debido a la implicación personal y al aprendizaje mediante la serie de exámenes orales y la similitud de procedimiento de todos los exámenes, en los que el habla constituye un componente de una prueba más amplia en todas las destrezas, se decidió que el baremo de evaluación oral para *Lengua BII* se basara ampliamente en el *Cambridge ESOL*. El nombre de las categorías y el mismo baremo de evaluación fueron modificados según nuestras propias circunstancias y necesidades, y la definición de las características que tenían que ser evaluadas en cada categoría es genuina.

La descripción de características, tal y como estas aparecen en las instrucciones de los examinadores de *Lengua BII*, se reproduce a continuación y del siguiente modo:

Gramática y vocabulario

En esta categoría, el propósito consiste en evaluar la precisión gramatical de las proposiciones. Las imprecisiones ocasionales y de poca consideración no son importantes, especialmente en la fase de “adaptación y acomodamiento al examen”, si bien las imprecisiones frecuentes y repetidas deben tenerse en cuenta, especialmente si impiden la comprensión del discurso.

La variedad y propiedad del vocabulario empleado por el candidato también se evalúa en este caso. A los estudiantes se les pide poseer un buen dominio de vocabulario para hablar sobre ellos mismos, y un nivel adecuado para tratar otros temas de conversación. Se reconoce el uso de la paráfrasis para expresar los conceptos más complejos, pero no se acepta este recurso cuando una palabra o frase forma parte del vocabulario exigido en este nivel.

Pronunciación

En este caso se tienen en cuenta tanto la pronunciación de palabras individuales como los patrones generales de ritmo y entonación. A los candidatos no se les penaliza por tener un acento influido por su lengua materna, a no ser que impida el entendimiento del discurso. Sin embargo, se reconoce como positivo el hecho de que se esfuercen en aproximarse a una pronunciación de nativo e intenten pronunciar rasgos como las formas débiles y fuertes (acentos).

Al considerar los patrones de ritmo y entonación de los candidatos, los evaluadores deben desempeñar el papel de hablantes del inglés tolerantes, al objeto de decidir cuánto esfuerzo realiza el candidato al pronunciar bien y cuánto impiden dichos patrones la comprensión del discurso.

Estructura del discurso

Esta categoría guarda relación con la coherencia interna, es decir, la habilidad del estudiante en organizar coherentemente el discurso mediante el uso apropiado de recursos cohesivos y los tiempos verbales, de manera que el estudiante sea capaz de exponer un argumento en este nivel (no necesariamente extenso) o una afirmación, y apoyarla mediante recursos relevantes sin dejar los enunciados a medio acabar o pausarlos en un tiempo indefinido para buscar estructuras lingüísticas u otras ideas.

Interacción

El propósito evaluador en este apartado consiste en juzgar la habilidad de los candidatos en su interacción con los otros compañeros durante la conversación, desde el punto de vista del nivel sociolingüístico, haciendo especial hincapié, en primer lugar, en la sensibilidad que muestran en el cambio de los turnos de intervención, en segundo lugar, en el uso de las estrategias de comportamiento,

tanto en las desavenencias como en el aliento a que participen los otros compañeros, y, en tercer lugar, en el nivel de la coherencia externa cuando responden con lógica a las indicaciones y preguntas, y son consecuentes con la direccionalidad de la conversación.

El **baremo de evaluación** se presenta en primer lugar a través de la consideración de las categorías como componentes del constructo del habla, tal y como se señaló anteriormente, seguido del diseño de los descriptores que resumirían el nivel de la habilidad que tendría que ser medida. Dado que varios evaluadores y entrevistadores emplearían el baremo, su visualización debía ser clara, concisa y fácil de manejar, con un uso terminológico que hiciese posible y sin dificultad la diferenciación de las características que distinguen cada puntuación. Por esta razón, se representaron mediante una tabla, donde las categorías y las puntuaciones para evaluar al candidato podían ser relacionadas visualmente, y, además, en formato de una sola página, con el fin de no estar pasando páginas o remitirse a varias. Este aspecto es relevante en las evaluaciones realizadas en tiempo real, ya que observar a los examinadores hojeando un sinfín de papeles causa ansiedad en los sujetos evaluados.

Como hemos visto con anterioridad, definir con claridad y diferenciar diez puntuaciones resulta una tarea casi imposible, con los descriptores sustituyendo términos como “la mayoría” por “casi todos”, o recurriendo a los intensificadores como “muy” para cambiar “frecuente” a “muy frecuente” con el objetivo de justificar las distintas calificaciones. También hemos observado que los descriptores del baremo incluso se repiten regularmente, como ocurre con los criterios de evaluación de los niveles 11 y 12 del Trinity, donde muchos de los descriptores del nivel 11 se repiten de nuevo en el nivel 12. Por esta razón, y al

objeto de simplificar el proceso de elaboración del baremo, lo hemos reducido a cinco puntos, como ocurre en el modelo ESOL de Cambridge, en el que la banda media representa la obtención necesaria de puntuación satisfactoria para pasar de nivel. Con este proceder, el baremo resulta más fácil de aplicar, y de este modo se muestra claramente la diferenciación entre las puntuaciones.

Se decidió definir únicamente tres puntuaciones de las cinco posibles: la puntuación más baja (1), la obtención requerida para aprobar (3), y la puntuación más alta en el nivel examinado para Lengua BII (5). Con ello se conseguía simplificar el baremo para su empleo en tiempo real, puesto que los evaluadores no tendrían que enfrentarse a tantos descriptores y la puntuación “adecuada” sería en este sentido un claro punto de partida para la evaluación, a partir de la cual los examinadores podrían oscilar hacia arriba o hacia abajo según la actuación del candidato. Asimismo resultaría beneficioso para los candidatos, al suponer al principio que poseerían un nivel adecuado, más que empezar desde el nivel inferior del baremo y ver lo que queda por conseguir, lo cual no suele ser conveniente para lograr altas puntuaciones. Muchos exámenes orales que no están estandarizados únicamente emplean estrategias de puntuación negativas, que se centran en el número de errores cometidos durante el examen, sin tener en cuenta las características positivas de la actuación de los candidatos. Sostenemos que esta actitud constituye una manera poco realista de juzgar la producción oral, ya que todos los hablantes, incluidos los expertos, cometen errores en su primera lengua, y tienden a corregirse y rectificarse, por lo que resulta absurdo emitir un juicio sobre la oralidad de una lengua extranjera basándose en los errores aislados cometidos durante el examen, especialmente si la interacción y la comunicación eficaz se consiguen. Dado que la perfección en la oralidad es prácticamente imposible,

creemos plausible la obtención de la puntuación más alta en un nivel específico, si este nivel y la puntuación han sido definidos dentro de ciertos cánones factibles y se ha determinado un punto máximo para el nivel.

4.4 Cuestionarios

Tras los exámenes, se entregó un cuestionario a los entrevistadores y a los estudiantes, con el fin de recopilar información de sus opiniones sobre el procedimiento, formato y puntuación de cada examen. Los entrevistadores debían cumplimentar sus cuestionarios (cuestionarios 2 y 4) después de la realización de los exámenes en cada sesión, mientras los estudiantes debían responderlo una vez obtuviesen sus puntuaciones unos días después de la realización de los exámenes, con la finalidad de expresar sus opiniones sobre si habían entendido la puntuación adjudicada (cuestionarios 1 y 3).

El primer cuestionario de los entrevistadores (2) se centraba en dos áreas principales. La primera, denominada “Test Management”, contenía preguntas sobre la doble función de entrevistar y evaluar, el tamaño de la muestra de producción oral del candidato y la interacción producida durante la prueba. La otra área de interés en el cuestionario se vinculaba al “Global marking vs. Analytic rating”, que incluía preguntas específicas sobre la comprensión del proceso de evaluación, el foco de evaluación, la claridad y las repercusiones de la puntuación en los entrevistadores y estudiantes en cuanto a sus respectivas opiniones convergentes/divergentes.

El segundo cuestionario dirigido al entrevistador (4) y cumplimentado después de la ‘Prueba oral de grupo’ incluía igualmente todas estas áreas, pero se ampliaba hacia la consideración del papel de la autoevaluación del alumnado, con

especial atención a considerarla como componente de la puntuación final y a conocer si esas autoevaluaciones eran útiles y precisas.

El cuestionario 1, dirigido al estudiante, consideraba en primer lugar la experiencia de la entrevista individual, con especial atención a los aspectos socioafectivos, la actuación del estudiante y el procedimiento del examen. Asimismo indagaba en las opiniones sobre las características del examen y de las actividades en cuanto a la familiarización con la tarea, nivel de dificultad y tema. Finalmente consideraba las ventajas y desventajas de la puntuación global y de la analítica, con especial atención a la claridad y comprensión de la puntuación, al igual que a su posible utilidad respecto a la forma de mejorar la destreza oral.

El segundo cuestionario del estudiante (4) contenía preguntas similares, pero esta vez se pretendía averiguar ciertas opiniones sobre la experiencia del examen oral en grupo, incorporando la autoevaluación y considerando aspectos como la precisión, la utilidad y las ventajas de obtener el entrenamiento necesario en las técnicas de autoevaluación.

4.5 Recogida de datos

Para la recogida de datos y en particular para las puntuaciones de los candidatos, se utilizaron planillas de puntuación específicamente diseñadas para cada formato de examen. Posteriormente, estas hojas se utilizaban para transferir todos los datos al programa informático apropiado para el análisis, el SPSS.

Al objeto de acelerar el procedimiento del examen, a los candidatos se les entregó una planilla de puntuación antes de entrar en el aula de la entrevista, en la que cumplimentaban su nombre y, en el caso del examen oral en grupo, también el nombre de los otros candidatos que formaban parte del examen. Estas planillas se

entregaban al entrevistador, quien a su vez se las pasaba al evaluador, el cual sería el responsable exclusivo para completarlas. En ningún momento, el examinador, en su papel de entrevistador/interlocutor, debía anotar algo en las planillas de puntuación, y los examinadores otorgaban las puntuaciones de manera independiente, sin discutir o modificar nada según la evaluación del otro examinador.

Los estudiantes tuvieron que completar tres hojas de autoevaluación a lo largo de los diferentes estadios de la investigación descrita anteriormente. Obtuvieron ayuda al interpretar los criterios si la solicitaban, pero de ninguna manera se vieron influenciados en este sentido por otros estudiantes o profesores. A lo largo del estudio, las hojas poseían el mismo formato y se cumplimentaban aplicando los mismos criterios. Al final de cada proceso global de recogida de datos, las hojas fueron numeradas de la misma manera que anteriormente, y los datos se interpretaron asimismo mediante el programa informático SPSS.

5. Resumen de los resultados

Al analizar la recopilación de datos, nos remitimos a los dos formatos de examen empleados, al procedimiento de puntuación, junto con su eficacia y utilidad, y, finalmente, a las posibles implicaciones pedagógicas de la autoevaluación y al papel que podría desempeñar este procedimiento en la prueba.

5.1 Formato de examen – perspectiva del estudiante

En relación con el formato del examen, teníamos interés por conocer el **grado de ansiedad** del alumnado y saber si esta disminuyó en el examen oral en grupo, en contraposición a la situación de desequilibrio de poder que se origina en

la entrevista individual. Nuestra hipótesis inicial consistía en comprobar si la 'Prueba oral de grupo' reducía los niveles de ansiedad de los candidatos debido a aspectos de la prueba como la familiarización con el procedimiento del examen y su similitud con las tareas realizadas en clase, así como por el apoyo obtenido entre los participantes, y si, en consecuencia, esto contribuía a una mayor autoconfianza y seguridad con el examen y con la expresión de sus opiniones. Respecto a la familiarización con el procedimiento del examen, el 88.2% de los estudiantes asintió que las tareas de la 'Prueba oral de grupo' eran similares a las realizadas en clase, mientras que solo un 37.5% encontró una similitud entre la 'Entrevista individual de competencia oral' y el procedimiento realizado en el aula. Del mismo modo, el 92.2% de esos sujetos preguntados respondieron afirmativamente a la pregunta "supe exactamente lo que tenía que hacer" en el examen en grupo. Se aprecia una reducción de la ansiedad, ya que el 70.5% de los candidatos declararon estar nerviosos a lo largo de la realización del examen, comparado con el 82.3% en la entrevista individual. No obstante, no podemos saber si esa amplia diferencia es estadísticamente significativa al tratarse de dos grupos de alumnos que no eran totalmente idénticos, pues mientras que al primer examen sólo se presentaron 51 alumnos, al segundo lo hicieron 78.

Tanto en la entrevista como en el examen en grupo, el 78% de los estudiantes se sintió cómodo con el procedimiento del examen. Estas declaraciones resultaban sorprendentes, pues pensábamos que se produciría un mayor descontento con la entrevista, debido a la situación desequilibrada de poder (incluso mucho más en los exámenes realizados para nuestra investigación, debido a la presencia de un entrevistador y de un evaluador en el primer examen, con lo que se creaba de este modo una proporción de 2-1 examinador/candidato). Nuestra

hipótesis se basa en que la estructura social o de poder en la universidad está tan arraigada -incluso en todo el sistema educativo- que los estudiantes se conforman con aceptar cualquier tipo de examen razonable, guarde o no guarde relación con el programa de estudios impartido

Por otro lado, los estudiantes podrían concebir la entrevista como un medio válido para evaluar la lengua hablada simplemente porque constituye una prueba tradicional, tal y como la conciben también los profesores y examinadores: siempre se ha hecho así, por lo que se acepta sin más. Otra razón podría ser la concepción generalizada de que es un medio útil de ganar experiencia en sus futuras vidas profesionales, en las que la entrevista individual, con sus respectivas estructuras desequilibradas de poder, forma parte del trabajo. También es probable que los candidatos con una mayor confianza en sí mismos realmente disfrutaran de la oportunidad raramente frecuente de hablar en una segunda lengua de manera individual y de que no se sintieran intimidados por la situación. Además, los examinadores mostraron gran habilidad en propiciar a los candidatos un entorno cómodo y distendido, lo cual no fue posible en el examen en grupo, ya que en este último caso se delegó la interacción a los propios estudiantes.

Nuestro segundo interrogante tenía que ver con la **relación entre el procedimiento del examen y la actuación**. Al contrastar las puntuaciones otorgadas en los exámenes a los cuatro aspectos diferentes que identificamos como parte del constructo del habla que se puede evaluar objetivamente (“gramática y vocabulario”, “pronunciación”, “estructura del discurso” e “interacción”), nuestros resultados señalaron que, en el primer examen, la ‘Entrevista individual de competencia oral’, los estudiantes se adjudicaron la puntuación más baja en cada categoría, mientras que en la ‘Prueba oral de grupo’, los estudiantes se otorgaron la

puntuación más alta en todas las categorías. Ante esta evidencia, podríamos decir que la familiarización tanto con la tarea como con el formato de examen conduce a una percepción de mejora de la actuación. A pesar de las diferencias de puntuación que los alumnos se otorgaron en uno y otro examen, los porcentajes de estudiantes que consideraron que habían actuado bien casi se igualaron: 67.9% para el examen en grupo y 70% para la entrevista individual, lo que demuestra que sus impresiones sobre sus actuaciones no guardan un vínculo con la nota que se otorgaron. En definitiva, el formato del examen en grupo indujo a creer a más estudiantes que habían actuado hasta el límite óptimo de sus habilidades.

En tercer lugar quisimos averiguar si los estudiantes estaban de acuerdo con que el formato del examen les permitiría demostrar sus **habilidades en el habla**. En el caso del formato de examen en grupo, un 20% más de candidatos respondieron afirmativamente a esta pregunta que en la entrevista oral (37.7% comparado al 17.7%). Este hecho podría estar relacionado con la situación distendida en que se encontraron los candidatos para expresar sus opiniones sobre los temas que tenían que discutir (en el examen en grupo, el 61% declaró haber dicho bastante sobre el tema, comparado con el escaso 43.1% en la entrevista) y sobre la dificultad detectada de las preguntas (en el examen en grupo, el 63.6% declaró haber sido capaz de responder a las preguntas sin dificultad, comparado con el 52.9% en la entrevista).

5.2 Formato de examen – perspectiva del entrevistador/examinador

La primera cuestión era comprobar si los examinadores eran capaces de **gestionar los materiales y la interacción del examen simultáneamente con la concesión de puntuaciones objetivas** en los exámenes orales. Los datos

recopilados en el cuestionario 2 (la entrevista individual, puntuada según el sistema tradicional e intuitivo de 0-10) muestran que todos los entrevistadores estaban convencidos de su control de la situación. Lo mismo sucedió en la gestión de la entrevista oral en grupo (cuestionario 4), en la que se otorgó a los tres candidatos una puntuación basada en el baremo de evaluación descriptivo y analítico de 5 puntos, y donde los tres entrevistadores estaban de acuerdo y otro totalmente de acuerdo en su capacidad para gestionar la interacción y evaluación.

En el caso de la entrevista oral, tres entrevistadores creyeron haber podido gestionar la interacción de la entrevista y otorgar una puntuación adecuada del baremo analítico al final del examen, mientras que uno de los entrevistadores no estaba seguro de haber sido capaz de llevar a cabo ambos aspectos, pues, según sus declaraciones, estaba más atento en dirigir la entrevista que en centrarse en los detalles de la competencia oral de los estudiantes. Sin embargo, el mismo entrevistador creyó haber dirigido la entrevista y haber otorgado la puntuación de 0-10 de una manera competente, aunque más tarde dudó haber concedido una puntuación justa en el mismo cuestionario. Estas contradicciones nos conducen precisamente a ciertos interrogantes sobre si lo que pensamos que estamos haciendo es *realmente* lo que estamos haciendo al examinar y evaluar en tiempo real.

En las entrevistas individuales, solo dos de los cuatro entrevistadores se sintieron satisfechos con el procedimiento del examen y todos coincidieron en que estaban más pendientes de dirigir la interacción que en los criterios de evaluación; sin embargo, tres pensaban que habían dado una puntuación precisa y justa, según el baremo tanto intuitivo como analítico. Estas respuestas parecen contradictorias, puesto que, si nuestra atención está centrada en una cosa (en este caso en conducir

apropiadamente la tarea inmediata *ad hoc*, es decir, la interacción), no podemos centrarnos por lógica en algo más al mismo tiempo. En el caso de la entrevista (donde el interlocutor necesita obligatoriamente no solo atender al contenido de lo que está diciendo el candidato, sino también dirigir la interacción), no es posible anotar y, al mismo tiempo, sopesar con eficacia las distintas características del constructo del habla manifestadas en la muestra de actuación.

Para nuestras preguntas de investigación referidas a las dificultades que presenta la **gestión del examen con un grupo de examinados en comparación con las del examen individual**, solo hay respuestas parciales, ya que se detectó un defecto en el diseño del cuestionario, el cual no aludía específicamente a la comparación entre ambos exámenes. Tres de los cuatro entrevistadores no estuvieron de acuerdo o discreparon rotundamente con el enunciado “fue difícil conducir la prueba con la participación de tres estudiantes”, mientras que solo uno sí estaba de acuerdo. Sin embargo, un examinador comentó al final del cuestionario 4 lo siguiente: “Fue mucho más fácil controlar una prueba oral con tres alumnos que una entrevista oral individual; había más interacción, menos tensión para los estudiantes y para el entrevistador (¡aunque mucho más para el evaluador!) y se asemejaba más a una situación de comunicación real”. Estas posturas indican en general que los entrevistadores son capaces de dirigir un examen en grupo sin dificultad, si bien no podemos hacer una constatación global sobre sus preferencias por un formato u otro, ya que un fallo del estudio fue la omisión de una pregunta que se centrara precisamente en este punto.

Sin embargo, dos de los entrevistadores quisieron indicar que estuvieron incómodos con el doble papel de entrevistador y evaluador, mientras que los dos restantes declararon no haber encontrado dificultades en realizar ambas tareas

simultáneamente. En el examen en grupo, los cuatro entrevistadores se sintieron cómodos en sus papeles de interlocutor y evaluador global, pero no podemos establecer conclusiones definitivas de esta diferencia sin llevar a cabo antes más experimentos con más examinadores. No obstante, sí podríamos concluir que estos resultados indican que el control de la interacción y la adjudicación de una puntuación global al final del examen mediante la aplicación de un baremo descriptivo resulta menos complejo que juzgar la habilidad de un candidato mediante un baremo analítico y controlar el examen a la misma vez.

Sólo uno de los entrevistadores encontró difícil evaluar la habilidad interactiva del candidato en la situación individual, quizás por ser más consciente de que los candidatos estaban actuando únicamente de una manera receptiva y no tenían la oportunidad de iniciar o cambiar el tema o formular una pregunta, lo cual significa probablemente que no estaban lo suficientemente seguros para discrepar. Podemos constatar que esta creencia de que es posible juzgar la habilidad interactiva en una situación de entrevista por la mera soltura al responder en un diálogo dirigido, está estrechamente relacionado con la falta de una definición apropiada de lo que el constructo de habla realmente significa. Un examen en que el candidato responde voluntariamente y con relativa seguridad puede ser percibido como “buena” actuación interactiva sin la consideración de todas las características implicadas en la comunicación auténtica; la entrevista constituye relativamente una situación comunicativa poco común y no requiere una amplio número de estrategias comunicativas. Otro entrevistador, si bien estaba de acuerdo en que fue fácil evaluar la interacción del candidato, comentó que “la interacción pudo ser evaluada, pero no era natural, pues solo el entrevistador formulaba preguntas, y la interacción se define como la habilidad para mantener la conversación”. En esta

constatación podemos entrever el intento de definir justo lo que el habla conlleva de manera implícita, lo cual es un paso necesario si pretendemos evaluar la expresión oral de una manera fiable y certera.

Nuestra pregunta final en este epígrafe pretendía averiguar hasta qué punto el formato de examen ha afectado **el tamaño de la muestra de habla** producida por los candidatos. Los cuatro evaluadores consideraron que dicha muestra fue suficiente para evaluar las habilidades orales en ambos exámenes. Dos entrevistadores marcaron la opción “totalmente de acuerdo” para esta pregunta (ítem 6) en el examen en grupo, indicando que los candidatos hablaron probablemente más en dicho formato de examen. Esto contrasta considerablemente con las impresiones de los estudiantes, con un 45% de candidatos que consideraron que hablaron bastante durante la entrevista oral, y un 48% del examen en grupo con la misma opinión. Aunque esta diferencia entre las percepciones de los estudiantes sobre cuánto hablaron en ambos exámenes no es significativa, parece indicar que los estudiantes no se hallaban en desventaja respecto a la cantidad de tiempo disponible para hablar al realizar un examen oral con sus compañeros, puesto que el formato individual no significa necesariamente que se hable más.

5.3 Evaluación – perspectiva del entrevistador/evaluador

Nuestro primer planteamiento en el presente epígrafe era averiguar si los examinadores se sentían más seguros **al aplicar un baremo de evaluación descriptivo** para otorgar puntuaciones que al aplicar un baremo de evaluación tradicional de 0-10. Nuestra hipótesis era que, al reducir la serie numérica de las puntuaciones a 0-5 y proveerla de definiciones para cada uno de los puntos, los examinadores serían capaces de identificar las características de la producción oral

que les permitiría conceder puntuaciones de una manera más objetiva y, por tanto, obtener un mayor grado de confianza en el procedimiento de examen. Sin embargo, también era probable que los entrevistadores implicados en la evaluación simultánea continuasen pensando que era más fácil otorgar una puntuación según el baremo 0-10 (cuya aplicación está muy arraigada en las mentes de los evaluadores, hasta el punto de hacerles, consecuentemente, sentirse seguros al aplicarlo) que usar el nuevo baremo analítico. Los resultados mostraron un equilibrio en cuanto a la preferencia de ambos procedimientos, con dos entrevistadores/evaluadores que preferían el baremo tradicional, mientras los otros dos declaraban encontrar más fácil el baremo descriptivo de 5 puntos.

La siguiente cuestión apuntaba al modo en que los examinadores interpretaron el **significado de los dos tipos de baremos de evaluación**. Con vistas al procedimiento de evaluación para la entrevista oral, los entrevistadores tenían varias opiniones sobre la comprensión de ambos baremos de evaluación (el baremo intuitivo global y el baremo analítico detallado). Tan solo uno de los entrevistadores/evaluadores dijo que no estaba plenamente seguro de lo que se estaba evaluando al aplicar el baremo tradicional de 0-10, mientras que los tres restantes estaban totalmente seguros de saber lo que estaban evaluando al otorgar la puntuación global. Dado que este baremo no proporcionaba criterios estables en los que los examinadores pudiesen apoyarse, pensamos inicialmente que la concepción del baremo 0-10 de los evaluadores se asemejaba a la evaluación normativa (las actuaciones se comparan entre sí y se adjudica la correspondiente puntuación), más que a la evaluación criterial (las puntuaciones se otorgan en función de previos criterios consensuados sobre el nivel y las expectativas, y no son dependientes de la comparación de las actuaciones de los estudiantes). En realidad fue un defecto

detectado en el primer cuestionario lo que no nos condujo a formular esta pregunta de modo explícito, y ninguno de los examinadores reflexionó sobre ello en sus comentarios generales sobre la experiencia de la entrevista. Sin embargo, uno de los examinadores observó lo siguiente en los comentarios del examen oral en grupo: “La razón por la que encontré un poco difícil evaluar sus destrezas orales fue el hecho de tener que prestar atención a tres personas a la vez y, en ocasiones, comparaba sus actuaciones en vez de otorgar una puntuación objetiva”. Dado que una de las características fundamentales de nuestra vida diaria radica en tomar decisiones basadas en la comparación, es probable que, en este sentido, nuestros procesos inconscientes nos lleven a comparar constantemente las actuaciones de los estudiantes en la evaluación de todas sus destrezas lingüísticas y que, con el fin de emitir un juicio basado en la evaluación criterial de dichas destrezas, necesitemos ser más conscientes de nuestros procedimientos y enfoques en la evaluación.

Los cuatro entrevistadores/evaluadores estaban de acuerdo o totalmente de acuerdo en haber comprendido las características del habla que se estaban evaluando al usar el baremo analítico. Sin embargo, resulta interesante señalar la manera en que interpretaron y adaptaron los baremos analíticos según su propia comprensión interiorizada.

Todos los evaluadores emplearon un baremo modificado para evaluar a los candidatos, en vez de guiarse estrictamente según el que se había proporcionado. Dos de los evaluadores comentaron la siguiente tendencia: “A veces sentí la necesidad de emplear puntuaciones decimales, como 1.5., 1.75...”; “A veces otorgaba puntuaciones que incluían medio punto, quizás porque estoy acostumbrado al baremo de 0-10”. No olvidemos que estos comentarios proceden

de haber eliminado las puntuaciones de decimales (0.5) del diseño del baremo inicial, con el fin de simplificar su aplicación (aunque estaban incluidas en un principio, pero sin descriptores). Los examinadores no solo las reincorporaron voluntariamente, sino que ampliaron la escala al incluir valores de precisión, como las puntuaciones (.25) y (.75). Ahora bien, lo que continúa siendo una incógnita es el tipo de características de la actuación que estaban evaluando para exigir dichas puntuaciones. Probablemente intentaban establecer una calificación de las actuaciones de los estudiantes de una manera normativa, tal y como se describió anteriormente, en vez de establecerlas en función de criterios reales. La puntuación tradicional del baremo 0-10 tiende con frecuencia a distribuir a los estudiantes a lo largo del baremo con la mayoría de las puntuaciones en el medio de la secuencia y con pocas puntuaciones en los extremos (especialmente en el final más alto del baremo). Consideramos que el logro de puntuaciones altas es inusual en la evaluación y los estudiantes que las consiguen son alumnos destacados y excepcionales, lo que significa en realidad que esperamos un nivel más allá del requerido en un momento dado y en un determinado estadio del proceso de aprendizaje para obtener la puntuación más alta. Por tanto, existen dos maneras de enfocar la interpretación de la evaluación, y creemos que ambas necesitan de una mayor discusión, debate y consideración en su aplicabilidad. Podemos observar que, aunque se disponga de descriptores de los niveles y características de la actuación, la tendencia principal de los examinadores es la de interpretarlos e interiorizarlos según sus propios criterios, lo cual implica, en consecuencia, su adaptación según los modelos de evaluación previamente asimilados.

Nuestro tercer planteamiento en cuanto a la puntuación alude a la manera en que un baremo descriptivo es capaz de guiar el procedimiento de evaluación al

ayudar a los evaluadores a **centrarse en una serie de características distintivas del constructo**. Las pruebas y los resultados muestran que el evaluador otorgaba una puntuación más alta que el entrevistador en todos los aspectos de la destreza del habla que fueron evaluados. Aunque este hecho no es significativo, los resultados son sorprendentes, puesto que esperábamos que la persona que tenía acceso directo y constante a los descriptores del baremo de evaluación y por tanto una visión más objetiva de las muestra de los candidatos, sería mucho más estricta. Esto podría tener su causa en que el entrevistador evaluaba el examen retrospectivamente, fijándose de alguna manera en errores (especialmente de forma) que consideraba destacados en el habla de los candidatos, lo que supuso una puntuación con “una actitud negativa”, más que el resultado de una apreciación reflexiva y equilibrada. También podría suceder que el evaluador, quien únicamente centraba su atención en la actuación de los estudiantes, pudo detectar características lingüísticas distintivas, así como los puntos fuertes y débiles de los candidatos a través de la constante referencia durante la prueba a los baremos descriptivos. De esta manera compensó o equilibró los aspectos positivos y negativos de la actuación.

Aunque nuestros resultados no indican directamente que el evaluador se estuviese centrande en un mayor número de características del habla al usar el baremo descriptivo, su actitud muestra indicios de una explicación del motivo por el que los evaluadores otorgaban puntuaciones más altas que los entrevistadores. La convicción general sobre ambos tipos de entrevista apuntaba a que era igual de fácil puntuar a un estudiante que expresaba una opinión contraria a la del entrevistador que evaluar los puntos de vista de un estudiante que coincidían con los suyos. Sin embargo, es interesante destacar que los tres entrevistadores

marcaron la casilla “en desacuerdo” en el primer cuestionario (en la entrevista oral; ítems 14 y 15) y que los cuatro marcaron “en completo desacuerdo” para el examen oral en grupo (ítem 10). Estas apreciaciones señalan de hecho que, si bien los entrevistadores creían ser objetivos en todas las ocasiones, había diferencias entre estar implicados en la interacción y estar al margen de ella como evaluador. Algunos comentarios generales de los evaluadores sobre el procedimiento del examen incluyó constataciones como “evidentemente, es más fácil evaluar al alumno ejerciendo de *evaluador* que de *entrevistador*”. El papel desempeñado por el evaluador, quien no estaba implicado en la interacción, parece indicar una mayor seguridad en el procedimiento de evaluación y una mayor convicción de que la puntuación finalmente dada es objetiva.

5.4 Evaluación – perspectiva del estudiante

La primera cuestión de este epígrafe es **el modo en que los estudiantes comprenden y responden a las puntuaciones obtenidas**. En este sentido nos interesaba averiguar si la puntuación analítica, vinculada a una serie de descriptores, resulta más significativa que la puntuación obtenida mediante el baremo tradicional de 0-10. En la entrevista oral, el 90% de los estudiantes respondieron afirmativamente a ambas preguntas “He entendido lo que significa la puntuación global que me han dado” y “he entendido lo que significa mi puntuación analítica”. Sin embargo, en la entrevista oral, el 55% de los estudiantes constataron que la puntuación global obtenida era más fácil de entender que la puntuación analítica. Teniendo en cuenta el esfuerzo que nos llevó adaptar y esclarecer los detalles del baremo de evaluación descriptivo, estas constataciones resultaron inesperadas y en cierto modo decepcionantes. Este hecho se debe en

parte a que los estudiantes, al igual que el profesorado, están acostumbrados al baremo 0-10, y, en esencia, resulta *claro* para los estudiantes: 4 significa “no he aprobado”, 4.5 significa “¿por qué el profesor no me ha aprobado?”, 5 representa “he aprobado”, y todas las puntuaciones que van más allá se conciben como ubicaciones de cada candidato dentro del grupo, basadas en juicios normativos sobre la propia habilidad oral.

El segundo planteamiento trata de averiguar si había una **relación entre las puntuaciones y el proceso de aprendizaje lingüístico** desde la perspectiva del estudiante. En este caso, encontramos una mayor complicidad en las respuestas a las preguntas “la puntuación global me ha ayudado a entender los pasos que he de seguir para mejorar mi competencia oral” (67%; probablemente “necesito mejorar”) comparado con un rotundo 94%, que consideró que la puntuación analítica les ayudó a comprender los pasos necesarios para mejorar su habla.

Los resultados del examen oral en grupo, en que solo se aplicaron baremos analíticos, fueron igualmente favorables: 92% de los estudiantes declararon haber entendido sus puntuaciones, y un 85% dijo haber comprendido lo que tenían que hacer para mejorar el habla. La razón por la que el porcentaje es más pequeño al responder afirmativamente a la misma pregunta en el examen oral en grupo, en contraposición a la entrevista oral, es incierto, ya que se emplearon los mismos baremos en ambas pruebas. Esto pudo deberse a que los estudiantes no comprendieron cómo podrían mejorar la interacción, ya que los estudiantes se evaluaron de manera significativa con puntuaciones más altas que las concedidas por el evaluador en dicha categoría, y por tanto consideraron haber interactuado mucho mejor en esta prueba que en la entrevista individual.

5.5 Autoevaluación – perspectiva del estudiante

Dado que el baremo de evaluación analítico pretendía definir algunas de las características del habla, aspecto que se consideró necesario para estimar la actuación del estudiante y establecer generalizaciones sobre su habilidad, decidimos que sería útil para los estudiantes analizar y usar dichos baremos como vehículos de autoevaluación y focalización de ciertos aspectos que debían mejorarse. Asimismo creemos que los estudiantes se centran mejor en las tareas encomendadas y reducen su ansiedad si conocen los criterios que van a ser utilizados para la evaluación. Este enfoque (por lo menos en nuestro contexto) es poco común. A los estudiantes raramente se les pide evaluar sus propias habilidades y actuaciones; y mucho menos que se consideren parte de su evaluación general. Normalmente, la evaluación se concibe como algo que viene del exterior, y, por tanto, constituye un informe objetivo de las habilidades de los estudiantes. Sin embargo, y como hemos visto con anterioridad, esto no siempre ocurre y, en cualquier caso, el aprendizaje de ser objetivos sobre nosotros mismos, sobre nuestros puntos fuertes y débiles, y usar ese aprendizaje como apoyo en áreas que requieren atención y mejora, constituye una destreza positiva y necesaria para la vida. Por este motivo se justifica la inclusión de la autoevaluación en nuestros diseños curriculares, tanto desde el punto de vista pedagógico como social.

Nuestro primer planteamiento tiene que ver con el **papel pedagógico de la autoevaluación**, su utilidad en el aprendizaje y en el progreso. Las respuestas demostraron una postura muy positiva hacia su utilidad. La amplia mayoría de los estudiantes (91%) pensaba que la autoevaluación desempeñaba en general un papel muy útil en el aprendizaje, y un aplastante 97.5% consideraba que debían ejercitar

más la autoevaluación en sus destrezas lingüísticas, con el fin de mejorar el aprendizaje.

Respecto a si la **autoevaluación debería formar parte de la puntuación final** de la asignatura Lengua BII, un 83% de los estudiantes opinaron que dichas puntuaciones deberían ser tomadas en cuenta en su calificación global, y un 92% señaló que la autoevaluación se debería incorporar en el programa de la asignatura y de manera continuada.

Finalmente, y con relación a si los **estudiantes estaban acertados en sus apreciaciones sobre sí mismos**, encontramos que el 82% pensaba que la puntuación otorgada a ellos mismos para el habla y fuera de la situación de la prueba era un reflejo directo de sus habilidades, en contraste con el 74.5%, que aseguró que su autoevaluación de la prueba oral no reflejaba realmente sus habilidades para hablar inglés. Este aspecto resulta interesante, ya que indica que algunos estudiantes piensan que sus actuaciones en la prueba no reflejan sus habilidades subyacentes, quizás debido al efecto de la ansiedad durante la actuación o a la tendencia de actuar con menor eficacia por las presiones causadas por el límite de tiempo.

Estos resultados aportan cierta evidencia acerca del fuerte deseo de los estudiantes de implicarse activamente en los procesos que evalúan sus progresos en el aprendizaje y sus resultados finales en la asignatura. Asimismo, los estudiantes señalaron que la autoevaluación y motivación están estrechamente vinculadas y que es posible un incremento de dicha motivación estudiantil si se introducen métodos de autoevaluación en los programas de estudios.

5.6 Autoevaluación – perspectiva del profesor/examinador

Desde el punto de vista de los profesores/examinadores, la opinión de incorporar la autoevaluación en los programas de las asignaturas parece entrar en completa contradicción con la de los estudiantes. Tan solo uno de los evaluadores (recordemos que también eran profesores) consideró que la **autoevaluación debería formar parte de la evaluación general del estudiante**; otro estaba en desacuerdo, y un tercero no expresó ninguna opinión. El cuarto sí que expresó su opinión, basada en ciertas condiciones que se señalan a continuación:

“La autoevaluación precisa de un entrenamiento de años de práctica por parte del alumno para que pueda tener un valor real en lo que se refiere a la medición de su progreso. El alumno encuentra dificultad para discernir y no tener en cuenta otros factores personales como el interés, el esfuerzo, el trabajo desplegado, la afectividad, etc. Por ello, en las preguntas 16 y 18 no pongo respuesta. Si el alumno estuviera convenientemente entrenado, estaría de acuerdo en los dos casos.”

En realidad, es difícil considerar esta constatación como una respuesta positiva. La indicación de los “años de práctica” no parece una proposición realista para introducir una innovación en nuestro programa de estudios. Asimismo, las dudas planteadas sobre la capacidad de los estudiantes respecto a la objetividad constituyen un rasgo destacado de los comentarios. Ello corrobora hasta cierto punto nuestra idea inicial de que nosotros, como profesores u “observadores desde el exterior”, creemos que somos capaces de establecer la objetividad, fiabilidad y veracidad en nuestras estimaciones sobre la habilidad lingüística, a pesar de nuestra falta de referencia a una definición del constructo, mientras que nuestros

estudiantes, comprometidos con el proceso de aprendizaje, se sienten incapaces de evaluar su progreso de una manera objetiva. Podríamos argumentar sin embargo que todos nosotros, como humanos que somos, estamos expuestos a la influencia de las “consideraciones personales” en los juicios que emitimos, y que, a menudo, la experiencia promueve, más que reduce, esta actitud.

Respecto a la cuestión sobre si **los estudiantes son capaces de ser objetivos en sus autoevaluaciones**, observamos que el mismo evaluador estaba en desacuerdo con la capacidad de los estudiantes en reflexionar con ecuanimidad sobre su habilidad oral general o sobre su actuación en la prueba. Esta opinión contrastaba con la de los tres evaluadores restantes, quienes sorprendentemente estaban todos de acuerdo con dicha posibilidad. Teniendo en cuenta esta constatación, resulta difícil comprender por qué estos profesores/evaluadores no estaban de acuerdo en incorporar las puntuaciones obtenidas de las autoevaluaciones en la calificación final de la asignatura.

Por último, nos planteamos el **papel pedagógico de la autoevaluación** desde el punto de vista del profesorado, que ahora es coincidente con la opinión de los estudiantes: tres de los cuatro evaluadores estaban totalmente de acuerdo en que la autoevaluación puede ser una herramienta útil de ayuda al estudiantado para mejorar su habilidad oral en inglés. Los cuatro también estaban de acuerdo en que la autoevaluación desempeña generalmente un papel esencial en el aprendizaje. Una vez más, estos resultados son contradictorios y confusos: si los profesores/examinadores no creen que los aprendices puedan ser objetivos y precisos al evaluar sus actuaciones o competencias, es difícil comprender por qué piensan que el progreso a través de la autoevaluación es posible.

A pesar de estos puntos de vistas conflictivos e inciertos, parece haber un argumento a favor tanto de introducir la autoevaluación en nuestros programas de estudios y diseños curriculares en un futuro, como de investigar con profundidad sus repercusiones y consecuencias.

5.7 Evidencia empírica con referencia a la autoevaluación

En un intento por averiguar si existe alguna **evidencia empírica para apoyar la tesis de introducir la autoevaluación** en nuestro programa de estudios para la asignatura *Lengua BII*, hemos comparado únicamente las puntuaciones otorgadas por los evaluadores con aquellas de los estudiantes en las diferentes categorías que completan el baremo de evaluación analítico, teniendo en cuenta que en el examen en grupo, el entrevistador solo otorgó una puntuación global sobre un baremo de 0-5.

En las categorías del baremo que corresponden a “gramática y vocabulario” y “estructura del discurso”, encontramos que en la entrevista individual de competencia oral (*individual oral proficiency interview*) había una diferencia significativa entre las puntuaciones del evaluador y las del estudiante, y que no se repitieron en el examen oral en grupo. En la entrevista oral, y en relación con ambas categorías, los estudiantes entendieron que su puntuación era mucho más baja que la otorgada por el evaluador. Este hecho contrastaba con el formato del examen en grupo, donde ocurrió lo contrario (las puntuaciones de los estudiantes fueron más altas que las del evaluador), dándose además la circunstancia de que las diferencias entre las puntuaciones del evaluador y del estudiante no fueron significativas desde el punto de vista estadístico.

Respecto a la “pronunciación”, hubo una diferencia estadística altamente significativa entre las puntuaciones en ambos exámenes, si bien con la puntuación más alta otorgada alternativamente por el evaluador y el estudiante en cada una de las pruebas. En la entrevista, los estudiantes se adjudicaron una puntuación mucho más baja en este aspecto del habla, mientras que en el examen en grupo, sus puntuaciones en la autoevaluación llegaron a ser mucho más altas que las del evaluador. Esto podría deberse a que los estudiantes tienden a compararse con sus compañeros en el examen en grupo, y en consecuencia consideran que su propia pronunciación tiene un parangón favorable a la de los otros miembros del grupo, mientras que en la entrevista se sienten inferiores con respecto al entrevistador en este aspecto de sus destrezas orales, debido a razones socioculturales.

En las puntuaciones sobre la “interacción” observamos que, al contrario de la entrevista, se detectó una diferencia muy significativa en el examen oral en grupo, en el que los estudiantes tenían la impresión de que estaban interactuando de una manera mucho más positiva que lo que realmente percibía el evaluador. Esta actitud resulta muy interesante, precisamente porque los juicios hechos en el caso del examen tipo entrevista podrían ser bastantes precisos, pero basados en el desequilibrio originado en la situación interaccional, es decir, donde las habilidades de interacción de los estudiantes están sujetas de manera restringida al propio formato de examen, con el consecuente reconocimiento y compensación del evaluador en las puntuaciones. En el examen en grupo, las percepciones de los estudiantes parecen inclinarse al hecho de estar interactuando de una manera mucho más natural y, en consecuencia, se adjudican una puntuación más alta.

Sin embargo, los evaluadores no solo les otorgaron una puntuación más baja a la que ellos se adjudicaron a sí mismos, sino que también otorgaron una

puntuación media más baja que la que habían dado en la entrevista. Es difícil postular una razón para ello; quizás haya una menor exigencia sobre la competencia interactiva de los estudiantes en la situación de la entrevista, en la que los candidatos solo tienen que responder en un diálogo ya iniciado y parecen interactuar con una mayor soltura o espontaneidad. En el examen en grupo, los estudiantes deben estar más atentos a las estrategias de cambio de turno, a la necesidad de incluir a los otros en la conversación y a cambiar la direccionalidad de la conversación, es decir, a iniciar nuevos temas. Las puntuaciones de los evaluadores parecen de hecho reflejar que los estudiantes no son muy diestros en estos aspectos al hablar inglés, debido a la falta de práctica o consciencia. Si este fuese el caso, entonces vemos la necesidad de incluirlo como objetivo específico de aprendizaje en nuestro programa de estudios.

Nuestros resultados muestran que en el examen en grupo hay una diferencia significativa entre las medias del entrevistador y del evaluador al compararlas con las medias de la autoevaluación del alumnado. Diferencia que puede estar motivada por diversos factores. Los estudiantes pudieron comprobar que en el primer examen (la entrevista oral) el evaluador les había dado calificaciones más altas que las que ellos se habían otorgado. Este aspecto pudo haber originado una modificación al alza al autoevaluarse de nuevo en el segundo examen. De igual modo, el formato del examen en grupo pudo haber contribuido a una valoración más positiva, pues el alumnado usaba como referente a sus compañeros en el intercambio oral, mientras que en la entrevista el único referente era el entrevistador. No obstante, también queremos señalar que se da una diferencia significativa entre las medias otorgadas por el entrevistador y el evaluador en el examen en grupo, de forma que siempre la nota del evaluador es la más alta. Lo

cual, a su vez, nos indica que la autoevaluación del alumnado se encuentra más cerca de la del evaluador que de la del entrevistador.

A pesar de estas diferencias en las puntuaciones, se observa en nuestro estudio un dato que consideramos muy importante. Nos referimos al altísimo índice de correlación que se da entre las valoraciones del evaluador y la autoevaluación del alumnado, hasta el extremo de registrarse además esta correlación positiva en todas y cada una de las categorías de la producción oral evaluadas. Ello nos facilita una sólida base a favor del argumento de incluir los criterios de autoevaluación en nuestros diseños curriculares y en instruir a nuestros alumnos al respecto.

6. Implicaciones prácticas del estudio

Si tenemos en cuenta los resultados que hemos presentado en este estudio, consideramos apropiado establecer algunas propuestas de cambio de nuestro proyecto docente de la asignatura Lengua BII. Si bien estas propuestas requieren una investigación más exhaustiva y una validación como pasos preliminares hacia la mejora de nuestro enfoque, creemos que están justificadas y que vale la pena implementarlas debido a los resultados de nuestro estudio.

Nuestra primera propuesta es utilizar el examen oral en grupo como un procedimiento de evaluación final del componente oral de la asignatura. Creemos que las puntuaciones resultantes constituyen una medida más precisa para evaluar la habilidad oral de nuestros estudiantes que las originadas en el formato de la entrevista individual, debido a nuestro intento de definir el constructo y de describir lo que queremos medir con los baremos. La implementación efectiva de dichos baremos requiere la presencia de un evaluador objetivo, que no esté implicado en la interacción, con el fin de compensar la estructura de poder que se origina en las

pruebas bajo términos socioafectivos. En este sentido, necesitamos dos candidatos como mínimo en cada prueba, para equilibrar la relación candidato – examinador. Debido a cuestiones administrativas, y dado que el número disponible de examinadores está limitado al de los profesores que imparten la asignatura y que el número de estudiantes es relativamente alto, resulta más práctico realizar las pruebas en grupos de tres que en parejas. También somos de la opinión de que la interacción en un grupo mayor de dos miembros, proporciona la posibilidad de evaluar más facetas de la habilidad interactiva, como permitir el cambio de turno, alentar a un compañero a participar, implicar a otros hablantes en la conversación mediante la formulación de preguntas, etc.

Como consecuencia, también proponemos un cambio en el sistema de puntuación que utilizamos para evaluar las destrezas orales, y que se recoge en el baremo descriptivo que hemos diseñado. Por el momento, no existe otra alternativa que la de convertir las puntuaciones sobre una secuencia de 10, pero incluso así, las puntuaciones poseen aún un significado, ya que aluden a descripciones específicas que corresponden a una puntuación numérica del baremo. Sería posible adaptar el baremo siguiendo las demandas de una mayor precisión al añadir las puntuaciones de decimales (.5), pero sin un descriptor. No obstante, es necesario continuar con la investigación y la consulta a evaluadores para confirmar que estos aspectos mejoran el baremo y permiten una mayor precisión en las puntuaciones.

Por último, nuestros alentadores resultados sobre el impacto y repercusión que podría tener la autoevaluación en el aumento de la motivación, aprendizaje y progreso en los estudiantes, nos lleva a proponer su inclusión como una importante innovación y experimento en nuestro proyecto docente de Lengua BII. Creemos igualmente que, al proporcionar a nuestros estudiantes ciertos criterios similares a

los que empleamos para evaluar sus destrezas lingüísticas y al invitarles a estimar el alcance de sus capacidades según lo requerido por el programa de la asignatura, podremos dirigir sus esfuerzos de una manera positiva y efectiva.

Estamos convencidos de que este cambio en nuestro enfoque sobre la enseñanza y aprendizaje de una lengua podría tener un mayor impacto en las actitudes de los estudiantes hacia el aprendizaje. En una sociedad donde casi todo es adquirible si se paga por ello, nuestras jóvenes generaciones han crecido en un mundo de proveedores de servicios, en el que prácticamente todas nuestras necesidades y tiempo de ocio se cubren mediante el intercambio monetario. La educación tampoco queda al margen, y es corriente observar cómo nuestros estudiantes, como otro sector cualquiera, igualan la asistencia a clase con el aprendizaje en sus valoraciones, inconscientes de su papel esencial en el proceso de aprendizaje. Si podemos implicar al estudiantado en sus propios procesos de aprendizaje mediante el refuerzo en ciertas áreas y vías que les dirijan hacia el progreso, estaremos encaminados hacia la solución del problema.

Uno de los aspectos aprovechables de describir lo que los discentes pueden hacer es centrar la atención en las cualidades y logros positivos. Somos de la opinión de que la crítica no puede ser nunca constructiva. La crítica en sí es desmotivadora por definición: solo indica lo que hemos hecho mal, los objetivos que quedan por cumplir. Con la finalidad de ser constructivos, necesitamos realizar *observaciones* que describan lo que podemos hacer y lo que hemos logrado, y que también proporcionen sugerencias para nuevos pasos hacia el progreso. Un reto futuro lo constituye el desarrollo y diseño de baremos de puntuación con descriptores útiles que representen un vehículo común para profesores/examinadores y estudiantes.

7. Contribuciones y limitaciones para nuestro estudio

Como hemos mencionado anteriormente, una de las mayores limitaciones de nuestro estudio es el reducido número de sujetos implicados en el experimento, lo cual repercute en el impacto del estudio y en la generalización de los resultados y de las conclusiones. Sobre todo en relación con el número de examinadores implicados, necesitamos llevar a cabo otros experimentos con un mayor número de evaluadores y entrevistadores, con el fin de confirmar las tendencias que hemos observado (generales y comunes a la condición de todo examinador) en algunas áreas, como la adaptación del baremo según las concepciones valorativas, la seguridad en la gestión de la entrevista y en la puntuación simultánea o la detección de las características del habla al mismo tiempo que se adjudicaban las puntuaciones.

Quizás tuvimos que haber incluido preguntas más directas en el cuestionario del entrevistador sobre el uso del baremo 0-10, con el fin de establecer como mínimo lo que los evaluadores *piensan* sobre lo que hacen cuando lo aplican a la evaluación de la actuación oral. Dado que no hemos hecho ninguna alusión específica a esta actividad en nuestro cuestionario, solo podemos hacer conjeturas sobre la manera en que los evaluadores parecen usar el baremo de una manera normativa.

También se ha apreciado una cierta limitación en la naturaleza voluntaria de participación de los sujetos en la primera prueba (entrevista individual). Esto significa que los dos grupos de estudiantes que realizaron las pruebas no eran homogéneos, por lo que fue imposible comparar directamente ciertos datos estadísticos, tales como las puntuaciones dadas por el evaluador en cada categoría

del constructo del habla en los dos exámenes. Ante este hecho, solo podemos limitarnos a comentar ciertas tendencias generales en vez de establecer afirmaciones sobre la objetividad del evaluador y la serie de características del constructo que se focalizaron durante la evaluación.

Aunque no hayamos hecho una contribución significativa en el ámbito del aprendizaje y evaluación de la lengua extranjera, sí hemos establecido algunas bases para un posible cambio en nuestro contexto evaluador, lo que ha posibilitado un avance más en el diseño de nuestro plan docente, basado en la evidencia más que en la intuición. Este avance probablemente influiría en la práctica de otras asignaturas de similares características, como mínimo en nuestra institución, con el posible efecto de ver aumentada la coordinación y continuidad de los proyectos docentes, otro de los objetivos de las modificaciones EEES.

Creemos que nuestros resultados nos proporcionan una evidencia sólida sobre el valor pedagógico de la autoevaluación, y de ahí la necesidad de su inclusión en los diseños curriculares en la didáctica de lenguas extranjeras. También hemos demostrado que los estudiantes *sí* se autoevalúan objetivamente y de manera similar a los evaluadores cuando se trabaja con los mismos baremos y descriptores. Esto es una buena razón para continuar con la investigación sobre la autoevaluación como herramienta útil para nuestros procedimientos evaluadores, y estamos seguros de incluirla como propuesta para modificar nuestro propio contexto docente.

8. Posibles áreas de investigación futura

Obviamente, cuanto más exploremos las áreas de conocimiento, mayor será el grado de consciencia sobre nuestra ignorancia. El presente estudio ha presentado

además otros puntos débiles relacionados con la comprensión del constructo del habla y la conveniencia de nuestras herramientas de medida, lo cual requiere de una mayor consideración y de una investigación futura.

Necesitamos asimismo seguir desarrollando nuevas definiciones sobre el constructo, con el fin de estar seguros de que estamos haciendo las mediciones apropiadas para la destreza del habla. Esto incide especialmente en la competencia interactiva, la cual fue la categoría con mayor discrepancia entre las puntuaciones de los estudiantes y evaluadores en el examen en grupo, y el aspecto al que se le ha prestado menor atención en el ámbito científico. Investigar sobre las razones por las que nuestros estudiantes se perciben a sí mismos interactuando de una manera mucho más competente que cómo la ven los evaluadores, arroja luz sobre la esencia de la interacción y lo que ella implica, y si esto difiere según las lenguas y culturas. Si este es el caso, entonces necesitaríamos incluirla en nuestros proyectos docentes, a fin de que el estudiantado se mantenga informado sobre lo que el profesorado considera positivo o negativo en la interacción en grupo o en parejas.

En lo que se refiere a las puntuaciones, nos gustaría proponer un estudio que arroje luz sobre los procesos cognitivos que tienen lugar cuando aplicamos el baremo 0-10, y cuyos resultados podrían contribuir a corroborar nuestra teoría sobre su naturaleza normativa, lo cual nos proporciona más fundamentos sobre la necesidad de la descripción en la evaluación. Ello nos aportaría una mayor transparencia y credibilidad en nuestras actitudes y un beneficio para la institución.

Es nuestra intención realizar un estudio sobre la inclusión de la autoevaluación, con el fin de llevar un seguimiento de sus procedimientos. Para corroborar su impacto positivo en el aprendizaje, necesitamos diseñar instrumentos de medida y recogida de datos sobre sus posibles efectos en la motivación y

aprendizaje. Ello requiere, por un lado, establecer una distinción entre ambos aspectos, y, por otro, intentar verificar mediante datos objetivos si están tan directa y positivamente relacionados como creemos.

Asimismo se podrían continuar otros estudios con los datos recopilados en la presente investigación. Al analizar las grabaciones en video y en las cintas de todas las entrevistas realizadas, sería posible comparar el tiempo medio disponible para la producción del habla de los candidatos en los formatos de la entrevista y del examen en grupo y así determinar si el examen en grupo, tal y como creemos, permite a los estudiantes un mayor tiempo para hablar, ya que el entrevistador no interviene tanto en la mayor parte de la interacción. También resultaría interesante averiguar a través de las grabaciones si el tamaño de la muestra producida por el candidato es positivo en función de la puntuación final obtenida (es decir, cuanto más hablan, mayor la puntuación obtenida). Si esto se llegara a probar, podría ser una indicación de que los evaluadores se centran más en la interacción o, como mínimo, en la “buena voluntad de participar”, puesto que otro indicio en este sentido sería que, cuanto más habla un candidato, mayor número de errores podría cometer.

Durante las sesiones de evaluación, las carpetas de materiales utilizados fueron recogidos en las hojas de puntuación de los candidatos. Al analizar las puntuaciones obtenidas en los exámenes en relación con el material, podríamos también determinar si existe una relación entre los temas de la discusión y las puntuaciones. Ello nos conduciría a diseñar materiales de examen que garanticen una igualdad de condiciones para todos los candidatos. También sería posible seleccionar en este caso qué tipo de temas les resulta más interesante a los estudiantes para expresar sus opiniones o cuáles están estrechamente vinculados a

sus ámbitos de experiencia. También se esclarecería si estos temas de mayor interés y accesibilidad coinciden con los impartidos durante el curso académico. Ello nos llevaría a establecer algunas deducciones provisionales sobre su repercusión en la práctica en el aula o, como mínimo, en la seguridad adquirida por los estudiantes al presentarles previamente los temas o el vocabulario.

Esperamos que nuestro estudio haya contribuido modestamente a aportar informaciones sobre nuestra práctica evaluadora, aunque es evidente que nuestra tarea en encontrar un modo válido y fiable de examinar y evaluar la competencia oral se encuentra todavía en el principio, es decir, tal como señalamos anteriormente, necesitada de otras investigaciones futuras. Nos esforzaremos en continuar nuestra investigación para recoger los datos científicos necesarios como base para nuestras decisiones docentes en la educación y en su desarrollo.

