UNIVERSIDAD DE JAÉN

FACULTAD DE HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN
DEPARTAMENTO DE FILOLOGÍA INGLESA

TESIS DOCTORAL

A PROTOCOL TO DESIGN A *CEFR*-LINKED PROFICIENCY RATING SCALE FOR ORAL PRODUCTION AND ITS APP IMPLEMENTATION

PRESENTADA POR:
JOAQUÍN MANUEL CRUZ TRAPERO

DIRIGIDA POR:
DR. D. ANTONIO BUENO GONZÁLEZ

JAÉN, 29 DE NOVIEMBRE DE 2016

ISBN 978-84-9159-028-6

A mis padres, que me dieron la vida.
A mi mujer e hijos, que han hecho que ésta tenga sentido.

A mis maestros, que me han enseñado a leer entre líneas.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr Antonio Bueno González, my supervisor, for his understanding, guidance and, most of all, for being an inspiration throughout my professional career. I am particularly indebted to engineer Raúl Pérez Fuentes for his technical support in the development of Rubrik©. I am most grateful to Dr Michael Linacre for his advice on the analytic scales developed in this dissertation. Special thanks as well must go to all the colleagues from the language centers of the nine Andalusian universities who provided me with the data on which I have built most of my research. In this respect, I am particularly obliged to the colleagues from the universities of Cádiz, Huelva, Jaén and Seville who also helped me during the statistical validation of the rubrics and who provided me with valuable insight into their design. Naturally, I am also thankful to my parents and spouse without whose support I would have been unable to finish this work.

TABLE OF CONTENTS

Acknowledgments iii

Introduction v

PART 1 – THEORETICAL ASPECTS

Chapter 1. **A CONSTRUCT OF LANGUAGE** 1

1.1. A modern definition of language: biolinguistics 3

1.1.1. Biology 9
1.1.2. Psychology 11
1.1.3. Neurolinguistics 17
1.1.4. Computer sciences 20
1.1.5. Linguistics 23

1.2. A theory of language proficiency 30

1.2.1. Acquisition and learning 31
1.2.2. Competence and performance 33
1.2.3. The *CEFR* (Council of Europe, 2001) and its levels 40

Chapter 2. **A CONSTRUCT OF TESTING** 45

2.1. Measurement, test and evaluation 46

2.2. An updated approach to testing 48

2.2.1. Validity 49
2.2.2. Reliability 53
2.2.3. Fairness 56
2.2.4. Practicality 59

2.3. Test methods 62

2.3.1. Direct, indirect, analytic and holistic methods 64
2.3.2. Rubrics: history and definition 69

2.4. Psychometrics 75

2.4.1. CTT and sample dependency 79
2.4.2. MTT and statistical models 84
2.4.3. One-parameter logistic models and multi-faceted Rasch 91
2.4.4. Facets (Linacre, 2014) 94

Chapter 3. **THE CONTEXT** 101

3.1. European policies 104

3.1.1. The Council of Europe 105
3.1.2. The Bologna process 107

3.2. Spanish policies 117

3.3. Andalusian policies 122

PART 2 – THE EXPERIMENT

Chapter 4. **DESIGN OF A NEW SET OF RUBRICS** 139

4.1. Revision of previous sets of Andalusian rubrics 140

4.2. Development of a protocol to design rubrics 152

4.2.1. Stage 1. Previous considerations 155
4.2.2. Stage 2. Writing the descriptors 159
4.2.3. Stage 3. Validation 1 (qualitative) 169
4.2.4. Stage 4. Validation 2 (quantitative) 170
4.2.4.1. Data fit 172
4.2.4.2. Vertical ruler 176
4.2.4.3. Rating scale category utility 182
4.2.5. Stage 5. Implementation 186
4.2.6. Stage 6. Revision 196

Chapter 5. **DESIGN OF A MOBILE APPLICATION: RUBRIK©** 197

5.1. Design concerns 199

5.1.1. Functionalities 199
5.1.2. Programming 201
5.1.3. UX analysis 204
5.1.4. UI design 205

5.2. Production 215

PART 3 – CODA

Chapter 6. **CONCLUSIONS** 225

6.1. Concluding remarks 225

6.2. Further work 228

6.2.1. Biolinguistic concerns 228
6.2.2. Additional improvements to the protocol 229
6.2.3. Extra functionalities for Rubrik© 230

6.3. Methodological implications 231

PART 4 – SECTIONS IN SPANISH

Título 235

Índice 237

Introducción 241

Resumen 253

Conclusiones 257

GLOSSARY 259

REFERENCES 289

APPENDIX 311

INDEX OF FIGURES

1.1.	Maddieson's (2009) comparison of phonemic systems	6
1.1.2.	Baddeley's central executive diagram	16
1.2.2.	Arrangement of competences in our set of rubrics	36
1.2.3.	The *CEFR* (Council of Europe, 2001) levels and sublevels	43
2.1.	Measurement, tests, and evaluation	47
2.2.	Main areas of interest in language testing	49
2.2.1.a	Sample reading passage from a real B1 proficiency exam	49
2.2.1.b.	Reading task linked to the passage in figure 2.2.1.a	50
2.2.2.	Conceptualization of the factors that affect rater cognition	56
2.2.4.	Ideal test development cycle by Green and Spoettl (2011)	60
2.3.1.	Typical arrangement of a set of rubrics	74
2.4.	Psychometrics: CTT vs. MTT	79
2.4.2.	ICC curve of a three-parameter logistic model	89
2.4.4.a.	Main interface of Facets (Linacre, 2014)	96
2.4.4.b.	Specifications file for Facets (Linacre, 2014) with instructions	97
4.2.2.a.	Different approaches to build *CEFR*-linked rubrics	161
4.2.2.b.	Scalability of the design model proposed in the protocol	163
4.2.4.2.a.	Vertical ruler for validation trial 1 (2 raters and 84 test takers)	177
4.2.4.2.b.	Vertical ruler for validation trial 2 (9 raters and 44 test takers)	178
4.2.4.3.a.	ICC curves for trial 1	185
4.2.4.3.b.	ICC curves for trial 2	186
5.1.2.	Integrated development environment	202
5.1.4.a.	Early mock-up of the application with a bug	206
5.1.4.b.	The logo/icon designed for Rubrik©	207
5.1.4.c.	Main menu screen with settings	207
5.1.4.d.	Screen for new interviews	208
5.1.4.e.	Screen to introduce new candidates in the application	209
5.1.4.f.	Screen generated by the *Manage Candidates* button	210
5.1.4.g.	Screen generated by the *Export Candidates* button	210
5.1.4.h.	Screen for the *Import Candidates* button	212
5.1.4.i.	Browsing system to collect data stored locally	212
5.1.4.j.	Screen generated by the *About* button	213
5.1.4.k.	Interview screen	214
5.1.4.l.	Interview screen with band 2 selected	215
5.2.	Rubrik© on its beta release, July 27th 2016, with typo in the icon	220

INDEX OF TABLES

2.4.1.a.	Data for discrimination index through the Pemberton formula	81
2.4.1.b.	Expected correlation between facility value and Pemberton index	82
2.4.3.	How raw scores can disguise real ability in performance	94
3.1.2.	Timeline of the Bologna process	108
3.2.	Main Spanish national laws on education from 1970 to 2016	119
3.3.a.	Tests at Andalusian universities (2013-2016)	131
3.3.b.	Andalusian timeline of regional laws and university meetings	132
4.1.a.	Stages of the first attempt to design a common set of rubrics	142
4.1.b.	Analysis of pre-existing rubrics	144
4.1.c.	Categories, linguistic features and frequency of pre-existing rubrics	147
4.2.	Protocol to design a *CEFR*-linked proficiency rating scale	154
4.2.1.	Main linguistic features and sub-features of our set of rubrics	157
4.2.4.1.a.	Example of data misfit (candidate 6)	172
4.2.4.1.b.	Infit and outfit mean squares of the 2 validation trials carried out	175
4.2.4.3.a.	Scale category statistics for validation trial 1	182
4.2.4.3.b.	Scale category statistics for validation trial 2	182
4.2.5.a.	Exact agreement observed and expected among raters	190
4.2.5.b.	Theoretical distribution of ratings used to obtain κ coefficient	192
4.2.5.c.	Real data used to obtain κ coefficient	194
5.2.	Feedback for Rubrik© from the EALTA *CEFR* SIG Meeting	217

INTRODUCTION

The British mathematician Alan Turing, born in London in 1912, spent long periods of his early childhood in India due to his father's civil service commission. Back in England he studied at King's College and specialized in mathematics. It is estimated that his contribution to deciphering encrypted messages from the main Axis powers contributed to saving millions of lives and to shortening World War II by as many as 2 to 4 years.

Four years after the war had finished, on a cold winter's evening, he met with chemist and philosopher Michael Polanyi and zoologist and neurophysiologist J. Z. Young at a philosophy seminar at Manchester University to discuss about the prospect of artificial intelligence, which had come to the attention of the general public in the West following the success of wartime scientific developments. This conversation would evolve into the now famous paper "Computing Machinery and Intelligence" published in 1950 in *Mind* (Turing, 1950). Roughly speaking, in his paper, Turing proposed the idea of a test that could help to solve the question of whether machines can think or, more precisely, whether machines would do well in one game in which they had to imitate human behavior. The film about his life was actually called *The Imitation Game* (Tyldum, 2014). The Turing test and its implications are still today among the most influential in the philosophy of artificial intelligence.

The Turing test is one modern example of something that is at the heart of this PhD: tests. Tests (more precisely oral tests) and the other central interest of the present PhD, language, merge in the following passage from the Book of Judges (12:4-6):

> Jephthah then called together the men of Gilead and fought against Ephraim. The Gileadites struck them down because the Ephraimites had said, "You Gileadites are renegades from Ephraim and Manasseh." The Gileadites captured the fords of the Jordan leading to Ephraim, and whenever a survivor of Ephraim said, "Let me cross over," the men of Gilead asked him, "Are you an Ephraimite?" If he replied, "No," they said, "All right, say 'Shibboleth.'" If he said, "Sibboleth," because he

> could not pronounce the word correctly, they seized him and killed him at the fords of the Jordan. Forty-two thousand Ephraimites were killed at that time.

To say the least, this pronunciation test is flawed and its impact dramatic. The fairness of this on-the-spot exam is debatable as well as the qualifications of the judges who decided who was to die.

This example is meant to be shocking, but as we will see later, there are contemporary examples of tests having a similar impact. On March 23 2016, for example, the United Kingdom's Upper Tribunal rejected a ruling to deport 48,000 students who had been wrongly detained and removed accused of having cheated in one of the language exams which immigrants must take to live in the United Kingdom (Menon, 2016). The scandal broke after a BBC documentary "claimed to have uncovered fraudulent activity at an East London school involving overseas students sitting the Test of English for International Communication (TOEIC)" (Ali, 2016). At the tribunal's judgment, the investigation that followed the break of the scandal was flawed and its results were unfairly applied to many candidates, mostly Indian, who had obtained their marks in due form. Such is the impact of high-stakes tests.

From the early Chinese competitive examinations developed during the Han dynasty (201 BCE to 8 CE) (Spolsky, 1995:16) to the modern industry of testing, going through medieval oral disputations, tests have taken many different forms and have been used for a great many purposes. Tests are not only used to ascertain whether machines can think or whether people have the required level in a foreign language. On a daily basis, we check with our lips the temperature of a steaming spoonful of soup or we weight oranges in the supermarket, and we do it because there is some information missing, because we want to compare something with a particular yardstick or for simple curiosity, the latter being at the genesis of science.

Language, which arose among human beings as a means of communication based on certain pre-existing biomechanical conditions, will also be at the core or the following chapters.

The view of language that we will present here is a function of biomechanical preconditions, observation, randomness and interaction. In our view, language must be understood as a human product, something which cannot occur without human beings. We see language, as Berwick and Chomsky (2011:20) do, as another organ of the body. The phenomenon of language acquisition, the principal interest of biolinguistics, is miraculously repeated generation after generation and yet we know very little about how it takes place in our early years of life. Take for example the case of newborns. While they are suckling at their mother's breast, their speech organs rest in such a position that the air which flows out of their larynx is involuntarily articulated as mid-central vowels (as in "about" or "but") due to the position of the speech organs of the baby at a particular point. At the same time, the lips of the baby open and close to suckle producing the accidental articulation of nasal and bilabial consonants (as in "mama" or "baba"). Although there is no way of confirming this, some have hypothesized that the origin of words like "mama" and "papa" (Jakobson, 1962) is in this very biomechanical movement, which might explain why the words for "father" and "mother" are frequently similar across distant and unrelated linguistic communities. When the mother first hears the "ma" or "pa" sound (or any of its variations) she is willing to interpret it as a voluntary call from the baby. Sound and meaning are thus randomly matched and human behavior (the behavior of the mother) is conditioned by previous responses or prompts (the sounds of the baby) in a cycle that has been refined throughout the history of mankind up to the present day. This is a marvelous example of biomechanical preconditions, observation, randomness and interaction all working at once to create a pristine instance of language.

The real nature and complexity of language are still far from being completely understood. This is of the utmost importance for those who teach languages and for those who must assess their learning processes. Scholars have proposed elegant and intricate definitions of the way in which languages work. Over the last 100 years medicine and computer sciences have made it possible to take pictures *in vivo* of what happens inside the human brain when words are being recalled. Yet, there is no comprehensive and unified theory able to explain

everything about languages, from their acquisition to the way in which grammar works. Then, one might wonder, is it possible to teach and assess something whose very nature is still, for the most part, unknown? In fact, it *is* possible, but not knowing the very nature of language forces us to work on the basis of assumptions rather than on sound rules or exact procedures. In many cases, language researchers work blindly because they do not have anything like a table of elements or an expression such as $g=9,8m/s^2$ to ascertain how much language their students know. It would be easier if languages could be weighed in kilos but, unfortunately, that is not the way in which they come.

In 1957, B. F. Skinner opened in his seminal work, *Verbal Behavior* (Skineer, 1957) with a reflection on why one truly structured science of human verbal behavior (the core of which he places in psychology) has been historically neglected. He argued that such responsibility rested with "certain fictional causes which psychology has been slow in disavowing" (Skinner, 1957:5). In fact, the same can be said about most traditional approaches to a science of language, not only of those departing from the grounds of psychology. The descriptors and analytic scales, referred to in the quotation below (Knoch, 2009:12) will be defined in depth later on, but we have decided to include this passage here since it describes the feeling of being clutching at straws that many of us have occasionally felt:

> I often found that the descriptors provided me with very little guidance. On what basis was I meant to, for example, decide that a student uses cohesive devices '*appropriately*' (*sic*) rather than '*adequately*' (*sic*) or that the style of a writing script '*is not appropriate to the task*' rather than displaying '*no apparent understanding of style*'? […] This lack of guidance by the rating scale often forced me to return to a more holistic form of marking where the choice of the different analytic categories was mostly informed by my first impression of a writing script […]. I often felt that this was not a legitimate way to rate and that important information might be lost in this process.

Describing or analyzing language is complex, let alone rating students on the basis of abstraction. If we want to feel legitimized to do such and other things, first we must develop a unified theory of language.

The earliest modern attempts of developing a scientific method to describe language beyond grammar arose in the first quarter of the 20th century, for some with Saussure's *Cours de Linguistique Générale* (Saussure, 1995), for others with logical positivism and Skinner's works, but the picture was then and is still now incomplete. That was definitively an enormous impulse which has gathered momentum in more recent times. "It is no exaggeration to say that more has been learned about languages in the past twenty-five years than in the earlier millennia of serious inquiry into language" (Berwick and Chomsky, 2011:29).

Skinner's behaviorist view is widely surpassed nowadays. That was just a beginning thanks to which linguistics has progressed up to the point at which it has gained the highest level of self-cognizance: self-consciousness. This evolution has helped to define the core interests of linguistics and to differentiate theoretical linguistics from applied linguistics or from language pedagogy. This evolution has given birth to cognitive psychology or to the principles and parameters theory and has made it possible to apply mathematical models through psychometrics. And, despite all this evolution, the remarks that Chomsky and others made in the 1960s questioning whether linguistics or psychology had achieved a level of theoretical understanding that might enable them to support a "technology" of language teaching (Lawler and Selinker, 1971:28; Savard and Laforge, 1981:74) are still haunting nowadays.

It is mystifying that the study of languages still lacks one unified theory of language. Every time linguists scratch the surface, they keep returning to some fundamental questions that remain unanswered, and so do all disciplines derived from linguistics. At times it is as if these questions could only be answered by creating a new theory, a new branch or a new branch of a new theory. This variegation hampers the evolution of a discipline, linguistics, whose most logical destiny is to be studied at the core of natural sciences, since language is a product of nature in general and a product of human mind in particular.

If there were something like one universally accepted theory about the functioning of human languages, every single scholar would be using it in the same way in which chemists use the periodic table. But there is not, and this turns

ours into a complex task. Yet, complexity should not be allowed as an excuse in the times of unprecedented scientific achievements that we are living. The future usually gives us clues and it is the time for linguists to follow them in order to anticipate one unified theory that explains the basis of the human ability of language. If Sturtevant's mapping of the genes of the fruit fly in 1911 can be likened to the Wright brothers' first flight, then, the complete mapping of the human genome achieved in 2003, which draws on Sturtevant's discoveries, can be compared to the Apollo program bringing humanity to the moon (NHGRI, 2016). After looking back to Saussure and Skinner's works, one might think that linguistics has only been flying for 12 seconds and that it is time now to move further. Paraphrasing Schrödinger (2013), we believe that linguistics must set off this uncertain voyage with the certainty that the inability of present-day linguistics to account for unresolved questions is no reason at all for doubting that they will be unveiled by our discipline some day.

This step forward in the field of linguistics will, at some point, boost or unify the progress of all linguistic branches, testing among them, and will help us to sort out useful approaches from those which are not. The present dissertation has been written with the belief that the necessary breakthrough is not only a scientific need but also a social one. Our world is no longer a static reality and cultural boundaries are nowadays less clear than ever among societies. Extended international mobility has introduced new challenges not only for scientists who need to communicate efficiently worldwide but also for those who seek to find a job abroad or to study at a foreign university. All this has given rise to one unprecedented demand of language training and official certifications worldwide. Such are the clues that testing is giving researchers. Test takers are not just language students any more. Test takers are now business executives who want to sign contracts with foreign companies or members of a family who want to obtain their residence permits from immigration offices. In this context, language testers are expected to be responsible for and responsive to theories derived from 2 unrelated fields such as linguistics (which describes language) and psychometrics (which measures certain human attributes) and their job is at the same time

"directed and constrained by rival practical institutional, economic, social, and even political demands" (Spolsky, 1995:4).

Not all the aforementioned problems will be solved at the end of this dissertation. Contrary to what Leonardo Grassi told the Duque of Urbino about *Hypnerotomachia Poliphili* (Colonna, 2013:66), the present dissertation does not contain all ancient books nor unveils the very mysteries of Nature. This dissertation departs with a more humble objective, but it still provides scientific, interesting and innovative insight into certain aspects of linguistics and oral language testing that we deem important. The following pages try to answer these research questions:

1. What disciplines can help to define language in the context of non-unified linguistics?
2. Is there a scientific method to assess 1 level of oral production proficiency?
3. If the scientific method in point 2 exists and can be validated, can it be the basis to assess various levels of oral production proficiency?
4. If the scientific method in point 2 exists and can be validated, is there any way of implementing it that can transcend traditional approaches and help raters and test takers?

In trying to answer these questions we have made some (we believe) relevant contributions. First, we have outlined an effective protocol to design rating analytic scales which follows a series of easy steps. Second, we have captured in real time the breakthrough that 9 public universities from 1 Spanish region, Andalusia, are experiencing on their way to mutual recognition in language proficiency exams. Finally, we have designed an innovative form to implement oral exams which we expect to be useful for the growing community of language testers around the world. All this is included in the present research paper, which is divided into 4 parts.

Part 1 contains the theoretical aspects through which we will try to answer questions 1 and 2, and a description of the context in which they occur. In this part, chapter 1 introduces the notion of construct and develops a construct of

language, while chapter 2 defines a construct of testing. The idea is to define in this first part of the dissertation the element to be tested, the ability of language, and the means that will be used to observe and measure it. Since none of these 2 occur in vacuum, the context in which they occur will be presented in chapter 3.

Part 2 contains the experiment that we designed to answer research questions 3 and 4. Chapter 4 describes the process followed to design and validate statistically a new set of rubrics to assess oral production. Chapter 5 sets forth the steps taken to design an innovative mobile application to implement the design which was previously validated in chapter 4.

Part 3 includes chapter 6 with the conclusions drawn from the previous chapters, points to some future works and discusses the methodological implications of the main aspects of the present dissertation.

Part 4 does not add any new content to the dissertation. It simply includes the translation of some excerpts which, due to official requirements, must be included here.

To finish this introduction, we would like to make some formal remarks on the conventions followed across pages. Acronyms are only expressed in full the first time that they appear. Besides what is customary, *italics* has been used for the linguistic features of rubrics, our analytic scales, for the name of the tables of the *CEFR* (Council of Europe, 2001) when these appear in the body of the text and for the name of buttons and screens of mobile applications. Words from languages other than English (including the abbreviation of textual markers such as *ibid*.) have been italicized. This does not apply to the word "etcetera". The acronym *CEFR* itself has been treated as a noun and thus it frequently appears either as the head of noun phrases or as a noun premodifying the head of noun phrases. Since at some points of the dissertation the use of numbers will be very frequent, we have chosen to use numerical notation for quantities rather than words (*i.e*. we have chosen to write "2 tests" rather than "two tests"), with the exception of those cases in which numbers appear at the beginning of a sentence or in hyphenated structures (*i.e*. we have preferred to write "three-parameter model" rather than "3-parameter model"). The ambiguous cases in which the numeral "one" might be

considered as an emphatic equivalent of the indefinite article "a" (Quirk *et al*. 1985:273-274) have been treated differently. In such cases the numerical notation has only been preserved when the meaning of the word clearly referred to quantities. Figures and tables are numbered according to the chapter and section in which they appear, which helps to retrieve them in case the reader needs to go back or forth to any of them. Additionally, as can be seen in the introduction, the present dissertation contains a glossary which uses the same corpus of bibliographical references that the PhD does. This glossary can be found after part 4, before the references section, which follows an adaptation of the Chicago referencing style. In the glossary we include all the terms marked with the symbol ► along the chapters, a symbol which will only appear the first time the term is used. The objective of this glossary is to avoid explanatory footnotes which hamper the natural flow of reading. If a term marked with ► is familiar to the reader, there is no need to look it up in the glossary. In this way readers will be able to set their own pace of reading according to their familiarity with the terminology used.

PART 1

THEORETICAL ASPECTS

We feel clearly that we are only now beginning to acquire reliable material for welding together the sum total of all that is known into a whole; but, on the other hand, it has become next to impossible for a single mind fully to command more than a small specialized portion of it. I can see no other escape from this dilemma (lest our true aim be lost for ever) than that some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them -and at the risk of making fools of ourselves. So much for my apology.

Schrödinger (2013)

CHAPTER 1. A CONSTRUCT OF LANGUAGE

Bachman (1995:81) pointed out that if we are to develop and use language tests appropriately for the purposes for which they are intended, test development must be based on clear definitions of both the abilities that we wish to measure and on the means by which we observe and measure these abilities. In the case of language testing this entails a definition of language (the abilities we wish to measure) and a definition of the test technology and rationale used (the means by which we observe and measure abilities).

Chapter 1 deals with the first point, with the definition of the ability we wish to measure which, in our opinion, comprises at least 2 levels of analysis namely 1) the definition of the human ability of speech in biomechanical and cognitive terms (*i.e.* how is it possible for humans to produce sounds that label ideas? and how is it possible for others and ourselves to interpret them?), and 2) a definition of the theory of language proficiency (*i.e.* what is a good and a bad user of the human ability described in the previous point?). These 2 levels of analysis are the sections into which chapter 1 is divided and, when combined, provide a definition of the characteristics of the skill and level assessed (*i.e.* oral skills at a B1 level in our particular case).

We will define our conception of language (as well as of assessment) through the notion of construct►. Defining a language construct is one exercise of scientific honesty that teachers, test designers, raters► and linguists are continuously revisiting or, at least, they should be revisiting. Explicitly or implicitly (*i.e.* through academic papers or on the grounds of self-reflection) we all make our own sense of what languages are and this view inevitably shapes the way in which we teach, design tests or rate candidates. Thus our underlying construct of language ability will shape our choice and definition of assessment criteria, criteria which should be based on principles of good measurement (Taylor, 2003:2).

But, what is a construct? A construct is an abstract idea of something and the meaning we make of it. Trying to define instances of "love", "intelligence", "anxiety" or "thoughtfulness" is a good way to understand the term "construct" (Fulcher, 2007:7). Constructs are not, however, simple abstractions of ideas. On the contrary, for any of these everyday concepts ("love", "intelligence", etc.) to become a construct, they must have 2 further properties. First, they must be defined in such a way that they become measurable and, second, any construct should be defined in such a way that it can have relationship with other constructs that are different (*ibid*.). Fulcher (*ibid*.) quotes Kerlinger and Lee (2000:40) to point that

> concepts become constructs when they are so defined that they can become "operational" – we can measure them in a test of some kind by linking the term to something observable (whether this is ticking a box or performing some communicative action), and we can establish the place of a construct in a theory that relates one construct to another.

The original words by Kerlinger and Lee (2000:40) were:

> [A]s a scientific construct, "intelligence" means both more and less than it may mean as a concept. It means that scientists consciously and systematically use it in two ways: (1) it enters into theoretical schemes and is related in various ways to other constructs (we may say, for example, that school achievement is in part a function of intelligence and motivation) and (2) "intelligence" is so defined and specified that it can be observed and measured (we can make observations of the intelligence of children by administering an intelligence test, or by asking teachers to tell us the relative degrees of intelligence of their pupils).

To establish the construct of language that underlies the present dissertation, we will first go through the different aspects that constitute the ability of communication to try then to establish their observable features and to relate the final whole to other constructs. Yet, this starting prospect is bound to be somewhat pretentious if we consider, yet again, the complexity of the human

ability of language. Establishing a construct is an enormous task and, as Alderson points out (2000:2):

> The fact is, however, that if we wait until we have a perfect understanding of our constructs before we begin to devise assessment instruments, then we will never begin test construction […]. Thus, testers have to get involved in test construction even though they know in advance that their understanding of the phenomenon – the construct – is faulty, partial and possible never perfectible.

And we do lack a perfect construct of language. This is not only an account of the missing gaps, but, paraphrasing Schrödinger (2013), an apology for our limitations.

1.1 A modern definition of language: biolinguistics

Biolinguistics is a relatively modern interdisciplinary field that

> sets out to explore the basic properties of human language and to investigate how it matures in the individual, how it is put to use in thought and communication, what brain circuits implement it, what combination of genes supports it, and how it emerged in our species.
>
> Di Scuillo and Boeckx (2011:vii)

In our opinion, there are 2 attractive facets in biolinguistics. Firstly, it studies language from a biological perspective, that is, as a product of (human) biology. Secondly, it fosters interdisciplinary dialogue.

Accepting the first assumption is, in our opinion, paramount for an integrative view of language. Chomsky (2000:106) has been postulating something similar and has considered "language and similar phenomena to be elements of the natural world to be studied by ordinary methods of empirical enquiry". We firmly subscribe Chomsky's naturalistic► approach. Considering language a product of the natural world has important implications since it generates an articulated construct in which language is likened to other "things in

the natural world, alongside of complex molecules, electrical fields, the human visual system, and so on" (*ibid*.) and thus it can be scrutinized with the same methods employed in the study of natural sciences.

The second assumption, interdisciplinary dialogue, is both a logical consequence of the first and a requirement to organize the parts that build the complex phenomenon of language.

This complexity has been tackled from many different perspectives, ranging from the early Platonic reflections on the nature of language and its signs (Plato, 2001) to highly complex systems of biometric analysis built upon a range of disciplines that include computer sciences, medicine or engineering to name but a few. For a discussion on the history of the biological basis of language we suggest Marx's contribution to Lenneberg (1967:443-469).

Descartes (1991:10-13), for example, was no exception to the spell of languages and he also tried to dissect and organize their complex nature. In a letter written in 1629 to his friend and former teacher, the French mathematician Marin Marsenne, Descartes argued against the feasibility of assembling a language universal to all human beings on the basis of the aforementioned complexity. Descartes wrote that, after the initial excitement provoked by the thought of creating a universal language, the project started to founder. He noticed that languages are something more than the meaning of the words which build them up. If languages were simply what words make of them, anyone would be able to learn in 6 hours a sort of universal system to communicate with the entire mankind. Descartes concluded at some point of the letter that words and languages, unlike numbers, do not have a 1 to 1 correspondence or order. Descartes thought that there was an underlying nature which prevented languages from being acquired at the same speed in which people can learn in a single day to name every one of the infinite series of numbers. He also thought that if that underlying nature were ever found, this experimental universal language would spread throughout the world very rapidly. Descartes then turns his arguments to more philosophical fields and ends up as follows:

> I maintain that such a language is possible and that the knowledge on which it depends can be discovered, thus enabling peasants to be better judges of the truth of things than philosophers are now. But I do not hope ever to see such a language in use. For that, the order of nature would have to change so that the world turned into a terrestrial paradise; and that is too much to suggest outside fairyland.
>
> Descartes (1991:13)

Descartes presented the idea of unveiling the complexity of languages as a difficult but yet feasible task. He was echoing the revival of the quest for a universal language that the Renaissance saw, which had already been suggested by Francis Bacon's criticism on the "unnecessary controversies caused by the inadequacies of existing languages" (Robins, 1976:112). Similar projects towards universal languages were carried out in England by Dalgarno or by Bishop John Wilkins (*ibid*.:114).

Beyond the Cartesian view of complexity in languages, more recently Maddieson (2009) has proved that there is not one single vowel which is common to all languages in the world. Maddieson reached this conclusion after comparing 317 different languages (L in figure 1.1 below). The analyzed languages included very exotic ones such as Mongolian, Ewe (Ghana) or Lahu (an ethnic group which spreads along China, Laos, Vietnam and Thailand). Among other things, Maddieson studied the frequency of occurrence of certain types of vowels. It was hypothesized that this would bring up the group of vowels that the 317 languages had in common (C in the figure below). Surprisingly enough, the result was that there was no sound present in all the languages studied (expressed by the mathematical symbol ∄).

The comparison of the samples of these 317 languages proved that, although certain major types of consonants are common to all analyzed languages, as it is the case of stops (Maddieson, 2009:25), there is no instance of one particular consonant realized in exactly the same way in all of them, the commonest form being a plain voiceless form found in 291 languages, which

stands for 91.8% of the total sample. Similarly, there is no vocalic sound common to all languages, the commonest being the high-front /i/ sound, present in 290 languages, which stands for 91.5% of the total 317 languages that Maddieson sampled. It is surprising and evokes the idea of linguistic universals► which are interesting in themselves because of the information that they can provide about the way in which cognitive processes are genetically encoded (Pintadosi and Gibson, 2014:736). Figure 1.1 is a graphical representation of this lack of common sounds in the study of Maddieson (2009), in which L stands for the different sampled languages and C represents the absence of a sound which is common to all of them.

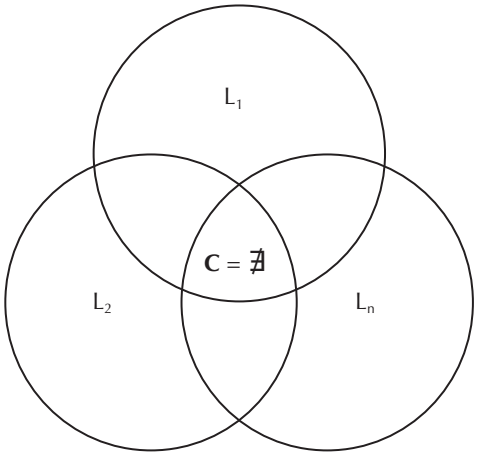

Figure 1.1. Maddieson's (2009) comparison of phonemic systems

Then, can biolinguistics put some order in all these findings and shed light on the complex nature of language? We think it can by aligning different fields of research, and it has been done so since the origins of the discipline.

As a branch of cognitive sciences, biolinguistics has its origin in the seminal works of Lenneberg and Chomsky during the second half of the 20th century, and has been extensively linked to the minimalist tradition. Biolinguistics and the generative or minimalist tradition are not the same thing but, to put it simply, generativist theories and minimalism have contributed to support biolinguistics as old MSDOS black-and-white shells contributed to modern

Windows operative systems. An average user does not need to be aware of the existence of the former to benefit from the possibilities of the latter.

In its strong version, biolinguistics is a multidisciplinary theoretical framework that "seeks to uncover the biological underpinnings of the human capacity to support language acquisition" (Boeckx, 2013:316) and is deeply concerned with genetics and the evo-devo► discussion. For us, biolinguistics will be a tool to understand the way in which various disciplines can create a construct of language departing from its biological nature. The evo-devo paradigm (a clipping of *evolution* and *development*) seeks to discover how an unprecedented novelty in the animal kingdom (language, for example) rises, survives and perpetuates itself. The evo-devo paradigm "tries to unveil, under an all-embracing conceptual umbrella, the rules and mechanisms which evolution has brought into play over time to generate the past and present biodiversity of life forms" (Baguñà and García-Fernández, 2003:465). Within the Darwinian paradigm we have assumed mutations and changes as operants in evolution, but what preconditions allow these changes to occur and prevent them from disappearing in the course of evolution? Which biological aspects of human brain store and preserve the basics of language generation after generation? Such are the questions that biolinguistics tries to answer. In our pursue of answers, new questions arise as we approach the very nature of language. Despite the fact that our definition of language is far from complete, the small steps taken through partial answers will one day be the giant leap of linguistics.

Using a biolinguistic framework as the basis of our construct of language is justified by the fact that all linguistic phenomena have their source in living beings and thus language can be analyzed as "a particular object of the biological world" (Berwick and Chomsky, 2011:19). The fact "(t)hat language acquisition requires a (possibly complex and multi-faceted) biological foundation cannot be seriously put into doubt" (Boeckx, 2013:316). However, this argument is not enough when we try to operationalize the framework, as must be done with constructs. In other words, if the framework cannot depart from abstraction to

yield practical conclusions, it is useless and must be discarded. On top of this, whatever biolinguistics is telling us about language acquisition may well not be applicable to language learning. Language learning requires different definitions from (and is at the same time linked to) biolinguistics. In this respect, if we can diagnose, for example, speech disorders through the analysis of gen $FOXP_2$, a gen whose mutation is known to be correlated with dyslexia and other language impairments (Jenkins, 2001), it would be logical to conclude that the problems derived from such mutations might affect individuals when acquiring a language and also when learning it. If one of our candidates were affected by $FOXP_2$ problems, we should treat him differently from other language students and language test-takers. We would not ask a deaf person to go through a listening test. This is, however, a residual case and there is no easy test to know if any of the candidates to our language proficiency tests are somehow affected by such problems. The example, however, is a taster of how biolinguistics might one day help language raters.

But biolinguistics is considered yet as a young discipline in the making, still in exploratory phase. Because of this, a truly integrated view of biolinguistics which favors some directions over others is a matter of placing one's bets, a necessary part in any scientific inquiry, "(i)t is more a matter of gut feelings than anything else" (Boeckx, 2013:320). In the present dissertation we will try to hedge our bets by enumerating and briefly defining how different fields of expertise have contributed to the sense that we make of language. "Biolinguistics is a fairly broad research program, and allows for the exploration of many avenues of research" (Di Scuillo and Boeckx, 2011:5) and it is precisely some of these many avenues that we will be pursuing from sections 1.1.1 to 1.1.5. The assumption is that by understanding the contributions that all these disciplines can make to our paradigm of language, we will have a clearer idea of how our tests should work and what cognitive processes these should elicit from candidates.

To finish the introduction to this first chapter, let us discuss the problem of unification►. A naturalistic approach such as ours "seeks to construct intelligible

explanatory theories, taking as 'real' what we are led to posit in this quest, and hoping for eventual unification with the core of natural sciences" (Chomsky, 2000:106). The myriad of disciplines that build our knowledge of language are disordered and related to each other, all at once, in a sort of quantum state. If biolinguistics is the leading force that we believe it to be, at some point it will not only reconcile all these variegated disciplines but will also generate a paradigm to integrate itself in "the core of natural sciences". Any attempt to bring about unification falls beyond the scope of these pages. A complete description of the human ability of speech based solely on formal analyses is like the description of one elephant based on the sounds it makes, it is only a part of the whole. At the other end of the continuum, analyzing the human ability of speech only as a product of human biology would be the same as saying that architecture must be studied by biology because it is a product of human beings as well. There must be some intermediate point at which not only classical linguistics and biology, but also other disciplines like psychology or medicine meet to provide a comprehensive explanation of the human ability of speech, a place where such disciplines are unified.

It is difficult to predict whether such unification will take place or not: "we do not know how eventual unification might proceed in this case, or if we have hit upon the right categories to seek to unify, or even if the question falls within our cognitive reach" (Chomsky, 2000:107). It is even difficult to envisage whether such unification will entail a reduction of the current constructs of linguistic theories (something which we suspect). Yet, there is "a good deal to learn from the history of the sciences since they abandoned common-sense foundations, always with some uneasiness about just what they were doing" (Chomsky, 2000:112). Let us explore new avenues now.

1.1.1 Biology

Many interesting findings have followed the biological approach started by Lenneberg's seminal *Biological Foundations of Language* (Lenneberg, 1967). The

book was ahead of its time in the perspectives it proposed for the study of language. Many linguists have found in Lenneberg's work a source of inspiration, among them, biolinguists. Di Scuillo and Boeckx (2011:1) describe the importance of Lenneberg's work as follows

> "Biolinguistics" expresses more transparently than any other term we know of what defines modern linguistics since the 'cognitive revolution' of the 1950s. Back then, under the impetus of Noam Chomsky, Eric Lenneberg, and Morris Halle, the field of linguistics abandoned its focus on external behavior and followed a path that was more decidedly cognitive –indeed biological– as it turned its attention to the organism that makes language possible.

Lenneberg was the first to describe the critical period, *i.e*. "the period during which a child can acquire language easily, rapidly, perfectly and without instruction" (Richards and Schmidt, 2002:134), but most importantly he presented and structured a series of scientific studies which, on the basis of biology, accounted for many unexplained linguistic phenomena. Lenneberg explained why biological considerations are necessary to understand major behavioral and linguistic facts, revised how human morphology and physiology correlate with language and proposed neurological considerations among other contributions. In his work, Lenneberg used approaches based on phylogeny► and ontogeny►. Phylogeny accounts for the way in which language evolves to become a distinctive adaptive trait in human beings as a species. Ontogeny, on the other hand, aims to describe the way in which this adaptive trait is developed individually in the lifespan of one individual organism. Roughly speaking, the phylogeny of language would thus define how the early forms of human communication arose in our ancestors while its ontogeny would focus on language acquisition. The ontogenetic approach is particularly relevant since it might account for the influence of mielinization◗ in early stages of language acquisition, it might explain the existence of a critical period and, in sum, a full understanding of it would help us understand language acquisition and learning processes.

Since this biological foundation the study of languages has evolved considerably over the past 70 years. Now it is known that we activate 1 muscle every 5 milliseconds when we speak. This means that about 225 muscle activations take place during every second of speech (Macneilage, 2008:4). We also know the way in which second-language learning may change the physiology of human brain (Osterhout *et al*. 2008) and we can see, even in real time, how language use activates certain cortical areas (Chen, 2006).

Genetics, generally considered a field of biology, is among the most important contributions of biology to biolinguistics. Genetics has shed light on, for example, the aforementioned gen $FOXP_2$, a gen whose mutation is known to be correlated with speech disorders. Genomic studies have made possible to zero in on the source of many disorders such as developmental dyslexia, specific language impairment or even some kinds of autism (Jenkins, 2001:128). One day, these findings might make it possible to find out through a simple blood test which linguistic impairment an individual is like to develop.

All in all, the main and most important contribution of biology to our construct of language is, perhaps, the fact that it has clearly pointed to the direction which our studies must follow. Biology should be the departing point of all our linguistic considerations. In essence, every linguist should be a bit of a surprised biologist. This requires a major change in traditional mindsets.

1.1.2 Psychology

Biology is the basis but, as we are going to see up to section 1.1.5, it is not the only component of our biolinguistic approach. Psychology itself sprang from the realms of philosophy during the second half of the nineteenth century with Wilhelm Wundt's foundational laboratory (Asthana, 2015) and ever since its early stages, psychology has been concerned with language.

In 1885, for example, the German psychologist Herman Ebbinghaus, who is credited with drafting the first standard research report (including one introduction, the methods, the results and the discussion sections), locked himself

in a room with the only company of a watch and series of non-words‣ to find out how far his memory could go in one unprecedented experiment that was probably deemed as crazy by his contemporaries. He studied series of meaningless words and calculated the time that it took him to forget them. After countless trials, Ebbinghaus came up with the forgetting curve, a statistical representation of the speed at which forgetfulness occurs in the human brain. Although Ebbinghaus's main concerns were not related to language, language was one tool used in the analysis.

Almost one century later, Skinner, another psychologist, attempted to create one of the first paradigms for the study of language leaving aside morpho-syntactic concerns. He claimed that his work was not intended to create a complete paradigm. He believed it was more "an exercise in interpretation rather than a quantitative extrapolation of rigorous experimental results" (Skinner, 1957:11). Notwithstanding this and assuming that his postulates have been largely expanded, criticized and surpassed in many aspects, he started one of the most prolific eras in the study of language. His intellectual enterprise was probably spirited up by a sense of dearth derived from the fact that early promises of a "science of verbal behavior" had never been fulfilled up to his days and it was also driven by the certainty that the final responsibility of such an enterprise had to rest, in his opinion, with psychology (Skinner, 1957:5).

The days of Skinner's behaviorism were also the days of logical positivism and the days of Luria, Vygotsky and Leontiev, among others, whose studies were not solely focused on languages (sometimes only marginally) but that had a great influence on the way in which the human ability of language was understood and even on the way in which it has been taught for decades. Behaviorism, for example, was greatly responsible for the development of the *Audio-lingual Method*. Then came cognitive psychology (which will be dealt with later in this chapter) with its theories on information processing, and constructivism and the work of Piaget, Bruner or Kelly. Humanistic approaches in psychology generated *The Silent Way*, *Suggestopaedia* and *Community Language Learning* as well. More

recently, social integrationism has presented a psychological approach that has provided a framework encompassing the insights of cognitive and humanistic perspectives. For a discussion of the evolution of all these trends, see the first chapter of Williams and Burden (2010).

Nowadays, almost 200 years after its inception, psychology is being strongly criticized during the first quarter of the 21st century. Over the last decades many scholars have raised their voices against the scientific foundations of psychology. Critics argue that the quantitative methods that rule psychology can never achieve the objectivity of natural sciences (Lane, 2012; Lilienfield, 2012; Berezow, 2012; Ferguson, 2015). For some, such criticism makes no sense at all. For others, it does and, as a matter of fact, many congresses, books and papers are now echoing with such criticism.

In the light of this, including psychology in our list of contributors to the biolinguistic interdisciplinary approach can be either controversial or a bold attempt. Criticism aside, psychology must be acknowledged as the origin of cognitive psychology, the branch of psychology that has contributed the most to the study of languages. Most likely, the cognitive processes accounted for by cognitive psychology will be some day explained by neurology, biology or any other "hard" science to reconcile (and perhaps unify) psychology with biology and linguistics. However, in the meantime, it is most honest to acknowledge the understanding of language that we have gained through cognitive psychology as well as its importance. If we accept the fact that many cognitive processes related to language (production, reception, memory retrieval, etc.) take place at a psychological level, it is also logical to see language as a means to represent the most important of these cognitive processes, as a tool for the construction of the real world or, in other words, as a tool for thought (Berwick and Chomsky, 2011:26) similar to Pinker's "mentalese" (Pinker, 2007).

There is "good evidence that the language faculty has at least two different components, a 'cognitive system' that stores information in some manner, and performance► systems that make use of this information for articulation,

perception, talking about the world, asking questions, telling jokes and so on" (Chomsky, 2005:117). The former (the cognitive system) cannot be articulated without the latter (the performance system) and the latter cannot be understood without the first. There is an "input receptive system" and an "output production system" both of which seem to access a common core body of information which links them. It makes sense to postulate that such core is common to both input and output systems for "no one speaks only Japanese and understands only Swahili" (*ibid*.). Again, all these systems are stored in the human body and if we are to understand the nature of language, we are first to understand the parts of the human body that produce it.

The theories of cognitive psychology have given rise to transformational grammar, the principle and parameters program and, more recently, to the minimalist program which, in a way, represents a highly sophisticated version of Chomsky's theories from the 1950s up to present day. The minimalist program (Chomsky, 1997) considers the existence of a universal grammar couched in the human brain that stores a computational system which restricts all forms of syntactic variations. In its "strong" variant, the minimalist thesis maintains that

> [i]n place of a complex rule system or accounts grounded on general notions of "culture" or "communication," it appears that human language syntax can be defined in a extremely simple way that makes conventional evolutionary explanations much simpler. In this view, human language syntax can be characterized via a single operation that takes exactly two (syntactic) elements *a* and *b* and puts them together to form the set {a, b}. We call this basic operation "*merge*". The "Strong Minimalist Thesis" (SMT) holds that *merge* along with a general cognitive requirement for computationally minimal or efficient search suffices to account for much of human language syntax.
>
> Bolhuis *et al*. (2014:1-2)

Cognitive psychology has unveiled elegant and scientific ways of responding to key questions about the human ability of language. Very few would challenge the idea that it owes a great deal to Chomsky, "the linguist who first

unmasked the intricacy of the system (of language)[1] and perhaps the person most responsible for the modern revolution in language and cognitive science" (Pinker, 1995:21). Social sciences had been dominated by Skinner and Watson's behaviorism until Chomsky literally confronted and challenged many of their principles in the decade of the 1950s of the 20th century. Chomsky and Skinner's confrontations are now among the most fruitful controversies in the history of linguistics. Despite their seemingly bitter disputes, the history of linguistics will acknowledge them both as the first linguists that tried to create a scientific, modern definition of the human ability of language.

The work of Baddeley (1988; 2000a; 2000b; 2003) is another remarkable example of the insight that psychology can contribute to the understanding of the human ability of language (cf. Field, 2011 and more specifically the model proposed by Levelt, 1999:87). His model of working memory► and his interpretation of what he defines as the central executive► is very helpful, for example, to understand how Chomsky's input system might store and juggle with inbound information to retain vocabulary.

In the eighties, Baddeley (1988) proposed a cognition system which departed from the traditional view of short-term memory► and described a memory model which was stable, productive and understandable (*ibid*.:70). In his model, Baddeley described a "tripartite system, comprising a supervisory controlling system, the Central Executive aided by two slave systems, one which was specialized for processing language material, the Articulatory Loop, and the other concerned with visuo-spatial memory, the Visuo-Spatial Sketch Pad or Sketch Pad" (see figure 1.1.2). The latter would be located at the right hemisphere, while the former, the phonological loop, which "is assumed to hold verbal and acoustic information using a temporary store and an articulatory rehearsal system", has been associated with Brodmann areas►, 40 and 44 (Baddeley, 2000a:417), both slightly peripheral to Broca's area►. This location of

[1] Brackets not in the original.

phonetic phenomena (derived from the use of articulatory loop) is reinforced by experiments such as Abdullaev's (2006:39), in which some subjects were asked to report if a pair of non-words rhymed. After the experiment, Abdullaev concluded that "an area of the left superior temporal lobe near the angular gyrus was found to be active", an area which coincides almost exactly with Brodmann's area 44, referred to by Baddeley. This is as much as saying that the combination of psychological (Brodmann) and medical experiments (Abdullaev) are starting to draw the map of the brain areas in which language is stored. By narrowing and studying these areas, one day we will be able to describe the basic principles that underlie language production and we will be able to use such knowledge to create more efficient teaching and assessment methods.

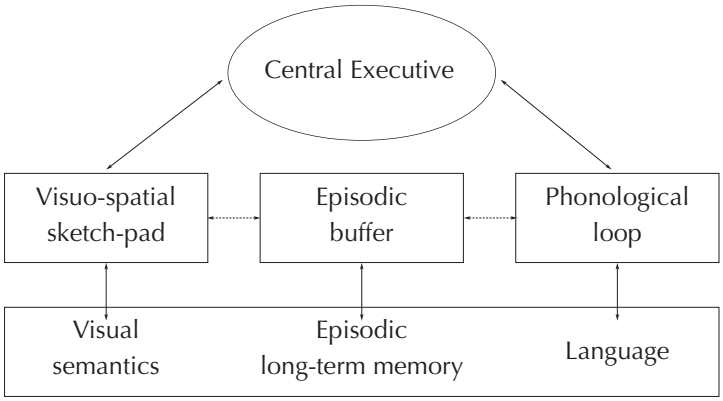

Figure 1.1.2. Baddeley's central executive diagram

The efficient use of the articulatory loop plays an important role in long-term storage of information during acquisition (Baddeley, 2000a:418) as can be drawn from the experience gathered in experiments with patients affected by short-term phonological problems: "research has shown that such patients have specific deficits in long-term phonological learning, for example, learning the vocabulary of a new language" (*ibid*.). The latter is an evident contribution of this cognitive model because, as we assumed when we mentioned gen $FOXP_2$, it would be logical to conclude that the problems derived from deficits in the phonological loop might affect individuals similarly when acquiring and when learning a language.

As we can see, psychology has evolved in parallel to the study of languages for more than 150 years. Perhaps because of its proximity to linguistics or perhaps because language output can be easily recorded, analyzed and mirrored to other brain processes, psychology has contributed enormously to the understanding of language. Psychology has given language professionals clues about attention, about the processes that underlie learning or about the role of memory in language learning and it has created powerful tools such as psychometrics►, a discipline that will be dealt with in section 2.4, which is of capital importance in our construct of assessment. Psychology has proposed models to integrate all the above as we have seen with the examples of Chomsky and Baddeley. Cognitive psychology now "combines tools from psychology, computer science, linguistics, philosophy and neurobiology to explain the workings of human intelligence" (Pinker, 1995:17) and continues to be a very powerful tool for the analysis of linguistic phenomena. As said before, most of these phenomena may be one day explained by neurology, biology or any other "hard" science but, for the time being, the relationship between linguistics and psychology seems to be in very good health.

1.1.3 Neurolinguistics

Some of the most relevant contributions that medicine has made to linguistics share 2 characteristics: on the one hand, they are based on conceptions that can be ultimately linked to human physical or psychological abilities and, on the other, these contributions are rooted in the study of impairments in patients with certain conditions. Neurolinguistics (not to be confused with the pseudo-science of Neuro-Linguistic Programming or NLP) is defined as "(t)he study of language through an analysis of deficits and impairments of language function resulting from neurological damage" (Colman, 2016). Neurolinguistics is closely related to medicine, the latter being a far wider discipline.

The contributions of Paul Broca to aphasiology and medicine are a well-known example of how physical conditions studied through neurolinguistics can

contribute to unveiling the mysteries of human language. Broca, a surgeon, physicist and anthropologist is credited with having provided the earliest proof of the localization of certain brain functions in the second half of the 19th century, on the basis of previous works by Gall and Bouillaud. Broca was the first person in history to identify which part of human brain is in charge of processing certain aspects of behavior and, more specifically, in charge of language.

By observing a patient named Tan, Broca came to the conclusion that brain functions are partially located in certain cortical areas. Leborgne, which was the real name of Tan, was nicknamed after the only word that he could pronounce: "tan". Broca observed that while his patient was disabled to articulate speech, the rest of his cognitive functions seemed to be relatively unaltered. After the death of the patient, Broca performed one autopsy and found out that Tan had a lesion in the frontal lobe of the left hemisphere. In 1861, Paul Broca wrote one article that, indirectly, changed linguistics forever:

> On 11 April 1861, a man of fifty-one years of age and named Leborgne was transported to the general infirmary at Bicêtre, into the care of surgery, afflicted with a gangrenous diffuse phlegmon of all his lower right extremity from the instep up to the buttocks. In response to the questions I asked him the following day, concerning the origin of his pain, he replied only using the monosyllable tan, which he repeated two times in a row, and accompanied by a movement of his left hand. I gathered all the information in the history of this man, who had been at Bicêtre for twenty-one years.
>
> Broca (1861:297)

Leborgne died some days after the surgery and Broca conducted his autopsy paying particular attention to his brain, the organ towards which the history of Leborgne pointed as the most probable source of his speech problems. In his article, Broca provides a detailed description of the autopsy and draws 6 conclusions from the case of Leborgne, the first 2 ones being that "Aphemia, *i.e.* the loss of speech, before all other intellectual disturbances and before any paralysis, was the result of a lesion in the anterior lobes of the brain" and that the

case of Leborgne "confirmed Mr. Bouillaud's opinion, which places the seat of the faculty of spoken language in these lobes" (Broca, 1861). Language was undoubtedly located in the human brain and the foundations of aphasiology (or aphemia, as Broca named it) had been laid.

In a broad sense, neurolinguistics attempts to describe the neural mechanisms in the brain of human beings that control the comprehension, production and acquisition of language. Neurolinguistics is highly interdisciplinary in the sense that it encompasses other disciplines such as engineering, computer sciences, neurology, neurobiology, psychology, neuropsychology and, of course, linguistics. As it is the case of biolinguistics, neurolinguistics is an umbrella term that draws on multiple disciplines. Linguists are frequently surprised when they meet for the first time such a discipline, which is normally left out from university linguistics syllabi, and they are even more mystified when, after scratching the surface, they find a well-developed scientific discipline in which linguistics plays a crucial role:

> With the rapid development of modern technology and research procedures, undreamt of or too costly in the 20th century, neurolinguistics enables scientists to make increasingly intriguing and stimulating insights into the processes governing language acquisition, functioning and production in the human brain. It is a field of research that, more than any other within the broadly defined field of linguistics, has developed significantly within the past decade and where updating one's knowledge is therefore an unquestionable necessity.
>
> Arabski and Wojtaszek (2010:xi)

The boundaries or scope of neurolinguistics have not been clearly defined yet and some other disciplines may well be included here. The mix of these disciplines has led to very important discoveries, particularly in what refers to the analysis of language processing through techniques like brain tomography or event-related potential, just to quote some.

In our opinion, neurolinguistics currently plays a very important role in the advance of our knowledge of language, at the same level of biolinguistics. Indeed,

it is difficult to predict whether one day neurolinguistics will take over biolinguistics as a unifying force or if it will happen the other way around. Eventually, none of them may be the leading force to unification. In history, unification between 2 disciplines has frequently entailed a radical revision of the more "fundamental" science, the case of chemistry and physics being a relatively recent example (Chomsky, 2000:106). Since we are inclined to think that biolinguistics is more "fundamental" than neurolinguistics, we also think that the former will absorb neurolinguistics and not the other way around.

1.1.4 Computer sciences

As used by generativists, for example, computer sciences have helped to analyze structural parameters such as recursion►. Computer sciences are nowadays oblivious in the work of any linguist, as well as in the work of virtually all types of professionals regardless of their field of expertise. From the computer that edits our papers and the books that we print, to the loads of cloud computing that allow us to process massive amounts of data, computer sciences are here to stay. In the next chapters, computer sciences will be a very important part of the present thesis, as we will see in the second part of the dissertation, where we will describe and create a computer-based tool for the assessment of oral skills.

First, in chapter 3 we will use computer sciences to carry out the complex mathematical calculations that are necessary for the statistical validation of our rubrics. To measure cognitive abilities, psychometrics (which will be described in section 2.4) requires the application of mathematical formulae thanks to which the results of tests can be interpreted from different perspectives.

There are various software packages used for such purposes, employed to obtain different types of psychometric results, the most famous of which is SPSS► (IBM Inc., 2016). SPSS (*ibid.*) is a software package originally released in 1968 by SPSS Inc., a software house headquartered in Chicago, and later incorporated in Delaware. After its release and success it was acquired in 2009 by the corporation IBM for a whopping $1.2B (Dicolo, 2009), which gives a taster of the importance

that this type of software may have. SPSS (IBM Inc., 2016) was not specifically designed for the study of languages but the psychometric analyses that it allows, also used in psychology, can be easily adapted to the study of tests. Thanks to this software, for example, test developers can get to know how difficult one test or item was when compared to others. This software also offers the possibility of analyzing the discriminatory properties of items in language exams (*i.e.* how good are items at pointing at strong or weak candidates). Another analysis frequently carried out through SPSS (*ibid.*) is that of internal reliability►, which consists in analyzing to what extent a particular language test would yield identical results if it were repeated different times in similar samples of population. All these are part of the descriptive statistics that the software allows, but it also is able of carrying out analyses of bivariate statistics (t-tests, ANOVA, correlations, nonparametric tests, etc.), predictions for numerical outcomes (linear regression) and prediction for identifying groups (factor analysis, cluster analysis, etc.).

The main advantage of SPSS (IBM Inc., 2016) is that it has a user-friendly interface and that it can be used similarly to conventional spreadsheets. The analysis of such data is complex but SPSS (*ibid.*) has a very wide community of users worldwide that has pushed the package to its current 24th installment. There are plenty of video tutorials online which help to avoid a steep learning curve, as well as online and printed manuals, the most famous of which is Field (2014), perhaps because of the humorous way in which it is written.

Still in the field of data analysis applied to the study of language and tests, we find a second group of software packages which are used to obtain data other than the above mentioned. These other packages include 3 Australian programs, Winsteps► (Linacre, 2016), Facets► (Linacre, 2014) and R► (RDCT, 2016a). These last 3 packages can perform powerful data analysis based on probabilistic theories that are not possible for SPSS (IBM Inc., 2016). On the downside, Winsteps (Linacre, 2016), Facets (Linacre, 2014) (which we will use extensively in section 4.2.4) and R are not user-friendly and require notions of mathematics and of programming.

R is a language and software environment for statistical computing and graphics developed through free GPL (General Public License), which allows free distribution. It first appeared in 1993 built on previous S language and environment software developed at Bell Laboratories:

> R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible […]. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
>
> RDCT (2016b)

Winsteps (Linacre, 2016) and Facets (Linacre, 2014) have been developed by Mike Linacre (McNamara and Knoch, 2012) and are extensively used in the analysis of data derived from language tests and for the validation of rubrics, as will be shown in 4.2.4. These are based on Rasch probabilistic models of computing which will be dealt with in more detail in sections 2.4.1 to 2.4.4.

As said before, the last 3 software packages are very powerful tools but pose difficulties in their use due to the complexity of their interface and due to the fact that their operation requires mathematical knowledge. They also have a very active community of users worldwide (more reduced than SPSS's community) in which researchers and developers keep developing the packages. The developer of Winsteps (Linacre, 2016) and Facets (Linacre, 2014), Mike Linacre, is one of the most active members of these communities and participates frequently in online debates about the software and its applications. While SPSS (IBM Inc., 2016) has penetrated in Spain, the other packages have not and are practically unknown in our country for most researchers.

In chapter 5 we will also make extensive use of computer sciences to build a mobile application to implement our rubrics. We will create an offline data storage model based on WebSQL►. The application will be created through Apache Cordova► (*Apache Cordova*, 2016), an open-source mobile development framework. Apache Cordova (*ibid*.) allows programmers to use standard, common

web technologies such as HTML5►, CSS3► and JavaScript► for cross-platform development. Thanks to Apache Cordova (*ibid*.) we have been able compile the code of the application and export it on to different platforms such as Android► or Linux► during the development stages. The application that we have developed also contains a second layer coded through AngularJS► (*AngularJS*, 2016). The use of this second layer was intended to make the application run smoothly by avoiding load times between screens. However, this will be dealt with in depth in chapter 5.

To finish, let us say that very little is taught in Spanish universities about computer sciences applied to linguistic investigation in general and to psychometric research in particular despite the fact that, as we can see, they are central to the study of languages. This probably derives from our proximity to the British tradition which, unlike the American one, has not emphasized psychometric considerations traditionally (McNamara and Knoch, 2012). A subject on computer sciences applied to psychometrics should be compulsory in linguistic syllabi in Spain. It would be beneficial not only for undergraduates who want to start a career on testing but for all types of students since it provides them with the most basic tool to measure language. It would also provide Spanish degrees in linguistics with one strategic advantage over similar European degrees in which psychometrics have not either penetrated.

1.1.5 Linguistics

Finally, in the compendium of sciences that build up our construct of biolinguistics, linguistics itself cannot be omitted. Linguistics, as well as other sciences, arises from the development of human self-awareness (Robins, 1976:2).

The Greeks were perhaps the first people with such a degree of self-awareness that they could enter the realms of reflection upon language. The Mycenaean, a Greek civilization, flourished between the 15th and the 13th century BC. Their linear B writing is considered today one of the earliest examples of script including individual sounds and not simply syllables, in contrast to

Babylonian or Assyrian cuneiform. In the time of Mycenaean, Egypt had expanded to control much of Syria-Palestine, the Hittites were controlling much of Anatolia and North Syria and the trade between the 3 vortexes of the triangle formed by Greece, Turkey and north Egypt was bustling. It is thrilling to imagine the hectic lives of the peoples that lived in such cities 35 centuries ago and the way in which linguistic exchange took place these days among different cultures.

Let us imagine that one hot day of June a young merchant from Thebes wakes up and heads for his job at the local market where he has to prepare an important journey to the palace of Knossos, in Egypt, where the king's court was preparing a summer celebration and where he thought that he would be able to sell his goods. Our imaginary merchant might have been famous for the quality of his food but also and most importantly for his ability to learn other languages beside his mother tongue, a sort of proto-Greek dialect.

If he had not been confident on his possibilities to communicate with other peoples, he would have probably not set off on such a long journey through land and sea. Thanks to his trading with the palace of Knossos he also learnt to use the symbols that the administrators of the palace employed to keep track of transactions. Our merchant learnt these symbols and became able to use them for his accounts in many other travels across Greece, Turkey and north Egypt, where he was from. Probably, more than once he had to trade with merchants from other places and with other languages and had the opportunity to use such symbols to make the best of his trading. Our imaginary trader may well have been one of the first persons not only to be able to speak different languages but also to write and to translate written symbols into these languages. Although our merchant is just an imaginary example, he illustrates how the early needs of communication in languages other that one's mother tongue might have risen for the Greeks.

But the invention of writing was previous to the Greek and it is logical to think that contact between tribes may have made possible for other people to speak different dialects before our merchant was born. The example is only meant

to exemplify that theoretical linguistics had its origin in Greece due to practical requirements and to the fact that Greece was the first culture to gain self-awareness in favor of linguistic speculation (Robins, 1997:13 and see also Howatt, 2004). The Greek did not only begin to democratize writing, they also spoke foreign languages► and reflected upon such phenomena, Plato's *Cratylus* (Plato, 2001) being a groundbreaking example.

Speculation about languages, a sort of proto-linguistics, was incipient and it had to go through its own mythical period. More or less 10 centuries after our merchant was born, the common belief that all languages were derived from a seminal one was established in the Ancient World. This first language would have been latent in any human being from their birth until a second one was learnt, overlapping the former, which might be never manifested, according to the tradition. In this respect, the Greek philosopher Herodotus (1996:117-118) narrates in the second book of his *Histories*, devoted to Euterpe, that the Egyptian king Psammetichus wanted to find out (after the above-mentioned belief about a seminal language) which one was the most ancient people of mankind. The assumption was that the most ancient people of mankind would have been the speakers of the mentioned seminal language. To find out which language was spoken by men if no mother tongue was taught, Psammetichus had a pair of newborns looked after by a shepherd under the order that nobody would talk to them. The idea for Psammetichus was that the first original language would appear sooner or later in the children if nobody talked to them and that this would demonstrate which one was the mother tongue of the first people ever. According to Herodotus, now regarded as the first scientific historian, the first word which the kids uttered was "becos", a Phrygian form for "bread", suspiciously similar to the bleat of sheep with which the children had been raised by the shepherd. However suspicious this was, Herodotus narrates that hearing the word "becos" was everything Psammetichus needed to confirm the superiority of the Phrygian civilization and their language as the mother tongue of mankind. If we found in the Book of Judges (12:4-6) the first instance of one pronunciation

test, Psammetichus's experiment may well be considered the first experiment in cognitive linguistics, and the newborns viewed as the first documented case of feral children.

Rome inherited a great deal from Greece and even so, in "passing from Greece to Rome we enter a very different world" (Robins, 1976:45). At the age of maximum power, some of Rome's eastern provinces were already largely under the influence of Greek administration and culture (*ibid*.:46). In the western part of the empire Latin took over most of the languages of conquered areas. The Romans thus welcomed multilingualism and we are even told about king Mithridates of Pontus (120-63 B.C.), who was able to speak in more than 20 different dialects or languages (*ibid*.:47). The introduction of linguistic studies into Rome is credited to one picturesque anecdote. It is known that Crates, a Stoic philosopher and grammarian, went to Rome on a political delegation in the middle of the second century B.C. In his travel he fell on an open drain and was detained in bed with a broken leg, where he passed the time while recovering in giving lectures on linguistic and literary themes:

> It is probable that Crates as a Stoic introduced mainly Stoic doctrine in his teaching; but Greek thinkers and Greek learning entered the Roman world increasingly in this period, and by the time of Varro (116-27 B.C), both Alexandrian and Stoic opinions on language were known and discussed. Varro is the first serious Latin writer on linguistic questions of whom we have any records […]. The number of his writings was celebrated by his contemporaries, and his *De Lingua Latina*, wherein he expounded his linguistic opinions, comprised twenty-five volumes, of which books 5 to 10 and some fragments of the others survive.
>
> Robins (1976:47)

Remmius Palaemon, Quintilian, Priscian, Donatus were other Latin scholars who devoted all or part of their work to different aspects of linguistic reflection and took the discipline to the Middle Ages.

After the fall of the Roman Empire, Latin remained to be the *lingua franca* of culture and thus it is not mystifying that many of the above mentioned reflections on language and linguistics made their way into the Middle Ages. In this context we find that "linguistic scholarship was supported by etymological and lexicographical work such as is well known from the pen of Isidore of Seville" (Robins, 1976:70). The Middle Ages see Ulfilia's translations of the New Testament and the birth of a new alphabet, envisaged by St. Cyril, which is the one used today by Russian and some other Slavic languages. All these are interesting and yet disordered examples of works on linguistic matters which, obviously, were oblivious to the feeling of belonging to a discipline that had not yet been defined. In England we have the works of the Bede the Venerable, Alcuin or King Aelfric's *Latin Grammar* and *Colloquium*.

> In the history of linguistic science, the second part of the Middle Ages, from around 1100 to the close of the period, is the more significant. This was the period of scholastic philosophy, in which linguistic studies had an important place and in which a very considerable amount of linguistic work was carried on. This same era is also marked by the flowering of mediaeval architecture (the so-called 'Gothic') and literature, and the founding of several of the earliest universities of Europe.
>
> Robins (1976:70)

Alexander of Villedieu's *Doctrinale* on Latin grammar, the *First Grammatical Treatise* by an unknown Icelandic scholar of the 12th century, the different *De Modis Significandi* (speculative grammars) of various medieval authors, Peter Helia's work in the application of logic to linguistic matters or Roger Bacon's appreciation of Hebrew grammar are other works that prove that the Middle Ages were not so dark an age. For a wider discussion see Robins (1976:66-93).

The Renaissance was influenced by the work on Hebrew grammar of scholars like Clénard and the first vernacular grammars in Italian, Spanish, Polish or Old Church Slavonic. In England, where French was systematically studied and

taught, Palsgrave's *L'esclarcissement de la Langue Françoyse* was very influential as well (Robins, 1976:100). The study of these new Romance languages, triggered by Dante's late mediaeval *De Vulgari Eloquentia*, became equally relevant for scholars who were ready to discuss the way in which languages changed. Nebrija (1984) wrote his *Gramática de la Lengua Castellana*, a very-well known book for Spanish linguists. In parallel to this new interest, Latin and Greek did not cease to be studied and the invention of printing popularized dictionaries as well as translations of the Bible into different European languages, all of which required implicit linguistic reflection.

During the Renaissance, Europe also realized that there were linguistic works in other parts of the world beyond Arab influences or beyond the aforementioned interest in Hebrew. Of course, this does not mean that these linguistic traditions were necessarily originated in the Renaissance. As an example, although the nature of Chinese is known in Europe only from the end of the sixteenth century, its character writing system had been in use at least from about 1400 to about 1200 B.C. (Robinson, 2003:183). Robins (1976:103) writes:

> From the New World, grammars of Nahualt (Mexico), Quechua (Peru), and Guarani (Brazil) were published in 1547, 1560, and 1639 respectively; in Europe a Basque grammar appeared in 1587 and the seventeenth century saw grammars of Japanese and Persian published.

Bright wrote about phonetics in the 16th century, Holder published his *Elements of Speech* in 1669. The 17th century also saw the rise of Port Royal grammarians and their works on the philosophy of language. In the 18th century we find Murray's influential *English Grammar* and then came the 19th century, which was very important for linguistics. Robins (1976:133-134) writes:

> In linguistics many of the scholars whose work was done in the nineteenth century are known to students well before they consciously delve into the history of the subject. Grimm, Whitney, Meyer-Lübke, Max Müller, Brugman and Sweet are just a few examples of nineteenth-century scholars who were partly

responsible for shaping their branches of linguistics in the broad patterns still taught in present-day textbooks.

The 20th century has been, nevertheless, the century in which more advances in linguistics have taken place (Howatt, 2004). The year 1916 saw the publication of the *Cours de Linguistique Générale*, Saussure's posthumous and highly influential work. Saussure was the first who attempted to delimit the object of study of linguistics. Then came Skinner, Chomsky and the array of traditions and schools that have characterized the last century, most of which have been referred to in section 1.1.2 and that we shall not repeat here.

As we can see, until the eve of the 20th century, virtually all the work on linguistics was formal and investigated only the most easily accessible manifestations of language such as morphology, syntax or the like. In this respect, if modern linguists continued to limit themselves to these concerns, in essence, it would be like perpetuating the linguistic mainstream of the Middle Ages or the Renaissance.

It might be surprising not to have mentioned linguistics up to this point in a section that aims at defining a construct of language, but it has been our intention to present a picture as wide as possible of the factors that lead to the present final remarks for section 1.1.

Linguistics, in its traditional form, has primarily been concerned with the factual materialization of language (*i.e.* words, phrases, sentences, sounds, meanings, etc.) and not until very recently in historical terms has it started to use the help of other disciplines or perspectives, integrating them into new paradigms for the analysis of languages. Disciplines related to or derived from modern linguistics are phonetics, phonology, syntax, cognitive linguistics, semantics, pragmatics, sociolinguistics, applied linguistics, anthropological linguistics, psycholinguistics, forensic linguistics (Richards and Schmidt, 2002:312), etc. On top of this, a "range of other disciplines, from the study of literature to computer science, deal with language in one way or another, and the boundaries between them and linguistics are not fixed" (Mathews, 2014). The relationship of

linguistics with other disciplines is so blurred that linguistics has been defined as "any investigation of language and languages if not clearly belonging to some other discipline, such as philosophy, the study of literature, etc." (*ibid*.).

Such a definition is discouraging but at the same time goes to show that our discipline is in dire need of a shift in what defines it. Drawing a map of all the disciplines that build what we consider linguistics or even defining what we consider to be language or its components would be difficult, tedious and, at its best, the final outcome would probably be very different from the map of other colleagues (Bachman, 1995:85-87; Field, 2011:77; Fulcher, 2015:199). Of course, all linguists have an idea of what linguistics is for them, but this conception draws upon the internal meaning that they make of their experience and beliefs. In other words, there are as many definitions of linguistics as linguists, and this is not operational.

It is precisely at this point that biolinguistics comes in handy ever since it offers a logical departure point for the study of language as a product of nature to be studied at the core of natural sciences. The real challenge that linguistics must face in the mid-run is whether to be oblivious to other disciplines and thus perpetuate itself as an incomplete tool for the analysis of languages or to accept the challenge of merging with such disciplines to provide more comprehensive results.

1.2 A theory of language proficiency

At the beginning of chapter 1 we quoted Bachman (1995:81) to underline that

> if we are to develop and use language tests appropriately, for the purposes for which they are intended, we must base them on clear definitions of both the abilities we wish to measure and the means by which we observe and measure these abilities". In the case of language testing this entails a definition of language (the abilities we wish to measure) and a definition of the test technology and rationale used (the means by which we observe and measure these abilities).

After having proposed a different look at the definitions of language and linguistics in section 1.1, in section 1.2 we will zero in on more specific characteristics of language such as the type of language that we want to measure and the levels at which we will try to do so. Section 1.2 is thus devoted to identify what bits of language we want and can measure, and at which levels we will do so.

Notice that there is not anything like a perfect test or a perfect task, let alone a test that is suitable for all purposes in all contexts. Due to this, it is necessary to define the language-specific factors that influence language production in order to make sense of our research on testing and to provide a framework (cf. McNamara 1996:54), in our case, the framework in which our rubrics operate.

It is our contention that the more accurately we define what we want to measure with our rubrics, the more scientific and measurable our results are likely to be. To separate the wheat from the chaff there are still certain questions that need to be answered as for example whether there is any difference between acquisition and learning from the biolinguistic perspective or what characteristics does oral language, the focus of our assessment, have when compared to other skills. Sections 1.2.1 to 1.2.3 narrow this framework a little bit more.

1.2.1 Acquisition and learning

Since biolinguistics is primarily concerned with language learning, this discipline has laid very little emphasis (if any) on differentiating language acquisition and language learning. In the realm of biolinguistics there are not conclusive studies that prove that language is stored or processed differently when acquired or when learnt. Those of us who have learnt languages know that it is virtually impossible to master a second (or a third) language as we master our mother tongue. This might be due to the fact that the same system, language, yields different levels of success depending on the moment at which input is received and not because

first or second languages► are stored differently in our brain. It might also be due to different cognitive processes or to a myriad other aspects.

The theories of Krashen (2002) are well-known in this respect. He establishes a distinction between the way in which one person learns a mother tongue and a second one. All the processes that he describes are quite logical in themselves. In our opinion, both language acquisition and learning must be considered, again, as different characteristics of the same human trait, language. To illustrate the point, let us consider the following question: is it possible to learn a second language without having learnt a first one? This seems a contradictory question, but also a question very easy to answer. Perhaps it is also contradictory to think that learning a second language requires biological abilities different from those necessary for a first one. If we go back to the idea posed in 1.1.2 about the faculty of language having a cognitive system that stores information and a performance system that makes use of this information (Chomsky, 2000:117), it makes sense to postulate that both systems share a core a common pool of information and mechanisms for "no one speaks only Japanese and understands only Swahili" (*ibid*.). Similarly, no one is able to speak a second language without speaking a first one.

Krashen (2002:70-82) tries to establish what he calls the "neurological correlates of language acquisition" by making reference to Lenneberg and others' experimental work but is not conclusive about the fact that there is a biological distinction between acquisition and learning, perhaps because establishing such difference was not among his goals. He departs from a cognitive perspective which is maintained all along his work. Bachman (1995:107) also speaks about "phychophysiological mechanisms" in his model of language, which is briefly discussed below.

As said before, there are no conclusive studies that prove that brain functions are different when first or second languages are activated by the same individuals. While some studies claim that different languages do not require the use of different cortical areas (Chee *et al*., 1999; Hasegawa *et al*., 2002; Klein *et*

al.), others maintain that certain cortical areas are specifically activated depending on the language used (Kim *et al*., 1997; Marian *et al*. 2003; Simos *et al*. 2001). See Roux and Lubrano (2006) for a wider discussion.

Determining whether second languages share the same cortical functions as first languages will clarify whether the former can be assessed through the same methods used for the latter. Some have suggested that the faculty of language might be based on a very limited number of core properties like "merge" (Bolhuis *et al*., 2014:1-2), referred to in 1.1.2 or recursion (Vries *et al*. 2011), referred to in 1.1.4.

Our assumption in this respect is that, for reasons of evolutionary economy, the basic neurological functions associated to language acquisition are likely to be the same as those of learnt languages. If this were proved true, we would be closer to understanding the core properties of language and closer too to the ultimate object of study. We would also have a clearer vision of what needs to be tested (for a discussion, see Bachman and Cohen, 1988).

1.2.2 Competence and performance

In section 1.2.1 we have seen that there are not conclusive theories about the difference between language learning and acquisition. There is not either common agreement on what components constitute it. "The current situation is that we have good and improving theories of some aspects of language and mind, but only rudimentary ideas about the relation of any of this to the brain" (Chomsky, 2000:116). The question at this point is whether it is possible to measure something whose ultimate nature we do not know. The answer is that, with certain limitations, we can.

In the absence of such a general and broadly-accepted framework or definition of language, tradition has put linguistic output at the core of language theories. In other words, since we do not know what language is made of, we have measured what we can perceive from language: what our candidates say, write or the degree of proficiency that they show in doing it. We may not know

what Coca-Cola is made of, but we know 1 liter from 2. In such a way, we have been (and still are) measuring liters and kilos of language. Fortunately, as we will see in the lines below, the study of what is perceivable from language has led to interesting findings and has proved that the inability of present-day linguistics to account for unresolved questions is no reason at all for doubting that they will be accounted for some day.

Tradition has been measuring liters and kilos of competence► and performance►. The distinction between performance and competence was first introduced by generativists in the second half of the last century (see the discussion in Chomsky, 1965:10-15) to distinguish between the idealized, abstract idea of being capable of speaking a language (competence) and the real way in which it is manifested (performance). Hymes (1972) made interesting contributions to the debate as well as Canale and Swain (1980), who broke down the knowledge of language into grammatical, sociolinguistic and strategic competence. Bachman (1995) took Canale and Swain's framework and established a linguistic model composed by language competence, strategic competence and psychophysiological mechanisms. Bachman would later include topical knowledge in a revision of his model (Bachman and Palmer, 1996). As said before, all these frameworks have not answered all the fundamental questions about language, but have helped in the progress. For a wider discussion of these and other models see Fulcher and Davidson (2007:36-51).

These works also influenced the *CEFR*► (Council of Europe, 2001), which will be analyzed in more detail in section 1.2.3. The *CEFR* (*ibid*.) proposes a similar but also distinct version of the above-mentioned frameworks of competences, which we have used to create a tool to measure performance. The description of communicative language competences proposed by the *CEFR* (*ibid*., 2001:101-130) is the following one:

- Linguistic competences
    - Lexical competence
    - Grammatical competence
    - Semantic competence
    - Phonological competence
    - Orthographic competence
    - Orthoepic competence
- Pragmatic competences
    - Design competence
    - Discourse competence
    - Functional competence
- Sociolinguistic competence
    - Linguistic markers of social relations
    - Politeness conventions
    - Expressions of folk wisdom
    - Register differences
    - Dialect and accent

How were all these categories of performance accounted for in our rubrics? As can be seen in figure 1.2.2, in designing our rubrics, whose content will be dealt with in depth in chapter 4, we have used communicative and pragmatic competences to define linguistic features (*Language*, *Pronunciation*, *Interaction* and *Discourse*) and sociolinguistic competences to define different levels of sophistication across bands.

As a graphic representation, if we considered our rubrics as a typical spreadsheet, linguistic and pragmatic competences would be columns and different levels of sociolinguistic competences would be the rows:

			LINGUISTIC COMPETENCES		PRAGMATIC COMPETENCES	
			Lexical Grammatical Semantic	Phonological	Design	Functional Discourse
			LANGUAGE	PRONUNCIATION	INTERACTION	DISCOURSE
SOCIOLINGUISTIC COMPETENCES	+ –	5				
		4				
		3				
		2				
		1				

Figure 1.2.2. Arrangement of competences in our set of rubrics

As can be seen in figure 1.2.2, linguistic competences have been useful to define the linguistic features of *Language* and *Pronunciation* while pragmatic competences were used to distinguish between *Interaction* and *Discourse*. Sociolinguistic competences establish a rank of sophistication of performance across levels. Some of the components of sociolinguistic competences are minimally present at our level of analysis or not present at all. Take for example the case of sensibility to *register differences*, one of the components of sociolinguistic competences. Since "(i)n early learning (say up to level B1), a relatively neutral register is appropriate" (Council of Europe, 2001:120) there is no room at our targeted level, B1, to incorporate definitions about a highly formal tenor. Similarly, more elaborate manifestations of *expressions of folk wisdom* or *dialect and accent* (2 other components of sociolinguistic competences) are only expected at levels beyond B1. In other words, sociolinguistic competences are more in number and more sophisticated as we approach the top levels of the *CEFR* (*ibid*.), and that is the reason why they are more useful to describe levels rather than categories.

We did not use everything from the *CEFR* (Council of Europe, 2001) and we had to compensate for some definitions that the *CEFR* (*ibid*.) lacks. We did not use, for example, orthographic or orthoepic competences since these are proper of skills other than speaking. At times, the references in the *CEFR* (*ibid*.) to competences like the phonological ones were very scarce. All the decisions taken in this respect are further developed in chapter 4.

Design competence, the knowledge of the principles required to sequence messages according to interactional schemata (Council of Europe, 2001:123), are not profusely described in the *CEFR* (*ibid*.) and yet they are very important in our theory of language proficiency. Oral production is in its initial and critical stages chiefly interactive. We considered that interaction deserved a place in our theory of language performance and for this reason, despite their being scarcely described in the *CEFR* (*ibid*.), we went a little bit further with them. The *CEFR* (*ibid*., 2001:84) describes some of the characteristics of oral interaction as follows:

> [T]he fact that spoken interaction entails the collective creation of meaning by the establishment of some degree of common mental context, defining what can be taken as given, working out where people are coming from, converging towards each other or defining and maintaining a comfortable distance, usually in real time, means that in addition to receptive and productive strategies there is a class of strategies exclusive to interaction concerned with the management of this process. In addition, the fact that interaction is primarily face to face tends to provide far greater redundancy both in textual, linguistic terms and with regard to paralinguistic features, contextual cues.

As a consequence, we tried to reflect all these peculiarities chiefly in the linguistic feature of *Interaction* and marginally in *Discourse*, where we tried to refer to the cohesion patterns necessary to sequence messages according to interactional schemata.

When we defined the linguistic feature of *Language* we were also concerned about the risks of some traditional approaches to grammatical competence. In our theory of language proficiency, oral production does not

follow the same strict syntactic patterns of writing. "It is obvious that conversational interchange, by its interactive or reciprocal nature, gives rise to or necessitates devices for organizing discourse that are unique to this genre of discourse" (Bachman, 1995:89). We knew from the beginning that this would have implications in teacher-training sessions, in which we would have to put forth an updated approach to grammatical competence based on "idea units":

> One important point, for example, is that people do not normally speak in sentences. Rather, spoken language, especially in informal situations, consists of short phrases or clauses, called idea units, strung together in a rather loose way, often connected more by the coherence of the ideas than by any formal grammatical relationship.
>
> Buck (2001:9)

Idea units provide oral syntax with a radically different structure from that of planned, written communication. As a consequence, since written and oral syntax are different, the latter cannot be measured exactly the same as the former. This is the type of implication that we would have to make raters aware of prior to the use of our new rubrics.

Idea units are often spontaneous, unplanned, and "usually contain about as much information as we can comfortably hold in working memory" (Buck, 2001:10). Luoma (2004:12) defines idea units as

> short phrases and clauses connected with and, or, but or that, or not joined by conjunctions at all but simply spoken next to each other, with possibly a short pause between them. The grammar of these strings of idea units is simpler than that of the written language with its long sentences and dependent and subordinate clauses (…). The units are usually spoken with a coherent intonation contour, and they are often limited on both sides by pauses or hesitation markers. Many idea units are clauses with a verb phrase, a noun phrase and a prepositional phrase, but some of them do not contain a verb, and sometimes an idea unit is started by one speaker and completed by another.

Idea units are more often than not unplanned speech acts although in more formal contexts (conferences, speeches) they are expected to resemble certain parameters of written, planned speech acts. Idea units also display other syntactic peculiarities, namely topicalisation and tails. Topicalisation is an alteration of the standard syntactic structures by means of which the initial element of a sentence is given special informational emphasis (as in "Joe, his name is"). Tails, on the other hand, are noun phrases that speakers put at the end of a clause to emphasize the comment that they make at the beginning of the clause (as in "It's very nice, that road through Skipton to the Dales") (Luoma, 2004:15-16).

If we transcribed any of our everyday conversations we would have a taster of how little its syntax resembles that of the dialogues of a play. As a consequence of these differences between oral and written language, one should not expect that the candidates to a test necessarily speak as if they were reading a written text, and this will be relevant in the interpretation of the rubrics during teacher training, as stated above. In the context for which our construct is devised, in which most of the test takers are university students from different backgrounds, idea units should be paid special attention, particularly at lower levels of competence. We cannot mark the oral production of our candidates with the criteria that we would use to mark a written test and should allow certain grammatical flexibility as long as it does not interfere with the intended meaning of the message. This may seem obvious in certain contexts, but definitely not in the Spanish one, in which there is still a strong influence of teaching methods strongly based on grammar. Grammar must be used, indeed, but with certain considerations:

> Learner grammar is handy for judging proficiency because it is easy to detect in speech and writing, and because the fully fledged grammars of most languages are well known and available for use as performance standards. However, the grammar that is evaluated in assessing speaking should be specifically related to the grammar of speech.
>
> Luoma (2004:12)

1.2.3 The *CEFR* (Council of Europe, 2001) and its levels

Languages are not learnt across well-delimited stages but in a continuum. Despite this, when we deal with language learning and testing concerns we need to establish levels of proficiency to be able to allocate candidates in one level or another. Clearly delimited language levels are necessary and, at the same time, contrary to the nature of language learning.

For many years, language levels were defined in Europe through broad classifications such as elementary, beginner, intermediate, advanced or proficient user. There were also more specific attempts to define proficiency scales, the best-known of which was probably the waystage, threshold and vantage scales (Ek and Trim, 1991). Language levels in Europe were also set taking as reference reputed tests. This way, for example, having a Cambridge First Certificate level was (and still is) a shared referent of proficiency. The lack of unification and the variegation of criteria hampered mutual recognition among institutions because not all certificates were as famous as Cambridge's suite of tests, as ETS's TOEFL, or because the skills that they assessed were not equally balanced or measured. There was no empirical guarantee that the levels of a board of examination could be unequivocally aligned with those of another institution and this led institutions in Europe to only recognize officially certificates with an extensive background and expertise. The circle of recognized certificates was difficult to expand basically because there were not any specifications as to what a test should look like to be included in the list of valid certifications.

All this changed with the development of the *CEFR* (Council of Europe, 2001), which is the yardstick that we are going to follow to distinguish levels of proficiency in the present dissertation. The *CEFR* (*ibid*.) is nowadays not only the mainstream in Europe, but also in many countries of South America and in a growing number of countries in Asia.

The *CEFR* (Council of Europe, 2001) has been the most influential milestone in the recent evolution of language teaching and assessment in Europe. Curiously enough, "(t)he move towards a European economic community

clarified, as no other theoretical approach would have done, the requirement to define language teaching goals" (Spolsky, 1995:3). The *CEFR* is another byproduct of this move towards a European economic community that completes the projects envisaged by the Council of Europe► in the 1970s, oriented to clarify the objectives of language students in our continent.

The *CEFR* (Council of Europe, 2001) was published in 2001 but some years before Bachman and Clark (as cited in Bachman, 1995:6) already hypothesized and, to a certain extent, foresaw the advantages of one system as the *CEFR* (*ibid*.) when they spoke about "a common metric":

> The obvious advantage of such a scale and tests developed from it is that it would provide a standard for defining and measuring language abilities that would be independent of specific languages, contexts and domains of discourse. Scores from tests based on this scale would thus be comparable across different languages and contexts.
>
> Bachman (1995:6)

The *CEFR* (Council of Europe, 2001) does not explicitly refer to any particular language construct but it is germane to our theory of proficiency because it introduces different levels to classify speakers' interlinguas in which they are not compared with each other but with certain scales based on criteria.

In the 1990s, Brian North, the man behind the *CEFR* (Council of Europe, 2001), was working with the Council of Europe (which will be analyzed in section 3.1.1) on a project to develop a scale of descriptors► (*i.e.* statements describing levels of performance within a proficiency scale) of communicative language proficiency in different categories, which could be scalable at an ascending series of levels. The idea was to aid different providers of language teaching services to describe and compare their systems (North, 2000:2). He noticed that virtually all language scales designed to that day seemed "to have been produced on the basis of intuition and/or subjective matching to samples of performance by a small authoring team" (*ibid*.:3). To correct for this, he

developed an example set of descriptors of communicative language proficiency which

> (a) bear some relation to the theory-based categories used for the description of communicative language competence in the Council of Europe Common European Framework, which (b) built explicitly on collective experience in the field of scales of language proficiency, which (c) were felt to be clear, comprehensible and relevant by practicing teachers, and which (d) were calibrated with a measurement model in relation to the achievement of learners in different educational sectors in a multi-lingual context.
>
> North (2000:2)

And to achieve his objectives, North followed 4 steps:

> (a) analyse existing scales of language proficiency in terms of categories which can be related to theories of language and to the emerging Council of Europe Common European Framework model, and write descriptors for those aspects of proficiency which appeared to be under-represented; (b) reduce and refine the descriptor set using groups of teachers as informants; (c) calibrate the descriptors felt to be the clearest and most relevant through an analysis of the judgements of Swiss teachers using a scalar version of the Rasch model from the Item Response Theory family of measurement models.
>
> North (2000:2)

The result was the groundbreaking *CEFR* (Council of Europe, 2001), which does not only serve the intended purpose but also many others. In the absence of clearer or better-calibrated proficiency scales, the *CEFR* (*ibid*.) has become the standard with which virtually all relevant proficiency exams in Europe are aligned. Even non-European boards of exams must align their scales of proficiency with the *CEFR* (*ibid*.) if they want to penetrate in the European market (Papageorgiou *et al*., 2016).

The *CEFR* (Council of Europe, 2001) proposes 3 main bands of levels, namely A for basic users, B for independent users and C for proficiency users. Each of these bands or levels can be in turn subdivided into up to 4 other levels

which are characterized through their own descriptors. Although the different levels can be broken down according to different teaching needs, contrary to what many people think, the *CEFR* (*ibid*.) provides descriptors for 9 different levels, not only 6. The different levels and sublevels for which the *CEFR* (*ibid*.) provides descriptors are shown in figure 1.2.3 below:

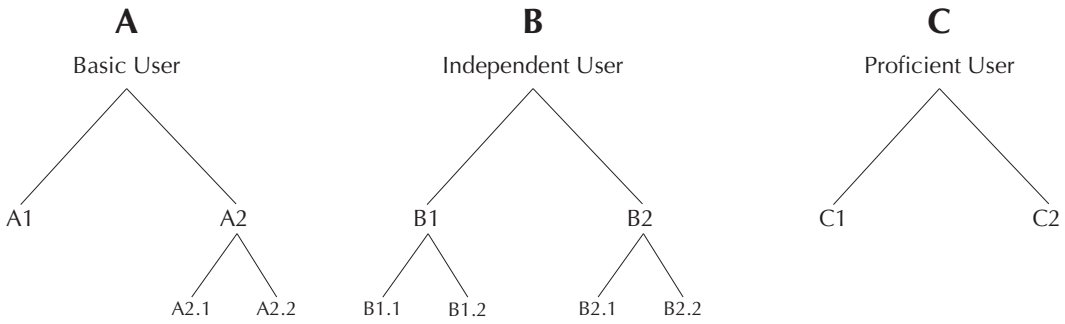

Figure 1.2.3. The *CEFR* (Council of Europe, 2001) levels and sublevels

All these levels are described according to different aspects of language (from global scales to speaking fluency, spoken interaction, turn taking, creative writing, orthographic control or sociolinguistic appropriateness to name but a few). The system constitutes a highly valuable and referential yardstick. Due to its comprehensive nature, the *CEFR* (Council of Europe, 2001) is many things at the same time. It is a reference manual that describes the implications of learning a foreign language but also a scale which separates different learning stages.

The numerous *CEFR* (Council of Europe, 2001) descriptors are particularly well known because they represent a referent for teaching and assessment. The *CEFR* (*ibid*.) is thus equally used by teaching professionals as well as by test designers. Test designers frequently try to link their items and tasks to the different levels that the *CEFR* (*ibid*.) establishes, and they do it with certain limitations.

But the *CEFR* (Council of Europe, 2001) is neither perfect nor complete. The theory of language proficiency that it presents, greatly influenced by Wilkins (1976), Canale and Swain (1980) and Bachman (1995), has been criticized by those who claim that it provides little or no information on key aspects of

assessment, something which is acknowledged by the *CEFR* (Council of Europe, 2001:xi) itself:

> One thing should be made clear right away. We have NOT set out to tell practitioners what to do, or how to do it. We are raising questions, not answering them. It is not the function of the Common European Framework to lay down the objectives that users should pursue or the methods they should employ.

To bridge this gap between the *CEFR* (Council of Europe, 2001) and language assessment, the Council of Europe and other assessment boards have published manuals to relate examinations to the framework (ALTE, 2011). These fields are now developing astonishingly fast in parallel to the use of the *CEFR* (Council of Europe, 2001) as one assessment tool, and they are generating very interesting literature in relation to language assessment and measurement.

CHAPTER 2. A CONSTRUCT OF TESTING

We opened chapter 1 with Bachman's (1995:81) assertion about the fact that if we are to develop and use language tests appropriately for the purposes for which they are intended, test development must be based on clear definitions of both the abilities that we wish to measure and on the means by which we observe and measure these abilities. Chapter 1 has been devoted to the definition of such abilities, and we shall devote chapter 2 to the definition of the means necessary to measure them. How do our definitions of language and assessment relate to each other?

> For our purposes, we can consider a construct to be the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task. The construct definition for a particular assessment situation becomes the basis for the kinds of interpretations we can make from the assessment performance. In designing, developing, and using language assessments, we can define the construct from a number of perspectives, including everything from the content of a particular part of a language course to a needs analysis of the components of language ability that may be required to perform language use tasks in a target language use domain, to a theoretical model of language ability.
>
> Bachman and Palmer (2010:43)

Throughout chapter 2, our construct of language will assist us in the definition of assessment not only at the level which is defined by Bachman and Palmer (2010:43) above (*i.e.* at a test-developing level), but also to define the very nature of assessment, which is precisely what will be found in this part of the dissertation. In these sections we will delimit our construct of assessment firstly by defining the differences between assessment, test and evaluation, secondly by making reference to the particular tool that we are going to develop for our experiment, a set of rubrics, and finally by establishing the mathematical parameters of our system of measurement.

2.1 Measurement, test and evaluation

Time and again, the terms "measurement", "test" and "evaluation" are used as synonyms. Although they are frequently interchangeable, a more accurate definition of them is necessary to the proper development and use of language tests (Bachman, 1995:18).

Measuring involves the quantification (*i.e.* what numbers, letter grades, etc. we are going to use to measure), the definition of the characteristics that are to be measured (*i.e.* what we are going to measure) and the definition of the rules and procedures to be followed (*i.e.* how we are going to measure).

A *test*, on the other hand, may well be deemed as "a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual", according to Carroll (1968:46). A test is then the tool that assessors use to ascertain reliable samples of linguistic behavior. Tests will be the focus of the following pages, in which we will first define the possibilities that tests offers and the different ways in which they can help raters to make valid and fair inferences about the characteristics of individual test takers.

Finally, *evaluating* is to put together all the relevant information, either elicited from tests or through other means, to make decisions on (in our case) the degree of linguistic competence of the assessed person, a process similar to clinic diagnosis in medicine.

Although the boundary among the 3 terms is not always clear, there are some ways of establishing operational boundaries. Figure 2.1 presents *measurement*, *test* and *evaluation* as interconnected rather than as mingled methods. The figure is taken from Bachman (1995:23).

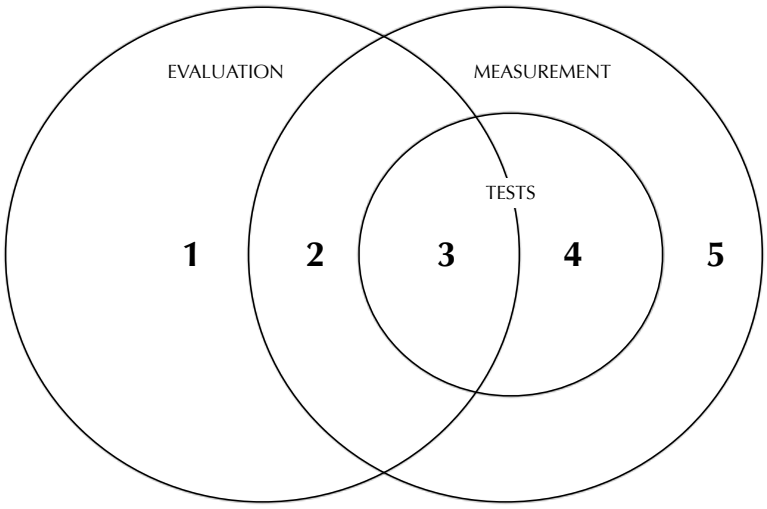

Figure 2.1. Measurement, tests, and evaluation

In this figure, as described by its author (Bachman, 1995:23-24):

> An example of evaluation that does not involve either tests or measures (area "1") is the use of qualitative descriptions of student performance for diagnosing learning problems. An example of a non-test measure for evaluation (area "2") is a teacher ranking used for assigning grades, while an example of a test used for purposes of evaluation (area "3") is the use of an achievement test to determine student progress. The most common non-evaluative uses of tests and measures are for research purposes. An example of tests that are not used for evaluation (area "4") is the use of a proficiency test as a criterion in second language acquisition research. Finally, assigning code numbers to subjects in second language research according to native language is an example of a non-test measure that is not used for evaluation (area "5").

Our research is placed in area 4. We will be analyzing the outcome of oral proficiency tests designed for candidates to prove that they have a B1 proficiency level in English. As we will see in more detail in chapter 4, these tests are designed according to a particular set of specifications shared by 9 different universities in Andalusia (a region in Southern Spain), according to which each university develops one specific B1 test. The fact that the specifications are shared while tests are different across the 9 different universities determines the nature of

the rubrics because they must fit the aforementioned specifications at the same time that they allow room for flexibility to be applied in 9 different contexts.

2.2 An updated approach to testing

In the last decades, most approaches to language testing normally focus on 4 main areas, namely reliability, validity►, fairness and practicality. Reliability is concerned with the absence of measurement errors once repeated trials have been carried out with the same instrument. Validity focuses on the extent to which our tests actually measure what they are supposed to measure. Fairness guarantees equal conditions to all stakeholders► regardless of their sex, age or cultural background. Practicality is concerned with actual rather than theoretical possibilities of implementing test-design practices.

Tests designed according to these 4 main areas of concern will also yield results which, more often than not, due to their impact, are linked (or aligned) to marking scales, in our case, the scales that we discussed in section 1.2.3, the *CEFR* (Council of Europe, 2001).

Figure 2.2 provides a visual guide of the 4 main areas, summarizes the structure of sections 2.2.1 to 2.2.4, and displays different aspects which are relevant within each of the 4 areas. Each area is broadly defined in the figure by the type of questions that they are intended to answer. As we will mention later on, very frequently the boundary between validity and reliability is not clear-cut (see section 2.2.2) and, depending on the model chosen, additional areas can be included in the conceptualization of tests.

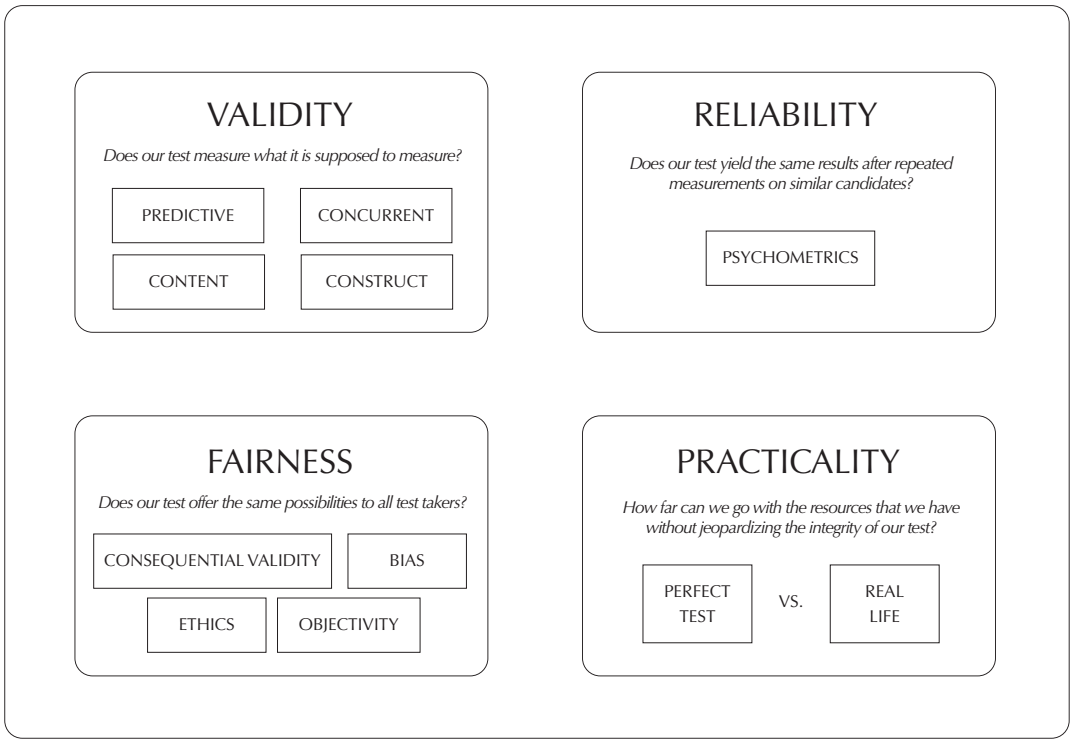

Figure 2.2. Main areas of interest in language testing

2.2.1 Validity

A test is valid if it does what it is intended to do. Consider for example the task below, taken from a real B1 proficiency test designed for undergraduate students at the University of Jaén. Candidates were given the following text to read:

The Universe, Dark Energy and Us

A Almost every scientific talk or seminar on astronomy today starts from the idea that we live in a universe in which a mysterious force known as *dark energy* makes up about 70 percent of the total cosmic amount of everything. A mysterious substance known as *dark matter* makes up about one fourth of the whole cosmos. And *ordinary matter* — the stuff of the periodic table, including interesting assemblies of matter like galaxies, stars, planets and people — is the insignificant remaining part of this cosmic equation.

Figure 2.2.1.a. Sample reading passage from a real B1 proficiency exam

The excerpt is part of a longer reading passage which candidates had to read to answer some items within the task. With the information contained in the passage in figure 2.2.1.a above, candidates had to complete the following task:

Subsection 1 (**questions 16-20**). Complete the blanks in the graph below on the structure of the universe, with TWO WORDS or ONE NUMBER. Write your answers in boxes 16-20 on your answer sheet.

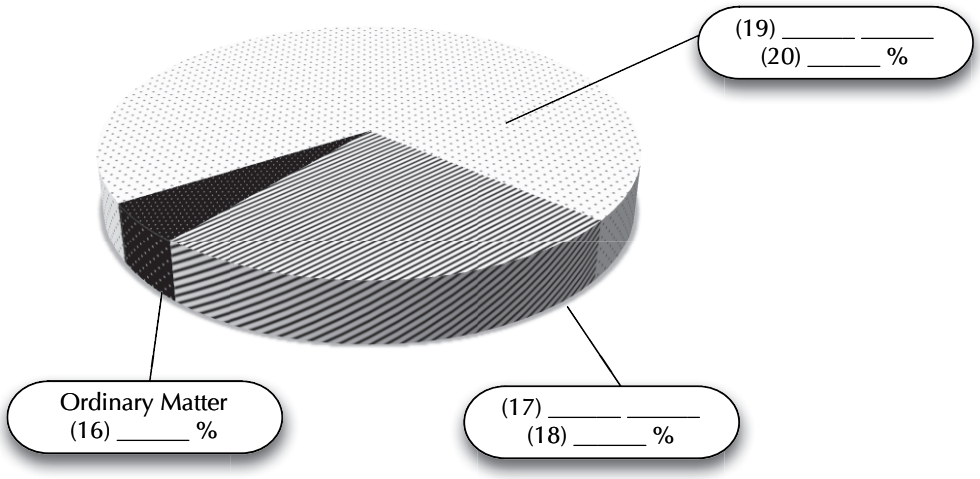

Figure 2.2.1.b. Reading task linked to the passage in figure 2.2.1.a

From the text, candidates know that the biggest area in the graph corresponds to dark energy: "*dark energy* makes up about 70 per cent of the total cosmic amount of everything". Thus the key to item 19 is "dark energy" and the key to item 20 is "70%". Candidates also know that "*dark matter* makes up about one fourth of the whole cosmos". Then, the answer for item 17 is "dark matter" and the key for item 18 is "25", which is "one fourth of the whole cosmos". Finally, the answer for item 16 is "5", which is obtained by resting 70% (dark energy) and 25% (dark matter) to 100% (whole cosmos).

The design of this task assumes that candidates will be able to identify the fact that "one fourth" equals 25% and also that they will be able to identify and solve the mathematical operation that is necessary to answer item 16. It is not just a question about the candidate's understanding of the text. There are other

cognitive processes involved like the ability to identify the parts of a problem or the ability to carry out small mathematical operations. If the mathematical operations are not in our construct of test, then it is not fair to produce a task with such characteristics and, consequently, the task cannot be considered as valid. This task may not be valid in, for example, an immigration test in which candidates are not supposed to belong to any particular educational background to enter an English-speaking country to work. On the other hand, this task may be valid in an academic proficiency test in which candidates will obtain scholarships to participate in scientific research programs. The tasks in a proficiency test for air traffic controllers are probably very different from the tasks of a business-oriented proficiency test. This is what we refer to as construct validity►, which is only 1 out of the different types of validity that can be found in the literature (see also construct irrelevant variance► and construct under-representation►).

Validity aims to identify whether our tests are measuring what we really want to measure. The idea seems obvious and the job easy to tackle. Of course, through language exams, we aim at measuring language, and to do it we want to elicit linguistic reactions from test-takers. However, if we take into consideration the complexity of language and of the construct that we make of it, and if we take into consideration that the same test may not be suitable for 2 different candidates even though the test aims to test the same thing in both of them, then, our initial perception of simplicity changes.

Early validity studies emerged after the Second World War (Fulcher, 2007:4). Crombach and Meehl (1955) are frequently considered among the first to tackle validity investigation. They described criterion-oriented validity, which included predictive validity► and concurrent validity►, (cf. Khalifa and Salamoura, 2011), content validity► and construct validity. This division is still the most extended today (Davies *et al*. 1999:221-222, Fulcher and Davidson, 2007:3-22) but other types of validity have been defined ever since, as for example face validity► (Davies *et al*. 1999:221), cognitive validity► (Field, 2011), context validity► (Galaczi and Ffrench, 2011), scoring validity► (Taylor and

Galaczi, 2011) or consequential validity► (Hawkey, 2011). Their definitions frequently overlap and this overlapping, yet again, gives the impression that there is no common ground of agreement among scholars. Despite the existence of so many types, psychometricians (see section 2.4) have increasingly come to view them all as part of a single, unitary concept of validity (Bachman, 1995:241-243; Fulcher and Davidson, 2007:12).

In the main scheme of predictive, concurrent, content and construct validity, construct and content validity can be considered as types of "internal" validity in the sense that they relate to the internal characteristics of the test (*i.e.*, the theory on which the test is based in the case of construct validity and how well the content of the test represents the targeted domain in the case of content validity), while concurrent and predictive validity can be considered examples of "external" validity in so far as they link our tests to predictions on external factors (for example, concurrent validity helps to predict how test takers' scores on one test relate to those on another and predictive validity establishes the relationship between the scores on one test and the actual ability of test takers to perform a particular task in real life).

The *CEFR* (Council of Europe, 2001:177) considers validity as 1 of the 3 concepts traditionally seen as fundamental to any discussion of assessment (together with reliability and feasibility), and defines validity as

> [t]he concept with which the Framework is concerned. A test or assessment procedure can be said to have validity to the degree that it can be demonstrated that what is actually assessed (the construct) is what, in the context concerned, should be assessed, and that the information gained is an accurate representation of the proficiency of the candidate(s) concerned.
>
> Council of Europe (2001:177)

2.2.2 Reliability

So, if we want to have valid and reliable tests, tasks and items, how can we be sure that they work as intended? How do we validate them?

> The investigation of reliability and validity can be viewed as complementary aspects of identifying, estimating, and interpreting different sources of variance in test scores […]. The investigation of reliability is concerned with answering the question, 'How much variance in test scores is due to measurement error?' and its complement question, ´How much variance is due to factors other than measurement error?' […] Validity, on the other hand, is concerned with identifying the factors that produce the reliable variance in test scores. That is, validation addresses the question, 'What specific abilities account for the reliable variance in test scores?' Thus, we might say that reliability is concerned with determining how much of the variance in test scores is reliable variance, while validity is concerned with determining what abilities contribute to this reliable variance […]. The process of validation thus must look beyond reliability and examine the relationship between test performance and factors outside the test itself. Despite this apparently clear demarcation of the domains of reliability and validity, distinguishing between the two for language tests is not always clear-cut.
>
> Bachman (1995:238-239)

In a simpler definition, reliability is concerned with measuring and with the absence of measurement errors after repeated measurements are carried out with the same instrument. The *CEFR* (Council of Europe, 2001:177) considers reliability a technical term and defines it as "the extent to which the same rank order of candidates is replicated in two separate (real or simulated) administrations of the same assessment". This is just a working idea, one abstract exemplification, for no test can be administered 2 times to the same group of candidates with the same results. If the same test is administered 2 times to the same candidates, these are likely to remember details from their previous experience. They may have learnt from their previous errors and the test will definitely be familiar for them. This way of presenting reliability is thus an intended simplification.

Reliability aims at establishing the most accurate possible measures in our tests, turning results into mathematical expressions that can be objectively analyzed and that make our construct of language operational, one of the characteristics of constructs (see the introduction to chapter 1). The results of our tests become interpretable data thanks to psychometrics and thanks to the analysis of aspects such as Crombach's alpha►, facility value►, discrimination index►, or as kappa coefficient►. The first 3 will be defined in detail in section 2.4. The kappa coefficient is presented as an inter-rater reliability► test in section 4.2.5. Fulcher and Davidson (2003:104) claim that

> [b]y ensuring that responses to individual items are not dependent upon the responses to other items, that they have good facility values and discrimination, and that we have enough items, we can ensure that such tests have the quality of reliability.

We will discuss the mathematical rationale of these and other reliability concerns in section 2.4. For the time being, let us just say that we want the results of our tests to be as reliable as possible and let us analyze several intuitive examples that demonstrate how important reliability may be.

Sometimes it is not relevant whether a test taker scores 90 or 91 in a test marked out of 100. Some other times, this point is the one that makes the difference. In general, the higher the stakes, the more important reliability becomes. Take for example the case of Freddie Lee Hall, who was sentenced to death in the United States in 1978 accused of having killed 2 people. According to the rules of the state of Florida, where he was inmate, he was eligible for death penalty because of his crimes and because of the fact that he scored 71 in IQ tests►. The high court of the United States establishes that mentally disabled cannot be put to death, the cut off score for such disability being at 70 in the state of Florida. Lee's sentence was finally thrown out after the United States Supreme Court found that the line was too rigid. The sentence of Jerome Bowden in Georgia in 1986 and the most recent case of Daryl Renard Atkins in Virginia in 2002 bear great similarity as well. One single point may matter, and so do

accurate and reliable measurements. How reliable were the psychological tests that these convicts took? Reliability arguments are the ones we use when we want to draw lines.

Reliability concerns, however, go beyond the tool used to test. There is a very important type of reliability which affects raters, particularly when it comes to productive skills. Hall, Bowden and Atkins results above might have been influenced by a too-strict (or simply wrong) interpretation of IQ tests. Since we want our tests to be reliable, we have to make sure that the professionals that administer them are reliable too, and this is when rater reliability (both inter-rater reliability and intra-rater reliability►) comes into play.

In a perfect model, our raters should apply consistently the same marking criteria to different exams so that different raters are able to measure the same candidates with no variation in the final results. In the case of productive skills, for example, we would expect rater 1 and rater 2 to apply the same rubric► to a given candidate with no variation in the results. However, we know that in real life this does not simply happen. Cognitively speaking, it is difficult for raters to hold all the definitions of their rubrics in their memory span at once (that is why they normally check the printed rubrics every now and then during assessment sessions).

Because of this, raters require mental representations of rubrics which are constantly matched with their interpretation of language during assessment sessions. The final mark of candidates is thus a function of this factor and others such as fatigue, prior information, environment conditions, previous performance, etc.). Kathrin Eberharter (personal communication) conceptualizes the process through the diagram below in her lectures at the University of Innsbruck:

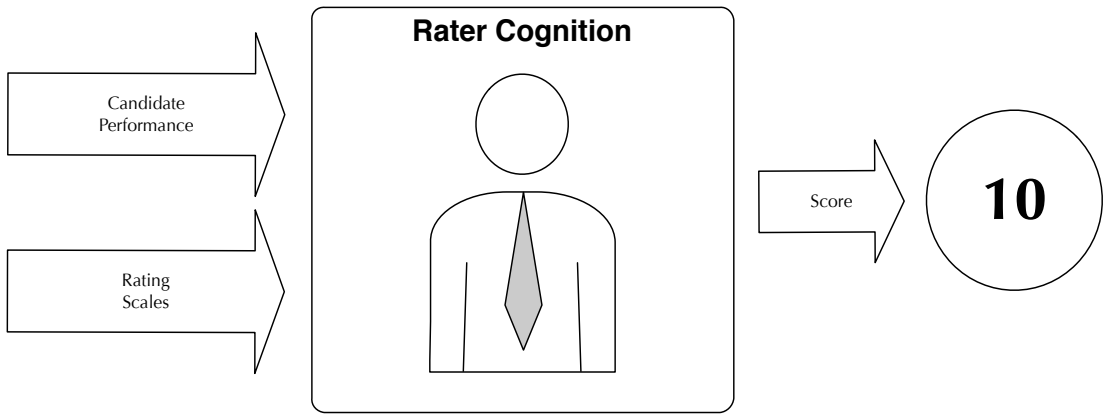

Figure 2.2.2. Conceptualization of the factors that affect rater cognition

In analyzing rater cognition, 2 aspects should be taken into account, a) the attributes of the raters (experience and expertise, native vs. expert, language background, L1, accent familiarity, knowledge of the rating scale, etc.) and b) the processes that they have to undergo (behaviors, general strategies, metacognitive strategies, etc.).

2.2.3 Fairness

If you judge a fish by its ability to climb a tree, or a duck by its ability to run, they will live their entire life believing that they are useless. Such is the moral drawn from *The Animal School*, the fable originally written in the 1940s by Reavis (1999) as a call to action against standardized tests during his days as superintendent of Cincinnati public schools in the United States. In the tale, a duck, a rabbit, a squirrel, an eagle and an eel have to take lessons on running, climbing, swimming and flying "to meet the problems of a new world". They all take all the subjects with, obviously, different results.

Testing fairness is not only about guaranteeing that all test takers are treated equally, but also about making sure that the results that texts yield are an index of candidates' abilities. In the fable, for example, the duck was not good at running but this does not mean that he is not good (and exceeds other competitors) at other activities. The fable of Reavis (1999) is not brought here to mean that tests must be more lenient with some candidates than with others, but

to exemplify that the fairness that standardized tests pursue is compromised if the picture is not seen as a whole.

At the University of Jaén, where we develop our professional activity, obtaining a B1 degree of proficiency in a second language is compulsory to finish any degree. In this context, we once had to consider the case of one student which we will call Lázaro. Lázaro suffered from infantile cerebral palsy, a brain disorder that affects muscle tone and motor skills and which makes muscle control and coordination difficult. This condition affected Lázaro's speech organs, which was particularly evident even during informal conversation in his mother tongue, Spanish. Lázaro was a hard working student, up to the point that he had been awarded some distinctions in his law degree. The day he decided to sit one of our tests we worked very hard to adapt the test to his special needs. We did not want to make the test easier for him because he had no cognitive impairment, we just wanted to make it accessible to know whether Lázaro had B1 competences in English. We prepared a special version of the exam for him, in which the contents were the same but in which we allowed extra time for him. We prepared a pdf version of the exam so that he could use a computer instead of paper, and so on. The test was suited for most other candidates, but not for Lázaro. Even after so much care was put into his version of the exam, we skipped one important thing. While his scores were similar to other test taker's in reading, writing and speaking (the last of which he faced with remarkable efforts), his scores at listening were surprisingly low. Where had the problem been? To answer this part of the test Lázaro had to write on a computer the correct answer to multiple choice items, on a Word document, through a mouse and a keyboard. Lázaro had to read the questions in a pdf file and to listen through headphones. Of course, he was allowed enough time to read the questions before the test started. Despite our efforts, we failed at providing a handy listening section that would not leave everything up to Lázaro's memory span: since he could not swift between the pdf (where he read) and the Word document (where he wrote) as quickly as a non-handicapped person, he tried to read, listen to the questions and answer them just

from what he remembered after listening to a 3 minutes recording. This task would have been challenging even for a non-handicapped candidate. Even though we failed to provide a fair test, Lázaro obtained his B1 certificate in English. Fairness here was not about making the test easier, but about making the test accessible. Standardized tests, as in the fable of Reavis (1999), may fail at eliciting a representative sample of candidates' performance if they are so rigid that cannot be adapted to special situations.

When designers prepare tests they have in mind an ideal test taker that somehow gathers a representative compilation of the many characteristics that all possible candidates might show. This is not always the case, as we learn from the case of Lázaro which, I am sure, is not unique.

Davies *et al*. (1999:199) point that

> Issues of fairness may arise at the test construction stage in relation to who has input into the test specifications and subsequently, which tasks or items are chosen as representative of the target domain. It might be considered unfair for example to include only multiple-choice items to test reading when it has been shown that boys generally do better than girls on such items.

The first part of this passage about specifications aligns with what Martínez (2011:59) claims when he says that it is useless to design a test without clarifying its purpose and the objectives of examinees. It would not be fair to administer a test in medical English to undergraduate students of law. Test designers that have "input into the test specifications" must consider this from the beginning.

As regards the second part of the excerpt, a test that contains only multiple choice questions might be unfair because of the reasons explained above or because it elicits only 1 type of cognitive behavior which does not represent all of the candidates' abilities.

Davies *et al*. (1999:199-200) go on putting forward the idea that

> [t]est development committees usually include representatives of relevant minority groups to ensure that the test content is sensitive to their interests and experience, and bias analyses are often undertaken to ensure that test items function uniformly across groups.

Washback►, fairness, bias, access equity in administration and several other ethic concerns (Kunnan, 2004) have been grouped under the aforementioned umbrella term of consequential validity. Consequential validity (Hawkey, 1999) is an overarching term about which Messick (1992:2) wrote back in the 1990s:

> With respect to consequences as validity evidence, I have argued for nearly 30 years that test validity and social values are intertwined and that evaluation of intended and unintended consequences of any testing is integral to the validation of test interpretation and use [...]. However, until the recent upsurge of renewed interest in performance assessment, there have been relatively few adherents to this position among measurement practitioners. Because they are now singing an old favorite song, the refrain of which intones that the consequences of measurement betoken its validity, I confess a certain fondness for performance assessors. But at the same time I am concerned that their enthusiastic embracing of the consequential basis of test validity might lead to a shortchanging of the evidential basis, including the need for evidence of the consequences.

Fortunately, Messik's fears did not come true and the "renewed interest" that he mentions has raised awareness on very important ethic aspects which very few would challenge nowadays. These aspects are present in virtually all codes of good practice (*Code Of Fair Testing Practices In Education*, 2004; EALTA, 2006; *ALTE Code of Practice*, 1994).

2.2.4 Practicality

As its name suggests, practicality frequently dictates the difference between how test designers would like to create their tests and how the have to do it. Practicality is the extent to which a test can get as close as possible to best practices in any possible aspect.

Modern test developers are subject go a great number (and frequently invisible) number of pressures. Quite frequently, "academics working in language testing in university contexts tend to be solitary figures; located as they are in applied linguistic programs for the most part, it is unlikely that any program will be large enough to have more than a single member or staff" (McNamara and Knoch, 2012:567). In this context, they have to toe the line on institutional policies and struggle to make the fairer test possible with limited budgets and pressing deadlines. The modern language tester, as Spolsky (1995:4) puts it,

> is expected to be responsible for and responsive to theories derived from two unrelated and fundamentally inharmonious fields, linguistics (which wants to describe language knowledge) and psychometrics (which hopes to measure it and other human attributes), and at the same time is directed and constrained by rival practical institutional, economic, social, and even political demands.

Take for example the test development cycle below proposed by Green and Spoettl (2011).

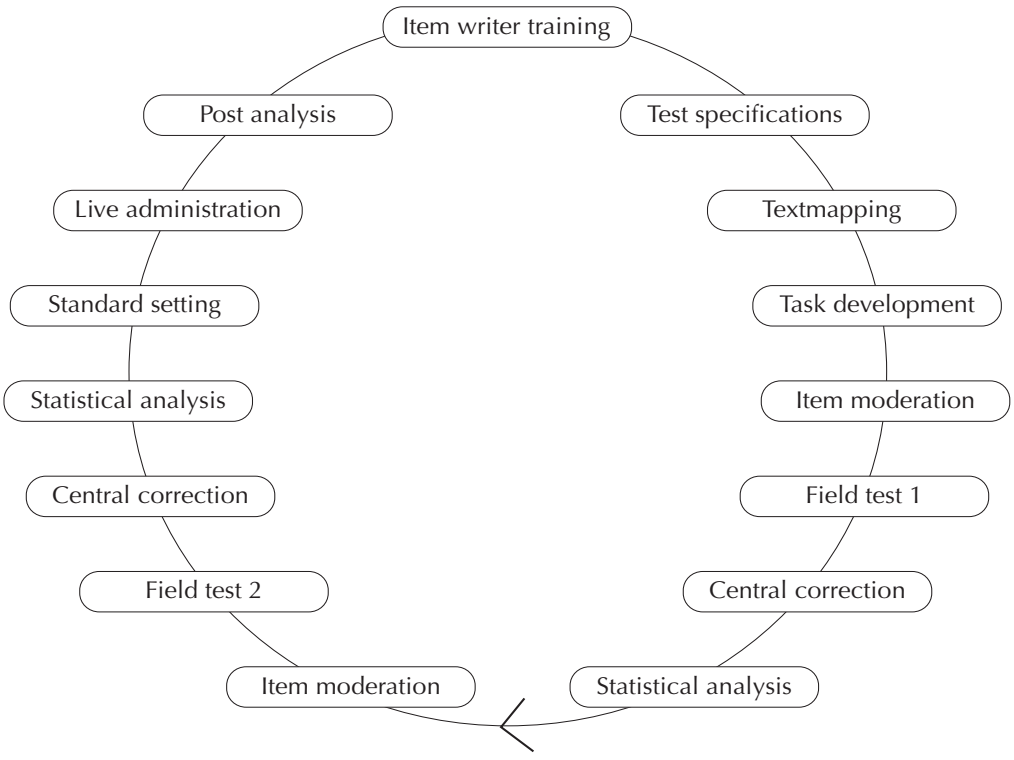

Figure 2.2.4. Ideal test development cycle by Green and Spoettl (2011)

This cycle starts with item writer training, moves clockwise towards test specifications, textmapping►, etc. and goes on across 15 different stages. Very few would challenge the idea that these stages are the minimal requirements for proper test design. Leaving aside founding problems that may hamper adequate item writer training, the first serious problem that test designers have to face is that of field test 1. Field tests require finding a group of candidates in which a preliminary version of the test, already revised by experts, is tested to check its reliability and to find out if the tasks and items are functioning as they are supposed to do. This entails agreements with fellow institutions which are willing to guarantee confidentiality of your tests and to allocate some time for your tests among their students. These are normally two-way agreements in which institutions test each other's items, and which require a lot of planning. Practicality here affects in so far as field tests are not always carried out with representative population samples, which may be not big enough or include members of all relevant minority groups.

When tests are field-tested these must be corrected centrally and its statistical properties must be analyzed. In small institutions it is difficult to find a full-time psychometrician (or even 1 at all) in charge of these tasks. It is frequent to find that test developers must specialize in unfamiliar areas such as statistical analysis at the same time that they juggle with other aspects of their jobs (lessons, administrative work, research, etc.).

Many items are frequently discarded after field test 1 and new ones must be designed to compensate for those which were binned. Then, the practical problem of finding representative population samples arises again at field test 2. Ideally, 2 or even 3 field tests should be carried out but practicality frequently leaves it in only 1.

Standard setting► also requires counting on a representative number of experts. After the tests have been tested for the second time and their statistical properties analyzed, a group of experts should meet to establish a cut score for

this particular version of the test. This procedure is costly and frequently substituted by a pre-set cut score for the sake of practicality.

Practicality in live administration entails using the minimal quantity of test invigilators, concentrating administration in the fewest possible number of sessions and balancing all this at the same time that the quality of the test is not put in jeopardy.

After the test goes live, sessions for productive skills must be arranged (not so in computer-based tests) and productive skills themselves should be ideally marked by at least 2 independent raters, which is not always possible for small institutions.

As we can see, there is a big (and frequently unnoticed) gap between good practices in test design and administration on the one hand and time and budgetary constraints on the other. Such is the daily work of test designers who are trusted with the responsibility of bridging the gap between the perfect test and real life.

2.3 Test methods

In 1935, Erwin Schrödinger, one of the fathers of quantum physics, envisaged the famous cat experiment that has become a staple of pop culture. Nowadays Schrödinger is famed not only for his cat experiment but also for his Nobel Price.

What not many people know is that Schrödinger (1935:812) described this experiment to criticize the Copenhagen interpretation of quantum mechanics. The Copenhagen indeterministic interpretation of quantum physics stated that a particle exists in all states at once and that it is only direct observation what makes that particle collapse into 1 of the multiple superposed states that it has. In other words, the action of the observer determines the state of particles.

In the above-mentioned thought experiment Schrödinger imagined that a cat, a small quantity of radioactive material, a Geiger counter connected to a hammer and a small quantity of poison were all put inside a box. The amount of

radioactive material was so small that it had a 50% probability of decaying within the hour. If the Geiger counter detected radiation it would trigger the hammer which, in turn, would smash the vessel containing poison, thus killing the cat. Given the fact that there is a 50% chance that the geiger triggers the hammer to smash the bottle containing poison, until someone opens and observes the system (or shakes the box!), it is impossible to predict if the cat was dead or alive. As a consequence, the cat is both dead and alive at the same time, in a zombie-like state. This paradox creates what Schrödinger himself called "ridiculous cases" (Schrödinger, 1935:812).

Obviously, this is not the place to discuss the principles of quantum mechanics. We just brought the experiment here to exemplify how important the observer may be in observations, something already mentioned in section 2.2.2 when we spoke about rater cognition. Observers are so important that they may radically change the results of the experiment. In the field of testing, there is another curious experiment which illustrates the importance of the observer more graphically.

During World War II, J. O. Roach, a Cambridge examiner (Weir, 2013:538-583) convinced some Polish military enrolled in the British Royal Air Force to take part in one testing experiment while Hitler's Germany was still bombing Europe. For the experiment he recruited in one military camp 22 Polish officers and men who had "to read a passage of English in turns before a panel of three or four examiners" (*ibid*.:542). After the experiment, Roach pointed that "the teachers and examiners agreed in retrospect that the candidates did not do themselves full justice before an audience of several inquisitors, particularly when some of these were their officers" (*ibid*.:543). In other words, the conclusion was that examinees assessed by officers who could decide about their position in the frontline experienced certain degrees of anxiety which led them to underscore in their tests. Roach called it one "abnormal feature of the test".

Both Schrödinger and Roach's examples illustrate common sense. The observer matters. The power position of examiners or "the frequent routine

variations in interlocutor behavior suggest that interlocutor cannot be considered a neutral factor in these assessments" (Lazaraton, 2001:59).

Speaking to and interacting with a person, a perfect stranger who is judging you, is a logical cause of anxiety. No one would expect to mark under such conditions without a possible margin of error in either direction. Developing assessment criteria that help standardize judgments while such and other "abnormal features of the tests" are compensated is among the earliest concerns of test designers. It is precisely here where rubrics are useful. Rubrics, most frequently used to assess productive skills, should be a set of descriptors able to compensate for all possible interferences between our candidates' performance and their final mark, thus enhancing scoring validity (Taylor and Galaczi, 2011). The objective of rubrics is to minimize the randomness in the behavior of the observer.

Section 2.3 is primarily intended to bring the discussion about test methods to the point in which rubrics come into play. For this we will first discuss different approaches to the assessment of productive skills in 2.3.1 and then we will define the very concept and history of rubrics in 2.3.2.

2.3.1 Direct, indirect, semi-direct, analytic and holistic methods

If one reading comprehension multiple-choice item is well designed, it can only be answered correctly or incorrectly. Tests may have open-ended questions, matching tasks or the like but, even so, if they are well designed and their key is clear, they are not difficult to mark. These items are frequently used to assess receptive skills. For productive skills, a different type of assessment is needed.

Things are far more complex when it comes to analyzing spoken performance. The performance of candidates in productive skills is not simply correct or incorrect and it frequently builds on multiple facets. There are other powerful forces involved in the final mark of candidates, like rater cognition (see section 2.2.2) or the analytic► scales, the rubrics which raters use to assess candidates. Rater cognition changes across raters and over time but can be

trained and standardized. Rubrics, on the other hand, offer a more stable standpoint because when they are finished they remain unchanged for very long periods of time.

But to define more exactly what rubrics are, we must first delve into the notions of direct and indirect assessment and second into the difference between holistic► and analytic approaches.

Clark (1979:36) was the first to discuss in 1975 a variety of techniques for measuring speaking ability and proposed the use of terms *direct* and *indirect* to distinguish 2 broad types of testing approaches. Four years later he included *semi-direct* methods in the classification. "*Direct* speaking tests were considered to include any and all procedures in which the examinee is asked to engage in face-to face communicative exchange with one or more human interlocutors" (*ibid*.). "*Indirect* speaking tests were considered to include both (1) those situations in which the examinee is not actually required to speak and (2) speech based on recorded or printed stimuli" (*ibid*.). Finally, the term *semi-direct* is used "to characterize those tests which, although eliciting active speech by the examinee, do so by means of tape recording, printed test booklets, or other 'nonhuman' elicitation procedures, rather than through face-to-face conversation with a live interlocutor" (*ibid*.). The type of tests that we are preparing our rubrics for is a hybrid of direct and semi-direct methods. It is direct in the sense that it requires candidates to engage in real, face-to-face communication in 2 of its 3 parts. They are semi-direct in so far as candidates are also given visual prompts to elicit from them a sustained monologue. Candidates always take the oral test in groups of 2 and, exceptionally, in groups of 3.

The tests whose outcome will be measured by our rubrics consist of 3 different parts and, basically, they bear the structure of a proficiency interview. In part 1 (direct) the rater asks introductory questions to candidates which are used to elicit personal information and to break the ice. In part 2 (semi-direct) candidates must describe pictures which normally present opposed views of the same matter (reading through digital books vs. reading through traditional books;

studying alone vs. studying in group, etc.). In part 3 (direct) candidates are also given visual prompts but, this time, with the objective of engaging them in conversation.

The proficiency interview method for testing oral performance enjoys a very high degree of face validity but also has its downside. This "technique does not perfectly reflect real-life conversational settings", it has difficulty in "eliciting certain fairly common language patterns typical of real-life conversation" and offers candidates little room "to demonstrate productive control of interrogative patterns unless the interviewer takes special pains to 'turn the conversation around' at one or more points during the interview" (Clark, 1979:38). Yet, "of the currently available testing procedures, the face-to-face interview appears to possess the greatest degree of validity as a measure of global speaking proficiency and is clearly superior in this regard to both the indirect (non-speaking) and semi-direct approaches" (*ibid*.).

As stated in the introduction of the present section,

> [f]or both direct and semi-direct speaking tests, the reliability question is somewhat more complicated in that examinee performance must be evaluated by human judges rather than through such mechanical means as answer key stencils or computer scoring devices. Two distinct types of reliability enter the picture here: intra rater reliability, which refers to the extent to which a given scorer is able to consistently assign the same scores to individual tests that he or she evaluates two or more times in succession; and inter-rater reliability, which refers to the extent to which two or more different raters assign the same scores to a given test performance.
>
> Clark (1979:41)

Modern advances in testing have made analyses beyond inter-rater and intra-rater reliability possible. Both types of analysis assume that judgments are made through one well-calibrated tool and that raters are the only factor that can introduce randomness in the observations made. In other words, they assume that the criteria that these raters use are adequate and that raters are the only possible

source of variability. Nowadays we know that it is not so. There may be a great degree of inter-rater unreliability if the assessment criteria used are not easy to understand or if they leave too much room for interpretation. Many raters will recognize themselves in the words of Knoch (2009:12) below, which we already quoted in the introduction:

> I often found that the descriptors provided me with very little guidance. On what basis was I meant to, for example, decide that a student uses cohesive devices 'appropriately' rather than '*adequately*' or that the style of a writing script '*is not appropriate to the task*' rather than displaying '*no apparent understanding of style*'? […] This lack of guidance by the rating scale often forced me to return to a more holistic form of marking where the choice of the different analytic categories was mostly informed by my first impression of a writing script […]. I often felt that this was not a legitimate way to rate and that important information might be lost in this process.

Although Knoch refers to writing rubrics, the same goes for speaking ones. Modern psychometric analyses have proved long-time-used rubrics to be faulty (Jansen *et al.*, 2015), and that is the reason why rubrics must also be (re)considered as an object of psychometric analysis.

Besides the consideration of direct, indirect and semi-direct methods and their reliability implications, in the mind of test designers there is also a never-ending good-evil struggle that confronts holistic versus analytical methods. With holistic approaches we think we know what we are assessing, but remain happily or unhappily uncertain about the accuracy or replicability of our assessment. With analytic approaches we tend to be sure enough of our measurement, but we may jeopardize our certainty as to what exactly we have measured.

The term *holistic* derives from the Greek word ὅλος (pronounced /ˈɔlɔs/) which means "whole". Holistic scoring thus requires raters to respond to oral performance as a whole and to base their score on a general impression of the candidate. This approach reflects the idea that oral performance is a single entity,

which is best captured by a single score that integrates the inherent qualities of writing (cf. Knoch, 2009:39). Davies *et al*. (1999:75) define holistic scoring as

> [a] type of marking procedure which is common in communicative language testing whereby raters judge a stretch of discourse (spoken or written) impressionistically according to its overall properties rather than providing separate scores for particular features of the language produced (eg (*sic*) accuracy, lexical range) […]. A problem with holistic judgements, however, is that different raters may choose to focus on different aspects of the performance, leading potentially to poor reliability if only one rater is used. For the sake of reliability, therefore, test performance is normally judged by several raters and their judgements pooled. A further drawback of holistic scoring is that it does not allow detailed diagnostic information to be reported.

While Jonsson and Svingby (2007:131-132) describe it as follows:

> Two main categories of rubrics may be distinguished: holistic and analytical. In holistic scoring, the rater makes an overall judgement about the quality of performance, while in analytic scoring, the rater assigns a score to each of the dimensions being assessed in the task. Holistic scoring is usually used for large-scale assessment because it is assumed to be easy, cheap and accurate. Analytical scoring is useful in the classroom since the results can help teachers and students identify students' strengths and learning needs. Furthermore, rubrics can be classified as task specific or generic.

In fact, it was the apparent lack of reliability mentioned above by Davies *et al*. (1999:75) what triggered the design of our rubrics. In our opinion, holistic scoring offers certain benefits in achievement or placement tests but, as a sole source of reference in proficiency tests, they leave too much room for individual interpretation.

As will be shown in chapter 4, our rubrics, we felt, had to be analytic and as accurate as possible, well-balanced, cognitive-friendly (so that they did not require too much working memory from raters) and leave little room for interpretation. This is easier said than done, but it was clear from the very

beginning that our rubrics had to be analytic. Knoch (2009:40) defines analytic scoring applied to the assessment of writing as follows:

> A common alternative to holistic scoring is *analytic scoring*. Analytic scoring makes use of separate scales, each assessing a different aspect of writing, for example vocabulary, content, grammar and organization. Sometimes scores are averaged so that the final score is more usable […]. A clear advantage of analytic scoring is that it protects raters from collapsing categories together as they have to assign separate scores for each category. Analytic scales help in the training of raters and in their standardization […] and are also more useful for ESL learners, as they often show a marked or uneven profile which a holistic rating scale cannot capture accurately.

Adapting the categories assessed, the same can be said about analytic scoring of oral performance. Generally speaking, trained raters feel more confident using analytic scales than using holistic ones. There is the underlying belief that "(j)ust as discrete-point test becomes more reliable when more items are added, a rating scale with multiple categories improves the reliability" Knoch (2009:40).

We did not consider the possibility of using primary trait scales► because this type of rubric is defined with respect to the specific task to be judged and to the degree of success in it (Weigle, 2000:110). In other words, primary trait scales are so task-specific that it would have been impossible to create a suitable rubric with such design for the context described in sections 3.3 and 4.1.

2.3.2 Rubrics: history and definition

In the previous section we have suggested that different methods should be used to assess different skills. We have mentioned the concept of rubric several times and have narrowed the scope of study to indicate that we have chosen rubrics as the method to assess oral performance in our proficiency tests. It is now the time to define more precisely what rubrics are, because this is precisely the tool that we want to create and validate for our experiment in chapter 4.

In medieval illuminated manuscripts, rubrics were indications written or printed in red for emphasis or for instructions in liturgical services. The term derives precisely from the Latin word *rubric*, which means red ochre, the color these indications were written with. Other terms like multi-trait rubric, analytic scale or scoring rubric bear the same meaning as rubric or a slightly different one depending on the author consulted. For some authors, for example, analytic scales are a more numeric tool, a naked version of rubrics that lend themselves to mathematical manipulation as opposed to rubrics, which are in turn considered to be a more descriptive way of scoring. Due to its etymology, the word "rubric" is also used to refer to the instructions given to candidates on how to complete a task or a test. Here we will refer to the word "rubric" as a synonym of "scoring rubric" or "analytic scale", that is to say, the set of descriptors against which raters will compare the performance of candidates in an oral test.

The use of rubrics is not new. The first analytic scale that we know of can be attributed to George Fisher. Fisher was born in Sunbury, England in 1794. Following the death of his father, he had to go out to work at an early age. His interest in science, however, won him recognition and in 1817 he was able to study at Cambridge University. Over the coming years, Fisher alternated his studies with different naval expeditions in which he sailed towards the Artic as astronomer in search of the North-West Passage. Back in England, in 1834, he got married and also accepted the Headmastership of the Royal Hospital School which, by that time, was located in Greenwich. In this school of naval tradition he had to educate the sons of many sailors who had lived most of their lives aboard ships sailing with their parents. The excerpt below (Chadwick, 1864:481) describes the labor of Fisher in the aforementioned institution. The passage reproduces part of a conversation about his students, held with the English social reformer Edwin Chadwick. The opening question is posed by Chadwick, the author of the article, and the answer is Fisher's:

> Of what class are they? –They may perhaps be best described in the words of Mr. Cannon Mosely, who reported on them, that the great majority of them are the sons of sailors; that they have not unfrequently passed their previous lives amongst the lowest haunts of a seafaring population and they come to the institution "at an age (about eleven) when the influence of evil example has already begun to acquire some hold upon them, and the power of evil habits has begun to be felt".

The text continues for some more pages in which Fisher describes his concern about recording the performances of his students and similar. It is the appendix of the article (Chadwick, 1864:484) that refers explicitly to the first analytic scale known in the history of education. Below we reproduce the first 2 paragraphs of the aforementioned appendix, in which all the original punctuation marks have been preserved:

> We quote, from a letter addressed to Mr. Chadwick by the Rev. George Fisher, the following description of the method of collecting educational statistics in use in the Greenwich Hospital School:
>
> "A book called the 'Scale-Book,' has been established, which contains the numbers assigned to each degree of proficiency in the various subjects of examination: for instance, if it be required to determine the numerical equivalent corresponding to any specimen of 'writing,' a comparison is made with various standard specimens, which are arranged in this book in order of merit; the highest being represented by the number 1, and the lowest by 5, and the intermediate values by affixing to these numbers the fractions ¼, ½, or ¾. So long as these standard specimens are preserved in the institution, so long will constant numerical values for proficiency in 'writing' be maintained. And since facsimiles can be multiplied without limit, the same principle might be generally adopted."[2]

After some paragraphs describing how the method is applied to other disciplines (mathematics, navigation, Scripture knowledge, grammar,

[2] Quotation marks in this excerpt and in the next one are reproduced as in the original.

composition, French, general history, drawing and practical science), Chadwick finishes the appendix quoting Fisher's words (*ibid*.) again:

> "Having stated thus much with regard to the plan pursued in this school, I may add, that the advantage derived from this numerical mode of valuation, as applied to educational subjects, is not confined to its being *a concise* method of registration, combined with a useful approximation to a *fixed standard* of estimation, applicable to each boy; but it affords also the means of determining the *sum total*, and therefrom the mean or average condition or value, of any given number of results."

Deygers and Van Gorp (2015:522) attribute mistakenly the first analytic scale to Thorndike (1910), who simply reproduces the same appendix that we have quoted above.

Chadwick's passage with Fisher's words contains some interesting concerns about what a rubric must be which remain present in modern definitions. Such concerns include the categorization of performances in numeric scales, the replicability of results or the "specimens" against which performances can be measured, which in Fisher's system were sample writings.

By the end of the 19th century there were already staticians such as Edgeworth (1888) who were concerned with enhancing the reliability of tests through "numerical marks". Edgeworth (1888:600) wrote:

> That examination is a very rough, yet not wholly inefficient, test of merit is generally admitted. But I do not know that anyone has attempted to appreciate with any approach to precision the degree of accuracy or inaccuracy which is to be ascribed to the modern method of estimating proficiency by means of numerical marks.

Some years later Cattell (1905:367) noticed the same problem when he wrote the following:

> In examinations and grades we attempt to determine individual differences and to select individuals for special purposes. It seems strange that no scientific study of

> any consequence has been made to determine the validity of our methods, to standardize and improve them.

However, neither Edgeworth (1888) nor Cattell (1905) refer to Fisher (1864). Fulcher (2015) provides a comprehensive account of the way in which rubrics evolved along the 20th century thanks to the work of Thorndlike, Yerkes, Kaulfers, Roach, Carroll, Adams, Bachman, Lantolf, Kramsch, Alderson or Linacre among other scholars.

After all such evolution, in a broad sense, rubrics are now agreed to be a set of competence descriptors embedded in an interval scale (Bachman, 1995:28) with different levels equidistant from each other. The description of these levels must be specific and accurate enough to elicit similar responses from different assessors when confronted with the same test sample. A good set of rubrics is one of the best tools for the assessor. There are also more specific descriptions of rubrics (Brindley, 1998:112, Bachman, 1995:35-37, Richards and Schmidt, 2002:471; Dean, 2012:1), out of which the one provided by Davies *et al*. (1999:153-4) is perhaps the most comprehensive:

> 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 characterized 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). Proficiency scales typically consist of sub-scales for the skills of speaking, reading, writing and listening […]. 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 test population and test purpose. Raters or judges are normally trained in the use of proficiency scales as to ensure the measure's reliability.

This definition is comprehensive and introduces important aspects of what rubrics should be, namely the concept of bands, the concept of the linguistic features to be analyzed, the descriptors associated to behaviors and the fact that one particular set of rubrics serves one particular construct. Although there are other arrangements (Dean, 2012:1-9), in figure 2.3 below we show the most popular outlay of rubrics, with its main parts labeled (cf. Appendix):

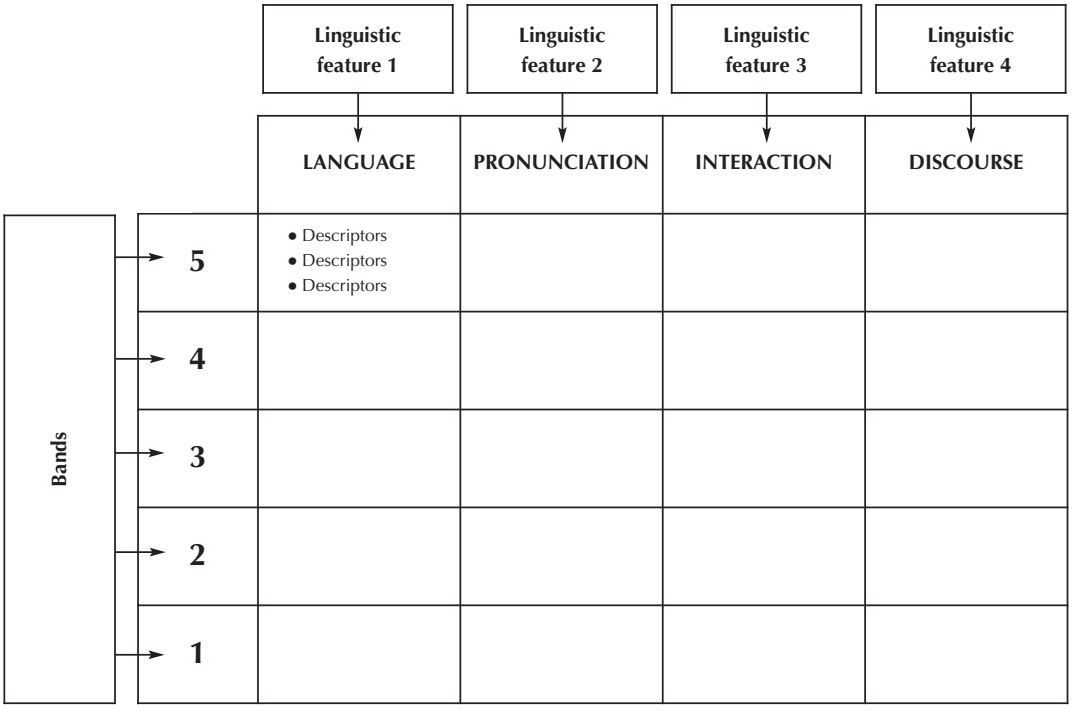

Figure 2.3.1. Typical arrangement of a set of rubrics

Along with this, Davies *et al.*'s (1999:153-4) definition above arguably identifies high levels of proficiency with native-like speech (an idea also present in Brindley, 1998:113) or focuses on proficiency rubrics leaving aside other typologies like diagnostic rubrics. It also refers to rater training and to reliability, but the latter appears in the description as based on inter and intra-rater concerns and does not mention the internal consistency of the rubrics, the progression of band descriptors or how the rubric itself must be empirically validated or linked to standards.

Davies *et al.*'s definition also lacks any reference to rubrics development. The question about the design of the descriptors and the linguistic features of the

rubric is also relevant to the perception that raters have of the rubrics since they will have to internalize and incorporate these descriptors to their own construct of language. In general terms, the more raters know about how rubrics are developed and the rationale that supports them, the better. When rubrics become a matter of opinion, raters frequently wonder if the rubric in question was developed by a particular body of experts, whose opinion was sought or if the rubric was validated after being developed. Answering such questions satisfactorily is the only way of mitigating any feeling of uncertainty, the "lack of guidance by the rating scale" mentioned by Knoch (2009:12). Otherwise, rubrics, which are intended to generate consensus, may become the source of bitter disagreement:

> [P]erformance assessment necessarily involves subjective judgements. This is appropriate: evaluation of any complex human performance can hardly be done automatically. Judgements that are worthwhile will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement.
>
> McNamara (1996:117)

In chapter 4 we will present a protocol which has proved to be successful to design and validate rubrics in collaboration with prospective users of the scoring rubrics. This protocol also offers a straightforward way of aligning the bands of the rubrics to the *CEFR* (Council of Europe, 2001) levels.

2.4 Psychometrics

Psychometrics has been frequently mentioned in the previous dissertation as a very important part of test development and which will be central for the present dissertation from this point. Unfortunately, for linguists and test developers within the European tradition psychometrics is still a "harsh" subject due to its mathematical component. Roughly speaking, psychometrics is a tool that allows test developers to prove through mathematics what claims about their tests are

valid and which ones are not. Psychometrics is defined by Davies *et al*. (1999:157) as:

> The measurement of psychological traits such as intelligence or language ability. In addition to deciding about item types and test content, the test developer needs to consider the psychometric or measurement properties of the test items, such as the level and range of item difficulty and discrimination. This information will typically be gathered during the trialling stage of test development, and decisions about the desired psychometric qualities of items will depend on the intended use of the test and interpretation of test scores.
>
> Psychometrics provides a framework for the development and evaluation of tests. It is based on the assumptions of normal distribution and of maximising the distinction between candidates. Psychometric tests have properties such as objective scoring, and are evaluated according to an established set of methods. The multiple-choice test is probably the best known of such tests, although all norm-referenced tests are generally based on psychometric principles.

The idea of measuring psychological traits is thrilling for language lecturers and test developers. For the former ones, psychometrics can provide numerical evidence of the progress of their students. For the latter, it can help to build more reliable, fairer tests. Just to name some examples, psychometric procedures can shed light on how difficult an item is (to know whether it is suited for the level it was originally designed) and on how well it discriminates between students who are in the targeted level and those who are not.

This is perhaps the first major step into the creation of operative tools for language teaching and test development. Surprisingly enough, psychometrics and its applications are unknown to a great number of lecturers on linguistics and test developers in Spain. When they approach this branch of psychology for the first time they are mystified by the fact that it has been there for more than 100 years. The psychometric models most extensively used in testing contexts, however, are not so old. Take for example the case of Rasch models (McNamara and Knoch, 2012). When professionals in the field of testing first encounter psychometrics

they feel as if they have been handed a very powerful microscope to examine the complexity of the rating process (*ibid*.:567).

Then, if it is so useful and it has been there for so long, why is it so unknown for language tests developers? The answer is simple. It is not an intuitive science *a priori* and requires certain knowledge and assumptions which are not always germane to language syllabi at universities, particularly in the European tradition. As McNamara and Knoch (2012:557) put it

> Language testing is a hybrid filed, with roots in applied linguistics and in measurement. Researchers (at least in the English-speaking world) frequently enter work in language testing following initial training careers in language teaching rather than in statistics or psychometrics. Their introduction to language testing is in specialist courses within graduate study in applied linguistics, or through practical exposure in their professional teaching careers, and they are likely to lack a strong background in mathematics or statistics in their prior education, typically being languages, linguistics or humanities and social science majors in their undergraduate degrees. They may even have consciously avoided and feel uncomfortable with numbers.

Quite so, many language lecturers and language test developers "consciously avoided and feel uncomfortable with numbers". Fortunately, this is changing in Europe due to the evident relevance of the analyses that psychometrics makes possible. Psychometrics can, for example, tell us whether one test is reliable or not through a very simple statistic analysis (Krombach's α is perhaps the most popular of such analyses). Psychometrics can even point at those items within one test which are less useful than others to discriminate among students or can measure rater leniency against different criteria along various test takers. Some of these analyses are more valuable than others, and some are more complex to carry out than others too. In the present dissertation we will use psychometrics not for the design of a test but for the validation of a set of rubrics. We expect to prove through numbers that the view of language that we reflect in our descriptors of performance is valid, reliable and consistent.

As we mentioned in section 1.1 when we defined our construct of language, psychometrics became a core part of language testing in the 20th century and, most precisely, in its second half. In the historical account of psychometrics applied to language testing, 2 traditional currents are frequently distinguished, Classical Test Theory (CTT►) and Modern Test Theory (MTT►). Green (2013:xii) puts it as follows:

> There are two broad ways of analysing test data: one uses what is referred to as the classical test theory (CTT) approach, and the other the modern test theory (MTT, also referred to as IRT) approach. Both have their advantages and disadvantages […]. In the field of language testing, CTT involves analysing test data in order to investigate such aspects as item difficulty, levels of discrimination, the contribution each item or part of a test makes to the test's internal reliability, the relationship between various parts of a test or tests, the relationship between test taker characteristics and their performance on a test, to name but a few […]. IRT is based on probability theory: the chances of a person answering an item correctly is a function of his/her ability and the item's difficulty (Henning 1987). In other words, a test taker with more ability has a better chance of answering an item correctly; similarly, an easy item is likely to be answered correctly by more people than a difficult one.

Thus the classical-modern distinction is not simply a chronological one. The inferences that can be made from classical and modern test theories are different as well. MTT, for example, is generally deemed as more robust than CTT and MTT is considered capable of yielding richer data.

In the following sections we will break down the main differences between both approaches in order to reach the goal of this section, which is to define the multi-faceted Rasch model, the psychometric model that we have used to validate our rubrics. For such purpose, we will use the following structure:

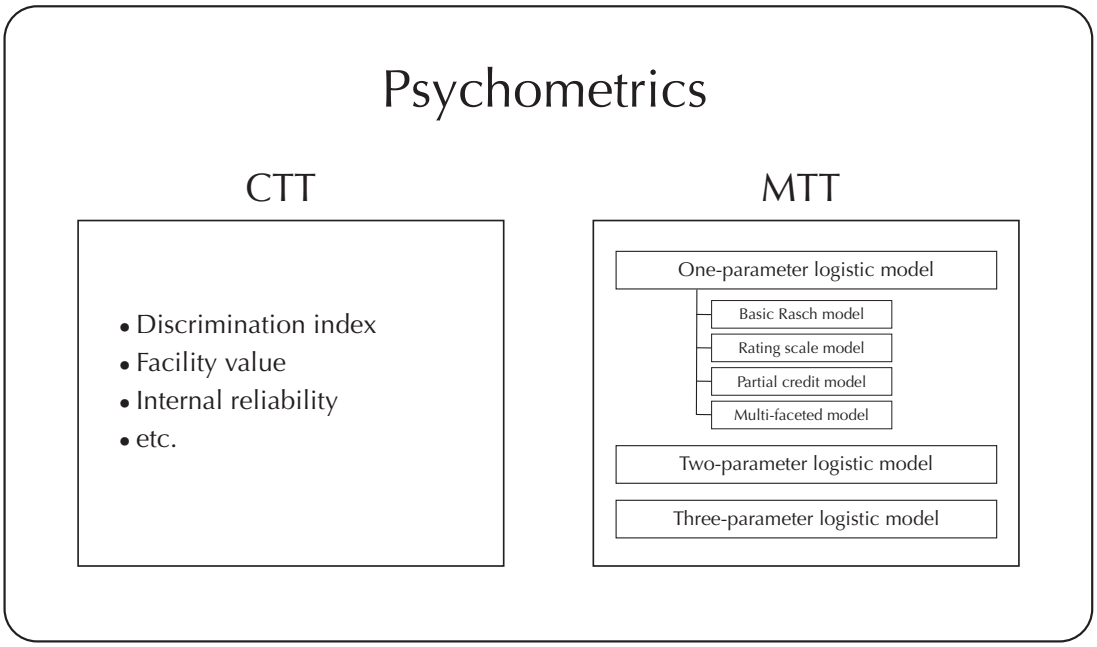

Figure 2.4. Psychometrics: CTT vs. MTT

2.4.1 CTT and sample dependency

Linden and Hambleton (1997:2) define CTT as a theory that

> [s]tarts from the assumption that systematic effects between responses of examinees are due only to variation in the ability (i.e. true score) of interest. All other potential sources of variation existing in the testing materials, external conditions, or internal to the examinees assumed either to be constant through rigorous standardization or to have an effect that is nonsystematic or "random by nature".

In a broad sense, the main distinction between CTT and MTT is that CTT statistics such as item difficulty (*i.e.* proportion correct), item discrimination (*i.e.* point biserial correlations) and internal reliability (*i.e.* the degree to which one particular test would yield identical results if applied to the same candidates in repeated iterations) are sample dependent (Hambleton and Jones, 1993:38; McNamara, 1996:151-152), while MTT statistics corrects for such dependency (Linden and Hambleton, 1997:2) through the use of probability models applied to data matrixes.

Sample dependency in CTT is not considered as a positive attribute because it is as much as saying that the psychometric characteristics of one item depend on the population it was tested on. If the inferences about the reliability of our results depend on the test-takers used during the trials, we will never be completely sure about its properties because these will change if the item is trialed in a different group of candidates. Such is the reason why CTT inferences are sometimes criticized. Let us see one example of what sample dependency means through the analysis of what happens in discrimination indexes when candidates vary.

Roughly speaking, the discrimination index of one item is a figure that tells us how well this particular item differentiates stronger from weaker performers. If the item in question is answered correctly by those that we identify as stronger candidates (*i.e.* those with higher overall scores) but it is not answered correctly by weaker candidates (*i.e.* those with lower overall scores), then we can say that this particular item is helping us to discriminate between stronger and weaker candidates because it predicts accurately higher scorers. However, there are times in which one "difficult" item is answered correctly by weaker candidates. The same goes the other way around and sometimes stronger candidates give wrong answers to "easy" items.

To calculate discrimination indexes we are going to use an imaginary item, Item A, which was answered by 300 candidates (n). To calculate the discrimination index of Item A we need 2 data from these 300 candidates, 1) the facility value, and 2) the Pemberton index (Martínez, 2011:68). By comparing the difficulty value and the result of the Pemberton formula we will know the discrimination index of the item, *i.e.* if it discriminates well strong and weak candidates. Then we will analyze what happens if both data are obtained from a different sample of candidates to exemplify sample dependency. The whole calculation is summarized in table 2.4.1.a below.

First, let us start by calculating the facility value if Item A. Since 75% of candidates answered Item A correctly (*i.e.* 225 candidates), it is said to have a facility value of 75%.

Second, let us calculate the Pemberton index. To apply the Pemberton formula (see the formula below) we are going to divide the 300 candidates who answered Item A into 3 groups of equal size (100 candidates each). In these 3 groups we will include candidates according to their overall score. The 100 candidates with the best overall marks in the exam will go to group 1 (G1), the 100 candidates with the lowest overall score will go to group 3 (G3) and group 2 (G2) will contain candidates with intermediate overall scores. The numbers on the right of the groups in the table below (100, 95 and 30) indicate how many candidates answered Item A correctly in each group. Thus, from the table we learn that 100 candidates out of the 100 candidates of G1 answered item A correctly, that 95 candidates out of the 100 candidates of G2 answered item A correctly and that only 30 candidates out of the 100 candidates of G3 answered item A correctly.

Item A (n = 300 samples)	**Correct overall answers (%)**
Facility value of Item A	225 of 300 = 75%
Discrimination index	
G1: 100 strongest candidates	100
G2: 100 intermediate candidates	95
G3: 100 weakest candidates	30

Table 2.4.1.a. Data for discrimination index through the Pemberton formula

The Pemberton formula being (Martínez, 2011:68):

$$\frac{G1 - G3}{n/3}$$

by substitution we find that:

$$\frac{100-30}{100}=\frac{70}{100}=0.7$$

Now, the facility value of Item A (75%) and the result from the Pemberton formula (0.7) must be correlated. A correlation is expected between the theoretical difficulty of Item A (75%) and its discrimination index (0.7). The more difficult an item is, the closer its facility value approaches 0% (because 0% candidates will be able to answer it correctly). The easier an item is, the closer its facility value approaches 100% (because 100% candidates will be able to answer it correctly provided it is very easy). This correlation is analyzed through the table below, which is adapted from Martínez (2011:68).

Facility value	Pemberton index
100	0.0
96	0.1
93	0.2
90	0.3
86	0.4
83	0.5
80	0.6
76	0.7
73	0.8
70	0.9
66	1.0
50	1.0

Table 2.4.1.b. Expected correlation between facility value and Pemberton index

As we can see in the table, a Pemberton index of 0.7 is expected for an item with facility values of 76%. Since the facility value of our item is 75% (very close to 76%), we can say that Item A is a good item because it displays the expected ratio between facility values and the Pemberton index.

The problem with this statistic (and the source of most criticism directed at CTT) is that the results are sample-dependent. This means that if, for example, the exam is taken by not-very-motivated students, the outcome and the conclusions drawn will vary as well. Not all groups of candidates will necessarily show the same degree of regularity in their answers, and this will also affect the final calculations.

To exemplify this inconsistency, let us imagine now that the same item is trialed in another group of 300 candidates who are slightly less motivated. In this case, only 65% of them (195 candidates) answer it correctly. In G1, 90 candidates give the correct answer to Item A, 65 in G2 and 30 in G3. By applying the formula we obtain a Pemberton index of 0.6 which, as we see in table 2.4.1.b, is not a good result because a Pemberton index close to 1 is expected for an item with a difficulty value of 65%.

The item remained the same in the first and the second calculation, but the sampled candidates changed and so did the statistic properties of the item. This is what is meant by sample dependency in CTT. Sample dependency affects the conclusions drawn from items in so far as the same item sampled in different candidates yields different results. Rita Green (personal communication) claims that reliability indexes may vary up to 14% depending on whether the exam is being taken by real candidates or by candidates who are trialing it.

Bachman (1991:203) also criticizes that CTT "does not provide a very satisfactory basis for predicting how a given individual will perform on a given item", chiefly because "it makes no assumptions about how an individual's level of ability affects the way he performs on a test" and because "the only information that is available for predicting an individual's performance on a given item is the

index of difficulty", that is to say "the proportion of individuals in a group that responded correctly to the item".

Set against this, MTT, as we will see in 2.4.2 and 2.4.3, uses mathematical models that balance the possible dependency and that provide accurate predictions of individual candidates on different items. Bachman (1991:203) points out the following:

> These models are based on the fundamental theorem that an individual's expected performance on a particular test question, or item, is a function of both the level of difficulty of the item and the individual's level of ability.

2.4.2 MTT and statistical models

MTT is also frequently referred to as Item Response Theory or Latent Trait Theory. However, for the present dissertation, we will use the MTT (modern) label since in mnemotechnic terms it is easier to remember the CTT (classical) vs. MTT (modern) distinction, as already stated in the introduction to section 2.4.

Defining what MTT is "sometimes verges on the nonsensical, and certainly on the irascible, because protagonists are using the term in very different senses" (Linacre, 2003:926). A precise definition of MTT draws on quite complex statistical models and obviously this is not the place to discuss such matters. However, most of the definitions of MTT (but not all) agree on the fact that it relates the probability of an examinee's response to a test item to an underlying ability (Linden and Hambleton, 1997:v; Green, 2013:xii) and on the fact that "it encompasses any mathematical model which attempts to predict observations on a latent variable" (Linacre, 2003:926) (hence its alternative name of "Latent Trait Theory"). The 2 assumptions above do not always go together but help us to understand that MTT has an eminently predictive nature. In other words, MTT can help us to predict how our language tests (or rubrics) will behave departing from a reduced data set. MTT tries to identify patterns in data which researchers or test designers can use to draw conclusions, even if such data sets are reduced in size, which is another advantage. As McNamara (1996:133) puts it "useful estimates

can still be derived from a 'holey' data matrix, although the more information available to the analysis the better these estimates will be".

Once the basic difference between CTT and MTT is stated (item-dependency vs. estimation of probability), it is necessary delve into MTT, which is, by far, more complex than CTT.

What characterizes MTT internally is the mathematical model used to estimate the aforementioned probabilities, the models used to find patterns in data sets. McNamara (1996:257-258) entangles the origin of MTT and the rise of its main mathematical currents as follows:

> Item Response Theory (IRT) is a powerful general measurement theory which was developed in the 1950s and 1960s independently, it seems, in two different locations: by Alan Birnbaum in the United States and by the Danish mathematician Georg Rasch in Denmark. Rasch's work was promoted and extended by an American, Ben Wright, who attended a series of invitational lectures given by Rasch in Chicago in 1960 and became his pupil and the advocate of his ideas in North America […]. Two main branches of Item Response Theory (or Latent Trait Theory as it is sometimes still known), stemming from these two developmental traditions, are recognized […]. They differ theoretically and practically. The essential feature of both is that they attempt to model statistically patterns in data from performances by candidates on test items, in order to draw conclusions about the underlying difficulty of items and the underlying ability of candidates. They differ mainly in the number of item parameters (characteristics of the interaction between a test taker and a test item) being estimated in the analysis: Rasch analysis considers one item parameter (item difficulty), while other models consider one or more further parameters (item discrimination, and a guessing factor).

Generally speaking, it is the number of parameters considered what establishes the current different forms of MTT. This way we find the one-parameter logistic model (also known as Rasch), the two-parameter logistic model and the three-parameter logistic model. For a deeper mathematical analysis of these models see Linden and Hambleton (1997), McNamara (1996) and Harris

(1989) and for a close up of their historical evolution see McNamara and Knoch (2012). We will be using the one-parameter logistic model in our analysis of rubrics although we will also use 1 example of three-parameter logistic model to provide the big picture of MTT.

In general, there is a series of reasons why researchers opt for MTT. Reise *et al*. (2005:100) claim that MTT methods are used because:

> [R]esearchers want to (a) more rigorously study how items function differently in different groups; (b) place individuals from different groups onto a common scale, even if they have responded to different items; (c) use individual scores that have good psychometric properties, so that statistical techniques (such as growth model) can be applied with greater accuracy and spurious results or invalid findings can be avoided; (d) thoroughly understand the psychometric properties of their instruments; (e) create more order in their fields by having a common metric for a construct, rather than many competing fixed-length instruments; and (f) develop CAT (*computerized adaptive testing*) [3] systems for more efficient assessment of individual differences.

Most scholars agree on the fact that that MTT has 4 intrinsic properties, namely sufficiency, separability, specific objectivity and latent additivity. Among these, the most interesting one is specific objectivity, which matches with reason (b) in the excerpt above. The property of specific objectivity in MTT allows, for example, for the comparison of persons without reference to the particular items taken and comparison of items without reference to the particular persons providing the responses, which compensates CTT's sample-dependency referred to in 2.4.1. Thus MTT models "place individuals from different groups onto a common scale, even if they have responded to different items", as Reise *et al*. (2005:100) pointed. In practical terms this means that if we have collated data properly and if these fit the parameters, MTT models create a suited scale in which all the measurements will be distributed accurately. This scale is our logits► scale which is explained below.

[3] The brackets and their content did not appear in the original passage.

The implications of specific objectivity of MTT models is particularly relevant in one study like ours, in which we are trying to validate a measuring instrument. As Thurstone (1928:547) puts it, a scale must transcend the group measured:

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement.

As a consequence of the underlying property of specific objectivity, if we can prove that our rubrics fit MTT models, by analyzing the data of a reduced number of trialed candidates, we will be able to ascertain whether our rubrics are valid to rate any prospective test-taker. The corresponding fitting analyses are shown in section 4.2.4.

The one-parameter model MTT is the model that we will be using for validation purposes, but there are others. The existing mathematical models can be basically broken down into the one-parameter logistic model (or Rasch model), the two-parameter logistic model and the three-parameter logistic model (Harris, 1989:35). Although we will be using one-parameter logistic models in the validation of our rubrics, we will define first what makes the one-parameter different from the two-parameter and the three-parameter logistic models to provide a general view of this type of statistics.

When establishing the degree of probability of a test taker answering correctly one item, each of the 3 models considers, as their name suggest, a different number of parameters. Parameters can be defined as the characteristics of an item which, according to the model being used, may or may not be taken into account. The 3 characteristics that models may consider are item difficulty,

item discriminability and the effect of guessing (Davies *et al*. 2006:140). As it is obvious from their name, the one-parameter model takes into account 1 parameter, the two-parameter model takes into account 2 and the three-parameter 3. These parameters are taken into account by the mathematical models that define item characteristic curves (ICC)►, which are the cornerstone of MTT models (Bachman, 1991:203).

ICC and many other statistics coming from MTT analyses are measured in logits (Green, 2013:151), as we can see in figure 2.4.2 across the X axis. It is very important at this point to become acquainted with such concept. Bearing in mind that one of the main characteristics of MTT is that it allows us to make predictions on candidates' answers based on probability theory (Green, 2013:xii), let us consider what McNamara (1996:165) writes about the probabilities or odds of a particular response:

> The odds are expressed as a logarithm ('log' for short) of the naturally occurring constant e. We thus speak of the 'log odds' of a response, rather than the odds of a response, and the units of measurement scale constructed in this way are called 'log odds units' or logits (pronounced 'LOH-jits'; stress on the first syllable). The logit scale has the advantage that it is an interval scale – that is, it can tell us not only that one item is more difficult than another, but also how much more difficult it is. The interval nature of the ability measurements means that growth in ability over time can be plotted on the scale; this has attractive implications for the evaluation of the effectiveness of teaching […]. By convention, the average difficulty of items in a test is set at zero logits. Items of above-average difficulty will thus be positive in sign, those of below-average difficulty negative in sign. Ability estimates in turn are related to item difficulty estimates, so that a person of an ability expressed as 0 logits would have a 50 per cent chance of getting right an item of average difficulty.

The most important thing about logits is that they will be our yardstick from now onwards. As we will see later in chapter 4 during the validation of the rubrics, logits will allow us to relate different aspects (or facets) of our

measurements to the same scale in the so-called vertical rulers, which will be very visual and convenient.

It is also important at this point to remark that 0 in a logit scale marks an average point and that this average point will vary from data set to data set depending on, for example, the average difficulty of items, the average ability of candidates, etc.

Once we know how our results will be scaled, it is time to go back to ICC. Since they are core to our analyses, let us see how they work through one example adapted from Bachman (1996:204-205), displayed as figure 2.4.2. For this example we will consider a three-parameter logistic graph, that is to say, a mathematical model which considers item difficulty, item discriminability and guessing (the 3 parameters) to tell us how likely one candidate is to answer a given item correctly. The graph below displays 3 different curves for 3 different items. The probability of one candidate answering one item correctly is displayed in the Y axis. The ability of candidates is displayed in the X axis and includes already a logit scale.

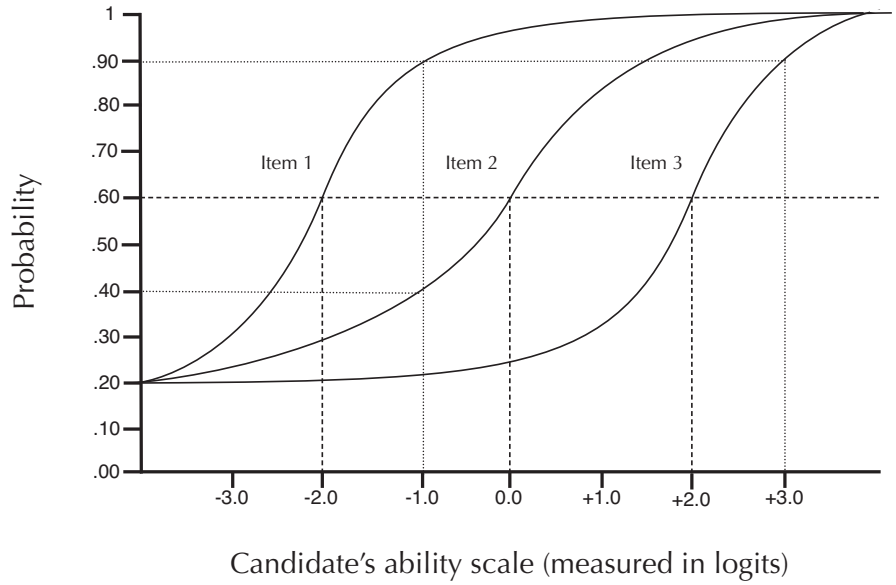

Figure 2.4.2. ICC curve of a three-parameter logistic model

The Y axis tells us how likely candidates are to answer one item correctly in relation to their ability (1 means 100% probability and 0 means 0% probability). The higher the ability of candidates (+1, +2, +3, etc.) the closer they will be to answering any item correctly, which is an intuitive idea. This way, for example, we see that a candidate with a logit ability of -2.0 will have 60% probability of answering item 1 correctly. Item 3 is clearly more difficult than item 1 because test takers must have an ability of +2.0 logits to hit 60% probability of correct answer. Likewise, a candidate with +3.0 logits ability will have 90% chance of answering item 3 correctly. Similarly, a candidate with a -1.0 logits ability will have 40% probability of answering item 2 correctly and 90% probability of answering item 1 correctly, etc.

Since this is a three-parameter logistic graph, it provides information regarding the 3 parameters above mentioned, which now we will present in a different order: guessing, item difficulty and discrimination.

The most interesting tenet here is the parameter of chance, since this model (the three-parameter model) is the only one that accounts for it. The Rasch model which we will be using later assumes that there is no chance in answers, which is as much as saying that there is no guessing. In contrast, in the graph above we see that the lower bound of the curves for the 3 items are asymptotic to .20. Being asymptotic, the lines will approach .20 but will never reach that value. This is the point at which the so-called pseudo-chance parameter is set (Bachman, 1991:205), which means that there is approximately 20% probability of candidates of very low levels answering the 3 items correctly as a result of (wild) guessing.

The middle point between the pseudo-chance parameter (.20) and 1 is called the difficulty parameter (Bachman, 1991:205), here set at .60 for the 3 items considered in the graph.

Finally, the discrimination of items is proportional to the slope of the curve at the point of the difficulty parameter. The steeper the slope is, the greater the discrimination of the item. Thus in our graph, item 2, which has the gentlest slope

will discriminate the least. Items 1 and 3, with much steeper slopes will discriminate much more effectively between individuals at different ability levels (Bachman, 1991:205).

If we used other probability models, the curves for these items would be different as well, describing different probabilities.

2.4.3 One-parameter logistic models and multi-faceted Rasch

There is no common agreement as to which MTT mathematical model is the best one. In fact, there may be no such thing as "one best model" since different models may be useful for different MTT analyses. The three-parameter logistic model, for example, is believed to be more adequate for data sets in which test takers at a very low proficiency level may get items correctly by chance, since guessing is one of the parameters that it accounts for, as we have seen in 2.4.2. On the other hand and most importantly for the present dissertation, as Bachman (1991:205) points out, the majority of MTT applications to language testing to date have used the Rasch model, which belongs to the one-parameter logistic family (see Deygers and Van Gorp, 2015; Ffrench, 2003 or Bruce and Hamp-Lyons, 2015 and, of course, the seminal McNamara and Adams, 1994). However, using the Rasch model for our validation is not a matter of tradition. It is a matter of adequacy, as we are going to see.

In the same fashion in which MTT displayed a variety of models, so does the Rasch model itself. Perhaps this is a good moment to have another look at figure 2.4 at the beginning of this section, where all these relationships are illustrated. McNamara (1996:254-257) distinguishes 3 main branches in the Rasch family, namely

1. the basic model for the analysis of dichotomous data,
2. models which can handle data from rating scales and the like (which includes the rating scale model or Andrich model and the partial credit model also known as Masters model) and
3. the multi-faceted Rasch model.

Out of the 3, we will be using for the validation of our rubrics the third one, the multi-faceted Rasch model because it allows us to analyze polytomous data, that is to say, data which are scored with a range of marks (McNamara 1996:254). When using our rubrics we will award each candidate a mark ranging from 1 to 5 depending on his performance. This range of marks is what we call polytomous data as opposed to dichotomous (*i.e.* true-false, right-wrong, etc.) data.

It seems counterintuitive not to use the second group of models (*i.e.* rating scale models) to analyze a set of rubrics but, as McNamara (1996:255) mentions, "the term *Rating Scale Model* is a technical label; it is not the only or even the most appropriate form of analysis for data from rating scales in general". In fact, as we shall see, the multi-faceted model provides a great deal of information and is more convenient for our purposes.

Rasch analyses offer all these and many other advantages when compared to CCT tools, which have different limitations as we have already discussed (see section 2.4.1). Knoch (2009:200-201) justifies the use of multi-faceted Rasch analyses over CTT methods as follows:

> [A]n ANOVA-based approach could be chosen to study group-level rater effects as well as rater-effect interactions. However, ANOVA has the limitation that possible interaction effects can contaminate main effects, making the interpretation of the main effects more difficult [...]. As mentioned earlier, multi-faceted Rasch measurement goes beyond the detection of main effects and interaction effects, as it allows for the detection of individual level effects. In this respect, multi-faceted Rasch measurement is superior to ANOVA-based approaches and regression approaches [...].
>
> Another approach possible when working with rating data is generalizability theory (or G-theory). One limitation of G-theory, which is addressed in multi-faceted Rasch measurement, is that although it identifies sources of variance attributed to each facet and its interactions, the impact of such differences on the candidates' scores during a particular examination is not corrected. Therefore, the candidates receive the raw scores they earn from the

> raters they encounter, and not an adjusted raw score due to rater differences or other attributes of the examination, as is produced in multi-faceted Rasch measurement.

And she finishes (Knoch, 2009:201) by assuring that Facets (Linacre, 2014)

> makes it possible to analyze data based on an analytic rating scale both as a whole (to see the functionality of the rating scale as a whole) or, by employing a partial credit model, with respect to each individual trait scale. It is also possible to investigate the rating behavior of all raters in the study as a group or individually or to investigate how each rater employs each individual trait scale.

The main advantage of Facets (Linacre, 1999) is thus that one analysis fits all the possible needs when validating analytic scales. In fact, by looking at particular data from the output provided by Facets (*ibid*.) we will be able to easily ascertain whether our rubrics are working properly. This will be done in section 4.2.4.

Let us finish this section about the benefits of multi-faceted Rasch analysis by quoting a very easy example that McNamara (1996:117-119) proposes to illustrate which type of problems Rasch can solve when compared to CTT.

Imagine that 2 candidates, Michael and Paula, are rated as regards their performance in one particular productive task. Michael is given a raw score of 5 in the task and Paula is given a 6. Apparently, Paula has proved to be more able at this particular task. However, a deeper look into the task, the raters that assessed it and the background of both candidates might provide us with a different perspective. Imagine for a moment that Michael's rater is stricter than Paula's (what McNamara labels as 'Hawk' vs. 'Dove') and imagine too that for some reason the task itself was tougher for Michael than it was for Paula (imagine that the task elicited a type of response that Paula was more used to, something likely in proficiency tests in which candidates from different backgrounds are concurrent). If all these aspects are taken into account, the conclusion drawn that Paula is more able than Michael (6 vs. 5) is, at least, arguable. McNamara even claims that in such a hypothetical situation we might consider that despite the

looks of the raw score, Michael has proved a higher level of ability than Paula. All this is illustrated the following table (McNamara, 1996:118):

Candidate	Rater	Topic	Rating	Ability
Michael	'Hawk'	Tough	5	(Higher)
Paula	'Dove'	Easy	6	(Lower)

Table 2.4.3. How raw scores can disguise real ability in performance

McNamara then wonders whether there is any way to account of and correct for the severity of raters, the difficulty of the task and similar *facets* impinging on measurements, to provide an accurate "picture of the ability of the candidate". He writes (1996:118):

> Answers to these questions can now be given in terms of the concepts and procedures of *multi-faceted measurement*, a new theory and method of measurement relevant to performance assessment situations such as the above.

2.4.4 Facets (Linacre, 2014)

In section 2.4 we have analyzed this far a lot of important aspects. We first saw the difference between CTT and MTT. We have seen the basics of probabilistic models and understood the rationale behind them. We have also described the characteristics of different mathematical models very superficially and we have seen how all this works through one three-parameter logistic model ICC. Finally, we have become acquainted with a new form of measurement, logits, which will be our most important yardstick in chapter 4.

Along this familiarization process we have been oblivious to the mathematical formulae upon which all these theories draw. Believe it or not, we have just scratched the surface. There are many such formula but their discussion is beyond the scope of the present dissertation. For the mathematical discussion of parametric models we suggest Harris (1989) and for the mathematical discussion of the Rasch model we suggest McNamara (1996), particularly chapter 9.

Luckily for linguists, it is possible to run complex data analyses without a deep knowledge of mathematics. This is indeed lucky since, after all, many of us moved into languages because we did not feel comfortable with mathematics. To establish a simple comparison, the relationship between statistics or mathematics and language testing is like the relationship between MSDOS programming and running Windows on your computer. Although the latter builds on the former, you can use Windows without any notion of software programming. Obviously, the more you know about MSDOS, the more advantage you can take of Windows. Even so, you can leave a long and peaceful life using Windows on a daily basis without having any idea of programming.

The same goes with language testing. There are several well-built software packages that can help us analyze our data from the CTT and MTT perspective, as we advanced in section 1.1.4. At the end of the day, these software packages are not more difficult to use than any Excel spreadsheet. While CTT is frequently linked to IBM SPSS (IBM Inc., 2016), MTT is linked to Quest, Winsteps (Linacre, 2016) and Facets (Linacre, 2014), although there are other packages like R (RDCT, 2016a), which are gaining popularity.

We will be using Facets (Linacre, 2014), which is particularly well suited for multi-faceted Rasch measurement. This software package was developed by the Australian researcher Mike Linacre as a result of his PhD in the late 80s of the last century (McNamara and Knoch, 2012:566). Some years later, McNamara and Adams (1994) wrote the first paper to ever use this software package applied to language testing. This first paper examined inter-rater consistency through data from the IELTS writing test. Since that moment on, the number of papers using Facets (Linacre, 2014) for the analysis of data applied to language testing has grown exponentially. Virtually all major conferences and meetings related to language testing around the world are likely to showcase 1 or various papers carried out through Facets (*ibid*.).

Facets (Linacre, 2014), through iterations of mathematical operations, tries to find patterns in the data it is fed with. By analyzing certain figures in the output

tables provided by the program, we can get to know if one individual response of our data set conforms to the general matrix of data as a whole. With all this information we can draw conclusions.

On the downside of these software packages we should say that, generally speaking, they are not very intuitive and have a very steep learning curve. At the same time they showcase a not very user-friendly interface which resembles MSDOS command boards. On top of this, Winsteps (Linacre, 2016) and Facets (Linacre, 2014) only run on Windows software. In the images below we reproduce some captions of the demo version of the package, which is available for free at www.winsteps.com.

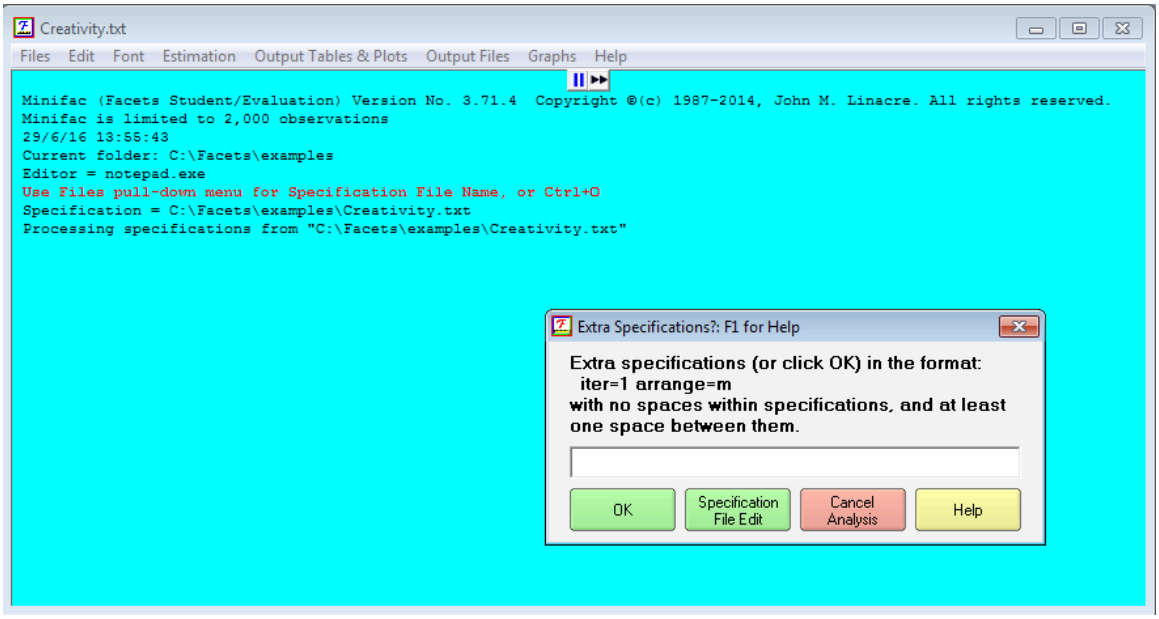

Figure 2.4.4.a. Main interface of Facets (Linacre, 2014)

Despite its many caveats in terms of design, Facets (Linacre, 2014) is a very powerful tool with a huge community of users around the world who are always willing to help each other. This is, perhaps, one of the factors that have boosted the popularity of the software. Its developer, Mike Linacre himself, is very supportive and active in discussions through the official website of the software package.

To set Facets (Linacre, 2014) to work, data must be fed to the main interface through a specification file►. Again, it is not a very intuitive system. The specification file, a txt document, contains all the instructions necessary for the program to yield its results. We can specify how many different aspects (or facets) we want to look at, which is going to be the referential element, etc. and, of course, it also includes the data themselves, which must be entered in a specific way.

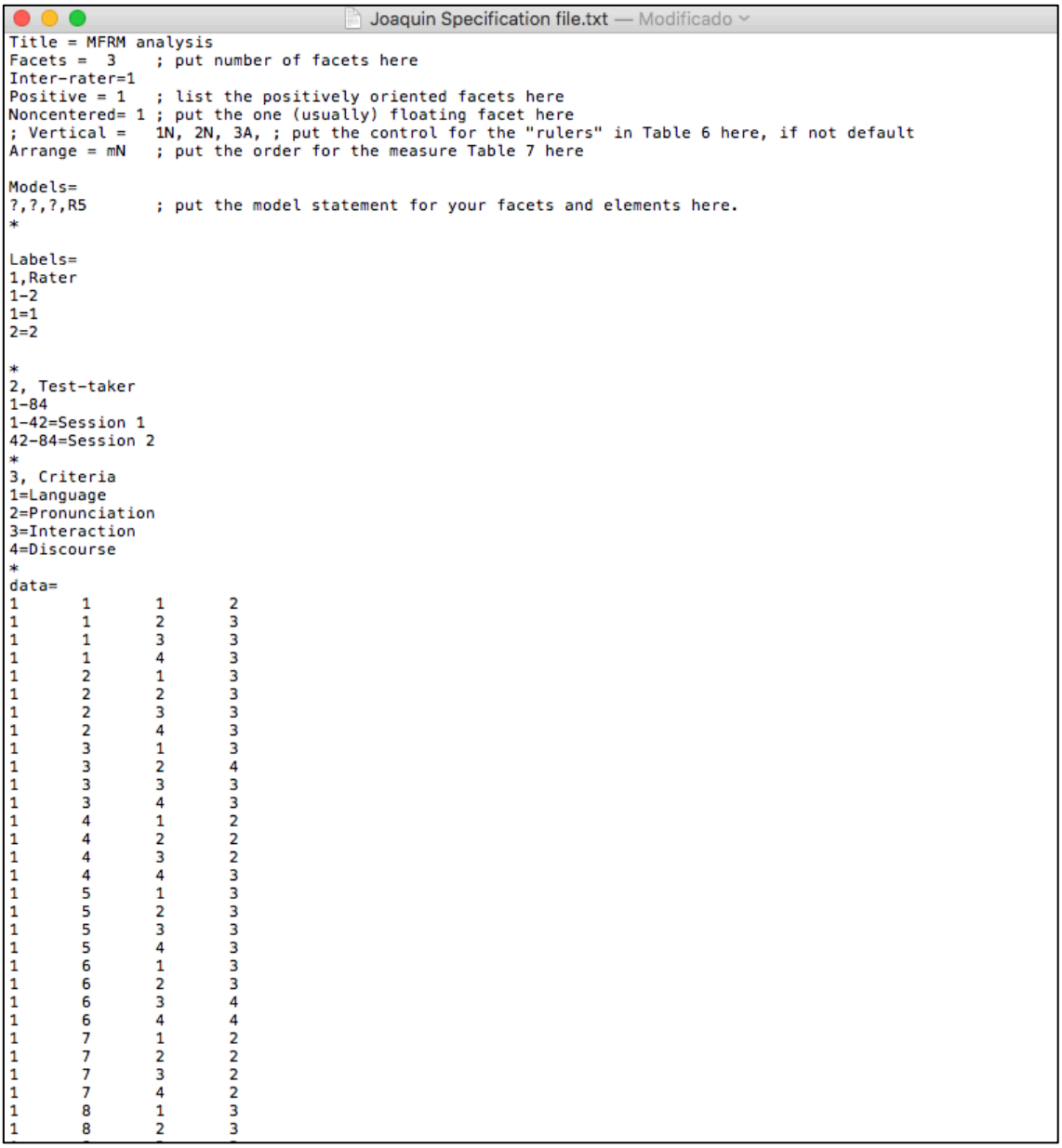

Figure 2.4.4.b Specifications file for Facets (Linacre, 2014) with instructions

The specification file is the first important step into the use of Facets. Data must be carefully constructed so that the results that the program yields are accurate. In figure 2.4.4.b we display the specification file of the first analysis of our rubrics. In the top part of the screenshot we find general instructions for the program, like the number of facets (or aspects) that we want to analyze or the order in which we want Facets to show the results in some of its tables. In the Model section of the screenshot, each question mark refers to the facets that data contain and R5 relates to the number of valid observations in our rating scale (Green, 2013:200). In our case, since our rubric contains 5 possible bands (or marks), that is what we are indicating through R5. In the lines below we can introduce the labels which characterize the different aspects of our data as for example the raters (here referred to as 1 and 2) or the name of the different linguistic criteria that our rubrics included (*Language*, *Pronunciation*, *Interaction* and *Discourse*). The final part of the screenshot contains the data to be analyzed and, as previously mentioned, these must be arranged according to a particular logic without which the results would not be reliable. In the image above, for example, we have 4 columns of data organized as follows:

1	1	1	2
1	1	2	3
1	1	3	3
1	1	4	3
1	2	1	3
1	2	2	3
1	2	3	3
1	2	4	3

The first row must be interpreted as Rater (1) > Test taker (1) > Criterion (1 - language) > Mark (2) (*i.e.* Rater 1 awarded candidate 1 a band 2 mark in language); the second row must be read as Rater (1) > Test taker (1) > Criterion (2 - pronunciation) > Mark (3) (*i.e.* rated 1 awarded candidate 1 a band 3 mark in pronunciation). In the fifth row we have Rater (1) > Test taker (2) > Criterion (1 - language) > Mark (3) (*i.e.* Rater 1 awarded candidate 2 a band 3 mark in

language), etc. If we could scroll down in the image, we would be able to see 672 such data rows (2 raters x 84 candidates x 4 criteria or linguistic features), which is the total number of snapshots that we used for our first validation trial. Then, as described in 4.2.4, we carried out a second trial with 9 raters and 44 candidates, which amounts a total of 1,584 rows of data to feed into Facets (Linacre, 2014).

All these data are our data matrix, the data matrix through which Facets will look for specific patterns. Facets will try to find out if individual marks conform to the general patterns underlying in the matrix and it will yield output tables pointing at those data which do not fit the mathematical model and cause "noise" in their interpretation. According to Deygers and Van Gorp (2015:528),

> (i)n a robust rating scale raters and rating criteria fit the Rasch model. The Infit Mean Square (Infit MnSq) value is a good indicator of such a model fit. The closer Infit MnSq approaches 1, the better a rater or a criterion fits the Rasch model.

As a consequence, Infit and Outfit MnSq will be among the first values that we will analyze in our data. All these parameters will be used profusely during chapter 4 and, at that point it will be a good idea to come back to section 2.4.4. to revise the most important concepts.

CHAPTER 3. THE CONTEXT

This chapter is not directly linked to our construct of language (chapter 1) or to our construct of assessment (chapter 2) but provides the big picture of the circumstances under which our experiment was envisaged and developed. At the same time, chapter 3 is a convenient historical account of some paramount events in the recent history of language assessment.

Modern high-stakes big-scale tests are operants in their context. Their environment influences them and they also shape it. Bachman points out that "testing almost never takes place in isolation. It is done for a particular purpose and in a specific context" (1995:2). Tests act upon the world and change it, and are changed in turn by the consequences of their own impact. This is partly what we referred to as consequential validity in section 2.2.1. Tests are modified not only by the consequences of their action but also by a plethora of other factors. Spolsky (1995:3), for example, mentions that "(t)he move towards a European economic community clarified, as no theoretical approach would have done, the requirement to define language teaching goals as precisely as did the notional-functional syllabus". Through this statement Spolsky is stressing the importance that shared economical, commercial goals have had in the post-modern history of language testing and in the evolution of testing industry nowadays. To a great extent, this was present at the genesis of the *CEFR* (Council of Europe, 2001).

Tests are critically influenced by politics but they also influence upon their environment. The technology of testing has been used as a control tool by many different societies. Characteristic examples are the first competitive exams in China (Spolsky, 1995:16), the solution given in the United Kingdom to the supposed degeneration in national intelligence after the Boer wars (Fulcher, 2009:6) or the most recent case of Theresa May's massive deportation of 48,000 students accused of cheating at their language exams (Menon, 2016; Ali, 2016).

The creation of the *CEFR* (Council of Europe, 2001) has helped from a local perspective areas like southern Europe to speed up their pace towards global convergence with the World's *lingua franca* and with other minority linguistic realities that would not have been paid attention to otherwise. In our opinion, there is, in fact, a positive pan-European rationale behind the *CEFR* (*ibid*.). This pan-European view, as we envisage it, aims at recognizing Europe's different linguistic realities without allowing privileges to none since the *CEFR* (*ibid*.) was not written having any particular language in mind. Test linkage to the *CEFR* (*ibid*.) around the world is not currently being used to achieve "collectivist goals" because of a sense of "external threat" co-occurring with an "insider-outsider mentality", as Fulcher (2009) puts it. On the contrary, the fact that some South American and Asian countries are voluntarily using the *CEFR* (*ibid*.) scales and methods might suggest that we are on the verge of different type of test use, a sort of global use of common references.

Spolsky (1995:313) was the first to use the expression "the English testing industry". Nowadays we could speak simply of "the industry of testing", a coinage that reflects the dangerous relationship that we see between testing and business nowadays. It is dangerous, indeed, to shape a society solely on the grounds of market needs. The money generated by this industry is already very attractive to competing old and new companies. This may sound a bit exaggerated, but it is not if we consider the impact of some tests. Some tests are a requirement for obtaining employment, or provide the key for immigrants wishing to gain entry to a country. Tests are likely to interest thousands of people hoping for a place at college, for a better job or for a better future. In the USA, for example, TOEFL tests are a must-have for foreign students who seek to access universities. In 2014, according to internal documentation of their Spanish branch, ETS implemented 50M tests in more than 180 countries. Cambridge tests, which celebrated their first hundred years in 2013, are taken by 4M candidates in more than 130 countries. Out of these candidates, 250K are Spanish, which goes to show the penetration of the suite of Cambridge exams in Spain. Such figures are a

remarkable success in marketing and adaptation for Cambridge, when we consider that the first modern Cambridge test, held in 1913, lasted 12 hours and was taken by only 3 candidates, all of whom failed (Hawkey and Milanovic, 2013:22). Cambridge and TOEFL are just 2 examples out of the surfeit of test boards existing in most important languages (ACTFL, TOEIC, EIKEN, IELTS, Aptis, Trinity, DELE, eLade, HSK, etc.) that are ubiquitous nowadays.

The industry of testing moves millions of euros worldwide and has even borrowed many aggressive practices typical of pharmaceutical marketing, which is perhaps one of the first indications of the power at stake. The expectation generated amongst platinum centers, distribution centers that conform to Cambridge's most ambitious business development programs, is just one measurement of the current industry of testing. Trinity tests, for example, have gathered momentum in the Spanish market following their recognition by a number of important institutions. In Spain, Trinity tests have surpassed in popularity ETS and even the most popular suite of exams in the past twenty years in Spain, Cambridge. Leaving aside reliability and construct validity concerns, Trinity has been quick at meeting the requirements of certain regions like Andalusia with a poor tradition in foreign language recognition, and has been able to build very powerful face validity through aggressive campaigns.

One of the most evident consequences of this proliferation of exams and examination boards in their environment is, perhaps, the washback that some English tests originate in China. In the fourth fiscal quarter of 2015, the Chinese provider of private educational services New Oriental increased its total net revenues by 14,4% year-over-year to $328.8M. Surprisingly, 2015 was not a good year for New Oriental due to the uncertainty about the implementation of the new policies relating to the English test for Gaokao, the Chinese National Higher Education Entrance Examination. Some of New Oriental's lecturers are treated like celebrities and they have earnings to match, for some have become millionaires and appear regularly on television shows not for being athletes or pop stars, but

for teaching English. According to its own web page (New Oriental, 2015), New Oriental boasts +20M students since its founding with a yearly average of 2.7M student enrolments. New Oriental has +720 learning centers distributed in 50 cities around China and employs 33,000 people including over 17,000 teachers. The preparation of candidates for English proficiency exams is at the core of New Oriental's services which has provoked hostility on behalf of other brands for having used, copied and sold its exams illegally allegedly. There is little doubt that, with different degrees of penetration, we are now living the heyday of the industry of language testing.

As we learn from the case of New Oriental, testing does not occur in a vacuum. In Chapter 3 we aim to describe the 3 levels (international, national and regional) that have shaped and that interact with our experiment, our rubrics and the exams they are designed for. This will provide the present dissertation, we hope, with a more specific context. Section 3.3 contains, we think, a relevant contribution. It does not only describe the most immediate context that affects our tests and on which our tests impinge (the regional context) but it also captures, almost in real time, the joint effort of 9 public Spanish universities towards mutual understanding and recognition. A joint effort at such a scale is unprecedented in Spain and section 3.3 has been conceived as a log to account for the achievements already obtained by these 9 public universities, as a record of their effort.

3.1 European policies

Nowadays, "(t)he international mobility of students and staff, and the desire to attract a global and diverse student body, appear to be making English the second language of many European universities" (Extra and Yağmur, 2012:10). Together with this, the status of English as the *lingua franca* of business worldwide partially explains the growing interest of Europeans in this language. These 2 facts are at the core of our decision of developing the rubrics that are described in chapter 4.

Notice too that, despite the fact that these rubrics were designed in English and to be used in English proficiency tests, they are easily adaptable to other languages.

European mobility blurs now national borders and those students and professionals that move from country to country as well as their employers demand certifications of language skills that can help screening prospective students and prospective workers prior to face-to-face interviews. These certifications are an indicator of the skills of candidates that, more often than not, determine their access to study programs or their employability.

Mutual recognition in Europe through the *CEFR* (Council of Europe, 2001) is nowadays a reality, but it has not always been like this. In fact, for many years the lack of common standards hampered internationalization within European borders in many ways. The *CEFR* (*ibid*.) is the result of a long way which started in the middle of the 20th century with an objective that had little to do with languages or testing in its inception. The *CEFR* (*ibid*.) is a product of the vision of Europe that some countries had after World War I. Curiously enough, one of such countries, England, the same country which gave the world the most widely spoken language in history, which is also the basis of our experiment, has recently decided to quit the enterprise started in 1949 with the foundation of the Council of Europe►. But Brexit, which is the name given to the British exit of Europe, is a topic for another type of dissertation, a political and economic dissertation which someone is surely writing right now. Let us leave this aside and consider first the way in which it all started with the foundation of the aforementioned Council of Europe.

3.1.1 The Council of Europe

The idea of Europe has not always been the same. The modern process of European construction was launched back in 1949 when the Council of Europe was founded. After 2 world wars, the Council of Europe aimed at protecting democracy and human rights and at promoting European unity by fostering cooperation on different matters. The European Cultural Convention of the

Council of Europe of 1953 was the first of a series of acts and treaties oriented to set forth the fundamental rights and freedoms which were the core concern of post-war Europe (Council of Europe, 2016). Culture and eventually language were among these concerns as well.

After the initial steps of the Council of Europe, it took European higher education almost 40 years to unify cultural goals. It was not until 1988 when the rectors of 388 universities signed the *Magna Charta Universitatum* (*MCU*, 1988) (*MCU* henceforward) on the 900th anniversary of the University of Bologna. The *MCU* was a two-page document designed to lead Europe into a culture-based new millennium which contained principles of academic freedom and autonomy as a guideline for good governance and mutual recognition of universities. Nine years later, in 1997, the Council of Europe and the UNESCO drafted the *Lisbon Convention* (1999), which was designed to streamline the legal framework at a European level and to replace 6 previous conventions in matters of higher education. In 1998, 4 education ministers (from France, Germany, the United Kingdom and Italy) participating in the celebration of the 800th anniversary of the University of Paris shared the view that the segmentation of the European higher education sector was outdated and harmful. As a consequence, they agreed to sign the *Sorbonne Declaration* (1998). The document put forward a number of ideas about the European Credit Transfer and Accumulation System (ECTS►) and distinguished between the 2 main cycles of the system, undergraduate and graduate. For the first time, the document officially called for a recognition system able "to remove barriers and to develop a framework for teaching and learning, which would enhance mobility and an ever closer cooperation" (*Sorbonne Declaration*, 1998:1). These lines have echoed in Europe, repeated as a mantra, ever since. One year later, in 1999, the European Higher Education Area (EHEA►) was finally shaped, through the signing of the *Bologna Declaration* (1999).

At the end of the road, the idea of mutual recognition was officially born in Europe in 1999 with the *Bologna Declaration*, three-quarters of a millennium after the first universities came into being in the Old Continent. In retrospect, the

Bologna Declaration has become one of the most influential documents in the modern history of European higher education. After 1999, other communiqués► have been issued (Prague, 2001; Berlin, 2003; Bergen, 2005; London, 2007; Leuven/Louvain-la-Neuve, 2009; Bucharest, 2012 and Yerevan, 2015), articulating the previous agreements, including some partners from beyond Europe, and weaving a brand new European network. EHEA policies have also been updated at the Budapest-Vienna Ministerial Conference (2010) and the Bucharest Ministerial Conference (2012) as well as through further Declarations (Salamanca, 2001; Graz, 2003; Glasgow, 2005; Lisbon, 2007 and Budapest-Vienna, 2010) and through 3 Bologna Policy Forums (2009, 2010 and 2012), as we shall mention in the next section.

As we see, the Council of Europe "provides a pan-European forum for sharing expertise and experience based on common values and respect for the diversity of contexts" (Extra and Yağmur, 2012:7). It acknowledges the particularities of the state members and lessens privileges among them.

3.1.2 The Bologna process

It is easier to understand the evolution of the current idea of education in Europe if the span of time that goes from 1949 to present is divided into 2 periods. First, from 1949 with the foundation of the Council of Europe until 1999, the year of the *Bologna Declaration*. Then, from 1999 to present day. The latter is the period in which the Bologna process has taken place.

The so-called Bologna Process has framed European educational policies since 1999, if not earlier. This process has given birth to realities like the EHEA or the *CEFR* (Council of Europe, 2001). Supported by the Council of Europe, the Bologna Process has, since its inception, aimed at recognizing the differences of the state members.

The evolution of the Bologna Process is best traced back through the different official documents that it has originated, the most important of which are listed in table 3.1.2 below:

The road to the Bologna process	
1949	The Council of Europe is founded
1953	European Cultural Convention of the Council of Europe
1988	*Bologna Magna Charta Universitatum*
1997	*Lisbon Convention*
1998	*Sorbonne Joint Declaration*
The Bologna process	
1999	*Bologna Declaration*
2001	Prague Higher Education Summit - *Prague Communiqué*
2003	Convention of European Higher Education Institutions in Graz - *Graz Declaration*
2003	Berlin Conference of Ministers - *Berlin Communiqué*
2005	Conference of Ministers in Bergen - *Bergen Communiqué*
2005	Convention of European Higher Education Institutions in Lisbon - *Lisbon Declaration*
2007	Conference of Ministers in London - *London Communiqué*
2009	Conference of Ministers in Leuven and Louvain-la-Neuve - *Leuven and Louvain-la-Neuve Communiqué*
2009	First Bologna Policy Forum - *Statement by the Bologna Policy Forum*
2010	Conference of Ministers in Budapest and Vienna - *Budapest and Vienna Declaration*
2010	Second Bologna Policy Forum - *Bologna Policy Forum Statement*
2012	Conference of Ministers in Bucharest - *Bucharest Communiqué*
2012	Third Bologna Policy Forum - *Statement of the Third Bologna Policy Forum*
2015	Conference of Ministers in Yerevan - *Yerevan Communiqué*
2015	Fourth Bologna Policy Forum - *Statement of the Fourth Bologna Policy Forum*

Table 3.1.2. Timeline of the Bologna process

The Bologna process was definitely launched with the *Bologna Declaration*, which is now "one of the main voluntary processes at European level, as it is nowadays implemented in 48 states, which define the European Higher Education Area" (EHEA, 2014a). As the *Bologna Declaration* (1999:2) itself claims, "the vitality and efficiency of any civilization can be measured by the appeal that its culture has for other countries". That appeal of European culture is perhaps the reason why so many countries have joined this venture along the way. The countries that signed the *Bologna Declaration* (1999) expressed their willingness to commit themselves to enhancing the competitiveness of the European Higher Education Area. Three-quarters of a millennium after the first universities were born in Europe and at the turn of a new one, the signing

countries emphasized the need to further the independence and autonomy of all Higher Education Institutions in the context of a "growing awareness […] of the need to establish a more complete and far-reaching Europe, in particular building upon and strengthening its intellectual, cultural, social and scientific technological dimensions" (*ibid*.:1). All the provisions of the *Bologna Declaration* were set as measures of a voluntary harmonization process, not as clauses of a binding contract, which somehow explains the fact that, even nowadays, different countries abide by the Declaration at different levels (EHEA, 2014b). The basic objectives considered of primary relevance by the *Bologna Declaration* in order to establish the EHEA were 1) the adoption of a system of easily readable and comparable degrees; 2) the adoption of a system based on 2 cycles, undergraduate and graduate; 3) the establishment of a system of credits, the European Credit Transfer System (ECTS) system; 4) the promotion of mobility for students and for teachers; 5) the promotion of European co-operation in quality assurance and 6) the promotion of the necessary European dimensions in higher education (*Bologna Declaration*, 1999).

From its inception, the *Bologna Declaration* has been followed up every 2 years by Ministerial Conferences after which the participating Ministers issue their *Communiqués*. In 2001, after the Prague Higher Education Summit and through the *Prague Communiqué* (2001), the number of member countries was taken to 32 at the same time that the objectives of the Declaration were extended to include students as active partners. Also, the participating ministers committed themselves to ensure further development of the 6 objectives that *The Bologna Declaration* (1999) had established 2 years before. The idea of social dimension was first introduced in the *Prague Communiqué* (2001) and encompassed, along with the whole document, with a strong Pan-European view which aimed at involving "the whole of Europe in the process in the light of enlargement of the European Union" (*ibid*.:1) in which the ongoing creation of the EHEA was considered a reality that had

> Ministers reaffirmed that efforts to promote mobility must be continued to enable students, teachers, researches and administrative staff to benefit from the richness of the European Higher Education Area including its democratic values, diversity of cultures and languages and the diversity of the higher education systems.
>
> *Prague Communiqué* (2001:1)

The *Prague Communiqué* (2001) also emphasized 3 more points, namely 1) the idea that lifelong learning is an essential element of the EHEA, 2) the aforementioned inclusion of students as active partners, together with other higher education institutions as constructive partners and 3) the promotion of the attractiveness of the EHEA.

The signatories accepted the applications from Croatia, Cyprus and Turkey, committed to celebrate a follow-up meeting 2 years later, in 2003 in Berlin, and, most importantly, referred for the first time in the historical series of documents to the Council of Europe as one institution that "should be consulted in the follow-up work" (*Prague Communiqué*, 2001:3). This last inclusion is particularly important if we take into consideration the key part that the Council of Europe has played in the design of joint European education policies ever since.

Two important events for the EHEA took place in 2003 almost in parallel. On the one hand, the Second Convention of European Higher Education Institutions celebrated in Graz and, on the other, the Ministerial Conference of Berlin. After the Convention, the participant institutions signed the *Graz Declaration* (2003) which principally focused on the role of universities within the process of European construction. The document described priorities for action and set goals for 1) maintaining universities as public responsibility; 2) consolidating research as an integral part of higher education; 3) improving academic quality by building strong institutions; 4) furthering mobility and the social dimension; 5) supporting the development of a policy framework for Europe in quality assurance, and 6) pushing forward the Bologna process.

On the other hand, the Ministerial Conference of Berlin (also in 2003) gave rise to the *Berlin Communiqué* (2003), which enlarged the number of countries to 40 and welcomed the presence of several representatives from European countries not party to the Bologna Process by that time, as well as the presence of the Committee of the European Union, Latin America and Caribbean Common Space for Higher Education. The Process was thus not only being extended to other European countries (Albania, Andorra, Bosnia and Herzegovina, Holy See, Russia, Serbia and Montenegro), but also opened to the world beyond Europe. The *Berlin Communiqué* (2003) also introduced economic concerns that time and liberalism have eroded, namely the idea of turning Europe into "the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion" (*ibid*.:2). The main provisions of *Berlin Communiqué* (2003) dealt with an expansion of the objectives, in terms of promotion of linking the EHEA to the European Research Area►, as well as the promotion of "effective quality assurance systems, to step up effective use of the system based on two cycles and to improve the recognition system of degrees and periods of studies" (*ibid*.:3). In the design of the recognition system, the importance of the *Lisbon Convention* (1997) was underlined. But most importantly, the *Berlin Communiqué* (2003) established the follow-up structures supporting the process in-between Ministerial meetings. This arrangement established the Bologna Follow-up Group, the Board and the Bologna Secretariat. Ministers also agreed that there should be created a national follow-up structure in each of the participating countries.

The *Bergen Communiqué* (2005), which followed a new Ministerial Conference, underlined the importance of partnerships which included stakeholders (students, higher education institutions, academic staff and employers) and highlighted the relevance of research, especially with regard to the third cycle and doctoral programs. The *Bergen Communiqué* (2005) also stressed the Ministers' will to provide a more accessible higher education, together with an increased attractiveness of the EHEA to other parts of the world.

The European University Association met in 2005 in Lisbon during the Convention of Higher Education Institutions that followed the aforementioned Convention held in Graz in 2003. During the 2005 Convention, the participating institutions agreed to sign the *Lisbon Declaration* (2007) which would be published 2 years later. The *Lisbon Declaration* (2007) revised the key goals in building the EHEA, its internationalization and also revisited the promotion of research and innovation in higher education. Quality, autonomy and funding were further aspects dealt with by the *Lisbon Declaration* (2007). Again, the internationalization of the EHEA was one of the main concerns in the final conclusions drawn from the meeting:

> Europe's universities are a major force in shaping the Europe of Knowledge. They accept the responsibilities which this brings and, in return, ask that governments, and civil society in general, should recognize their responsibility to enable universities to secure the resources which will permit them to fulfill their mission not just well, but with excellence and in a way which allows them to compete with the higher education systems of other continents. Not just Europe but the whole world, is becoming a "Knowledge Society".
>
> *Lisbon Declaration* (2007:7)

The *Lisbon Declaration* (2007:1) also "provides the basis for the message that EUA (European University Association) presented to Ministers of Education meeting in London on 17/18 May 2007" [4]. During this new Ministerial Conference held in London and through the *London Communiqué* (2007), the number of participating countries was enlarged to 46. The *London Communiqué* (2007) focused on evaluating the progress achieved by that time concerning mobility, degree structure, recognition, qualifications frameworks (both overarching and national), lifelong learning, quality assurance, social dimension, and it also set the priorities for 2009, these being, mainly, mobility, data collection, employability,

[4] Brackets not in the original.

EHEA in a global context, stock taking and the social dimension of the process, which was defined here for the first time:

> Higher education should play a strong role in fostering social cohesion, reducing inequalities and raising the level of knowledge, skills and competences in society. Policy should therefore aim to maximize the potential of individuals in terms of their personal development and their contribution to a sustainable and democratic knowledge-based society. We share the societal aspiration that the student body entering, participating in and completing higher education at all levels should reflect the diversity of our populations. We reaffirm the importance of students being able to complete their studies without obstacles related to their social and economic background. We therefore continue our efforts to provide adequate student services, create more flexible learning pathways into and within higher education, and to widen participation at all levels on the basis of equal opportunity.
>
> *London Communiqué* (2007:5)

The Bologna process was not only gaining momentum but also transcending European borders, as we have mentioned several times before. This was made clear in the full title of the *London Communiqué* (2007) (*Towards the European Higher Education Area: Responding to Challenges in a Globalized World*) and in the new global strategy adopted. For the goal of 2010 and beyond, further collaboration would be seen as an opportunity to reformulate visions and values.

The next Conference of Ministers took place in Leuven and in the planned city of Louvain-la-Neuve which was built to host the Université Catholique de Louvain after the linguistic Leuven crisis in the 1960s. In the *Leuven and Louvain-la-Neuve Communiqué* (2009) that followed the Conference, the main working areas for the next decade were set, again with emphasis on: social dimension, lifelong learning, employability, student-centered learning and the teaching mission of education, international openness, mobility, education, research and innovation. These main working areas showed a new path for the Bologna process oriented to ensuring the completion of the process itself.

As of 2010, according to the agreements of 2009, the Bologna process would also shift from a previous situation in which it was chaired by the country holding the European Union Presidency, to a situation in which it would be chaired by 2 countries: both the country holding the European Union Presidency and a non-EU country.

The Secretariat created after the Ministerial Conference held in Berlin in 2003 published, following the Leuven and Louvain-la-Neuve Ministerial Conference, the report *Bologna Beyond 2010* (2009), which served as background paper which summarized in 33 pages the achievements that the process had obtained up to that year.

The following Ministerial Conference took place only 1 year after the previous one, in March 2010. It took place in Budapest and Vienna and it was an anniversary conference, celebrating a decade of the Bologna process. With this occasion, there took place the official launching of the EHEA, which meant that the objective of developing a common European framework for higher education set by the *Bologna Declaration* (1999) had been accomplished. The conclusions of the conference were published in the *Budapest and Vienna Declaration on the European Higher Education Are*a (2010). Set against this, the existence of the EHEA, in itself, did not entail that all the objectives of the Bologna process had been achieved. It did mean, however, that the Bologna process and the EHEA have entered a new phase, a phase of consolidation and operationalization.

The main message of the Bucharest Ministerial Conference, which took place on 26-27 April 2012 and was attended by 47 European ministers, stated that Higher Education reforms could help to get Europe back on track and to generate sustainable growth and jobs (*Bucharest Communiqué*, 2012). Ministers agreed to focus on 3 main goals in the face of the economic crisis: to provide quality higher education to more students, to better equip students with employable skills, and to increase student mobility. The 47 signatories adopted a new European strategy to increase mobility with a specific target that at least 20 percent of those graduating in Europe in 2020 should have been on a study or

training period abroad.

The latest Ministerial Conference up to the printing of the present dissertation was the one held in Yerevan in May 2015. The *Yerevan Communiqué* (2015) acknowledges the impact of the global financial crisis that peaked in 2007 and 2008, whose consequences were in 2015 more visible than ever:

> Today, the EHEA faces serious challenges. It is confronted with a continuing economic and social crisis, dramatic levels of unemployment, increasing marginalization of young people, demographic changes, new migration patterns, and conflicts within and between countries, as well as extremism and radicalization.
>
> *Yerevan Communiqué* (2015:1)

This new vision of an endangered EHEA contrasts dramatically with the knowledge-centered society envisaged by the *Berlin Communiqué* (2003). On the bright side, the *Yerevan Communiqué* (2015:1) also acknowledges that "automatic recognition of qualifications has become a reality so that students and graduates can move easily throughout it". Despite there is still a long way ahead, this is a reality that can be easily perceived in universities all around Europe not only as regards degree studies but also, and most importantly for the present dissertation, as regards language proficiency recognition. The *CEFR* (Council of Europe, 2001), published at the turn of the century as an outgrowth of the Bologna process, has crept into universities across Europe and has provided a clear reference for the alignment of language proficiency tests across the continent as well as in other parts of the world.

Different Bologna Policy Forums have also been organized in parallel to the Ministerial Conferences described in the paragraphs above. The First Bologna Policy Forum took place during the Ministerial Conference of Leuven and Louvain-la-Neuve in 2009, and it was attended by the 46 members of the Bologna process at the time, as well as by a wide range of third countries and non-governmental organizations. The main issues agreed upon by the participants were, roughly speaking, the same ones that the Ministerial Conferenced dealt

with. The idea of a knowledge-based economy that had appeared in previous Communiqués is now partially discontinued and somehow replaced in the *Statement by the Bologna Policy Forum* (2009:1) by the reality that Europe is living at the moment in which the forum takes place: "(w)e underline the importance of public investment in higher education, and urge that this should remain priority despite the current economic crisis, in order to support sustainable economic recovery and development". The document uses the word "crisis" and declares that transnational exchanges in higher education should be governed on the basis of academic values, advocating a balanced exchange of teachers, researchers and students between countries, in order to promote fair and fruitful "brain circulation", as an alternative to brain drain.

The Second Bologna Policy Forum took place in Vienna, in March 2010, and it was attended by the 47 members and the 8 consultative members, as well as third countries and other relevant non-governmental organizations. The main topics of discussion included in the *Bologna Policy Forum Statement* (2010) refer to the manner in which higher education systems and institutions responded to the growing demands and multiple expectations and the balance between cooperation and competition in international higher education. The document also included some possible concrete feedback to be taken up by the participants, such as nominating contact persons for each participating country, which were intended to function as liaison points for a better flow of information and joint activities, including the preparation of the next Bologna Policy Forum at ministerial level. Also the need for supporting global student dialogue was acknowledged.

The Third Bologna Policy Forum was organized in conjunction to the Ministerial Conference held in Bucharest in 2012. It contributed to further the debate on the progress of the EHEA at a global scale. Members and delegations attended it from 47 EHEA countries and 19 non-EHEA countries along with representatives of international organizations from the field of Higher Education. The *Statement of the Third Bologna Policy Forum* (2012) focused on

creating and connecting national, regional and global Higher Education spaces, while deepening the discussions on the topics of public responsibility within national and regional contexts, global academic mobility and barriers, global and regional approaches to quality enhancement of Higher Education along with employability issues. The participants stated that the Bologna Policy Forum objectives and decisions should be further enriched and taken forward in order to maximize its potential for policy dialogue.

The Fourth Bologna Policy Forum took place during the Ministerial Conference held in Yerevan in 2015. The *Statement of the Fourth Bologna Policy Forum* (2015:1), again, acknowledges the problems that the Bologna process had (and still has) to face, "[p]olitical instability in many of our countries, a high level of unemployment and migration arising from economic and social crisis and lack of access to higher education […]". The text also aimed at "[i]mproving the mutual recognition of qualifications, through improved information, the joint development and dissemination of recognition practice and methodology" (*ibid*.:2).

The long way described in the previous paragraphs, ushered by the Council of Europe, has given birth to an unprecedented level of mutual understanding and mutual recognition in Europe. For the present dissertation, the most relevant achievement is, without any doubt, the *CEFR* (Council of Europe, 2001). This document has allowed millions of students of secondary and tertiary education to transcend their national borders. Thanks to the reference levels established by the *CEFR* (*ibid*.) we now have a common yardstick against which to compare the outcome of our teaching practices, the results of our tests in general and the results of our language proficiency examinations in particular.

3.2 Spanish policies

The impetus for tests of student competence (as language proficiency exams) is due in large part to the lack of public trust in the soundness of criteria in place

prior to such tests (Cizek and Bunch, 2007:8). Unfortunately, the popularity of proficiency language tests in Spain is no exception.

In Spain, where the boom in language testing is unprecedented, marketing arguments seem to be leading the choices of test takers in the first part of the 21st century. Some tests are held in massive venues such as hotels or trade fair parks which host thousands of candidates. The unaware observer may have trouble in saying whether candidates are actually going to sit a test or to watch the local football team as test takers make their way towards test venues. Such is the amount of people that language tests are able to bring together.

The washback from these tests has also been important in a country that relied heavily on traditional methods of language teaching, certification and accreditation. The language teaching methods used in Spain until very recently were inherited from the Grammar-Translation principles used extensively in the teaching of dead languages, which are decidedly unsuited to teaching modern, living languages. Catching up with the rest of Europe has necessitated profound changes in the mindsets of Spanish professionals and, even nowadays, at times, Spain seems to be stuck in second gear while the rest of Europe is working at full speed.

From the last major law on education passed by dictator Franco in 1970, Spanish policy makers have passed 6 other major laws over the past 36 years, each of them intended to substitute the previous one (*BOE*, 1970; 1980; 1985; 1990; 2002; 2006 and 2013).

Six of these laws are organic (see table 3.2 below) which means that, under the current Spanish Constitution, which dates back to 1978, they have an intermediate status between an ordinary law and the Constitution itself. This gives a taster of their importance and of the impact that these have had. To make this surfeit of laws even more complex, Spain hosts 17 autonomous communities (or regions), all of which have their own laws on education, as we will see in section 3.3.

Main Spanish national laws on education from 1970 to 2016	
1970	Ley General de Educación y Financiamiento de la Reforma Educativa (*BOE*, 1970)
1980	Ley Orgánica por la que se regula el Estatuto de Centros Escolares (*BOE*, 1980)
1985	Ley Orgánica reguladora del Derecho a la Educación (*BOE*, 1985)
1990	Ley Orgánica de Ordenación General del Sistema Educativo (*BOE*, 1990)
2002	Ley Orgánica de Calidad de la Educación (*BOE*, 2002)
2006	Ley Orgánica de Educación (*BOE*, 2006)
2013	Ley Orgánica para la mejora de la calidad educativa (*BOE*, 2013)

Table 3.2. Main Spanish national laws on education from 1970 to 2016

Unfortunately, educational regulation is seen in Spain as a political tool both by national and regional governments, which also hold devolved powers. The political party that wins the presidential elections every 4 years normally substitutes structural educational laws. This has lead to a scenario in which educational policies, whether bad or good, are not allowed to thrive. Confrontation among political parties frequently leads to national confusion, of which there are many examples. The LOMCE (*BOE*, 2013), to quote a case, which generated considerable turmoil, aimed to change higher study programs from a 4+1 configuration (4 years for degree studies and 1 for Master's studies) to a 3+2 structure. Being both quite extended in different countries around the world, where the LOMCE (*ibid*.) failed was at achieving consensus among stakeholders on which of the 2 configurations could be the most beneficial. Another sad example of how consensus is broken by Spanish policy makers was the implementation of the so-called *Reválida*, a school leaving exam. The *Reválida* of the LOMCE (*ibid*.) was an external test for 12-year-old students aimed at checking their progress in the areas of language, mathematics, science and technology. After the LOMCE (*ibid*.) was passed and widely criticized on December 2013, Spain held national elections in December 2015. Due to the tight margin of votes obtained by the main political parties, Spain's incumbent government had little authority to implement the mentioned progress test and 12 out of the 17 autonomous communities of Spain refused to bring the test live on the grounds

that it had not been properly designed and that it did not guarantee fairness to all test takers. The leaders of the 5 autonomous communities which accepted to administer the exam belonged to the same political party of the incumbent government. This situation affected +460K students, out of which +333K belonged to the 12 autonomous communities that refused to administer the exam (Álvarez, 2016).

The resulting variegation and uncertainty has hampered, among other things, mutual recognition of foreign language levels. As a result, depending on the community chosen, the linguistic proficiency of 2 different students may vary by up to 2 *CEFR* (Council of Europe, 2001) levels in the same academic year. It is because of this lack of intra-regional standardization that it has become necessary to establish external language tests whose results can clearly be linked to the *CEFR* (*ibid*.), which brings us back to Cizek and Bunch's (2007:8) assertion that the impetus for tests of student competence (as language proficiency exams) is due in large part to the lack of public trust in the soundness of criteria in place prior to such tests.

Our rubrics were designed for university language proficiency exams. The latest milestone in the history of Spanish university policies dates back to 2007, when the Spanish Ministry of Education and Science passed Royal Decree 1393/2007 (*BOE*, 2007) which regulated official higher education in Spain. This decree set the future for a series of new university degrees by recognizing the importance of the European educational policies generated following the *Bologna Declaration* (1999), dealt with in section 3.2 of this dissertation and referred to in the first paragraph of the decree. In this decree, the references to foreign languages are vague and yet, through the acceptance of the European policies and the Bologna process, it implicitly agrees upon the importance of foreign languages in higher education for transnational mobility of students and for their employability. In terms of mutual recognition, the decree proposes using ECTS credits.

At a different level, in November 18th 2010, following European regulations and recommendations, the CRUE► (*Conferencia de Rectores de las Universidades Españolas*, Spanish University Rectors' Conference) commissioned a report on language teaching and accreditation which was drafted in February 23rd 2010 and finally passed at the General Meeting held by the CRUE on September 8th, 2011 in Santander, Spain (*PAI*, 2011). This report, a type of unintended follow-up of Royal Decree 1393/2007 (*BOE*, 2007), pointed out that there existed considerable diversity in procedures and requirements for language recognition in Spain, and that this lack of homogeneity was leading to confusion. The report, entitled *Propuesta sobre la acreditación de idiomas* (*PAI*, 2011), also highlighted the fact that training and certification were not always differentiated in Spain, and wished that all universities integrated in the CRUE should issue language certificates which would be mutually recognized at both a national and an international level. To reach these conclusions, the CRUE took into account the experience of 50 Spanish universities and other educational institutions. At the same time, they agreed to work towards mutual recognition of language levels to access higher studies and to converge on accreditation mechanisms. For the latter purpose they established ACLES► (*Asociación de Centros de Lenguas en la Enseñanza Superior*, Association of Language Centers in Higher Education) as the standard of quality for language tests, and agreed to recognize other certifying institutions such as Cambridge, the Alliance Française and the Goethe Institut.

From this moment onwards, many regions in Spain have passed laws to meet the standards previously mentioned. Halback and Lázaro (2015), first published by the British Council in Spain in 2010 and updated in 2015, is the most up-to-date and comprehensive analysis of the impact of Spanish regional policies on higher education. This report gathers data from 50 Spanish universities and confirms that the coordination and homogenization of certification processes and mutual recognition has improved between 2010 and 2015. Likewise, the report looks forward to further clarification of standardization processes, clear

national linguistic policies, the implementation of quality standards and a more pro-active role on behalf of the Spanish central government.

3.3 Andalusian policies

Andalusia is 1 of the 17 autonomous communities that exist in Spain. Similar to the German federal *Länder* system, autonomous communities in Spain hold devolved powers over education. In practical terms this means that each community has exclusive competences in educational affairs, which leaves room for a great deal of heterogeneity in accreditation and certification. This variegation may eventually hamper mutual recognition among autonomous communities, particularly in the aforementioned desultory national context.

Over the last 15 years, Andalusia has also moved from a traditional method inherited from the teaching of classical dead languages (Greek and Latin), which chiefly relied on translation methods, on to a communicative approach that has increased the number of proficient speakers of different foreign languages in several orders of magnitude. In fact, Andalusia is living the *belle époque* of foreign languages assessment and teaching. Twenty years ago this would have been impossible, but now, proficiency examinations in Spain are accepted and looked forward to.

What are the reasons for this dramatic change? Weighted against outdated teaching traditions, the regional government of Andalusia, which has been in charge of the same political party since 1978, has decisively supported bilingualism at schools and high schools. In 2005 the regional government passed a three-year plan to promote plurilingualism in Andalusia (*BOJA*, 2005). Up to date, the plan has generated a network of 1335 bilingual schools among early childhood, primary and secondary education schools which use English, French and German as vehicles for content language integrated learning. Along with this, the number of publicly founded official language schools in which different languages are taught has increased to 51. The document sprang from all the

European and national recommendations (see sections 3.1 and 3.2) to provide Andalusian citizens with the necessary tools to live, travel, study and work in a knowledge-based society. Dissonant voices deemed the plan as a show for the gallery rather than as a real agenda. In fact, the plan had to struggle to convince policy makers and stakeholders alike. Mistakenly, policy makers and stakeholders believed that such a plan should be able to turn traditional schools into smooth-going plurilingual centers overnight. Reality, as usual, was stubborn and they both very soon realized that plurilingualism would take decades but that, as usual too, every journey needs a first step.

The plan was in a league of its own but at the same time it was supported by additional laws (*BOJA*, 2011a; 2011b and 2013) which extended the goals of the original plan, articulated the mechanisms through which syllabi should be designed and defined the competences of teachers and the means for non-plurilingual centers to join the network. Guides were created for those centers which wanted to join the plan and which necessitated guidance (Consejería de Educación, 2013).

The first generations of students who benefitted from the multilingualism plan at school arrived at Andalusian universities between 2012-2014. These new Andalusian university students are indeed better prepared in terms of languages than their counterparts were 15 years before, but this is not to say that homogeneity in levels has been achieved. When the students of this new generation of foreign-language Andalusian speakers enter university, they do it with different levels of proficiency, depending on the province or school they come from. The difference in levels is partly due to the fact that the plurilingualism plan has not penetrated homogeneously in all public schools and high-schools, and partly due to the fact that the high-school leaving exam that Andalusian students have to sit solely focuses on the skills of reading and writing. As a consequence, many teachers neglect listening and speaking skills during the last year of high-school. Their assumption is that, by focusing on reading and writing during the last year(s) of high school, they will boost their students' marks

in the aforementioned school leaving exam which, eventually, will determine the degree that university inbound students can access.

Even so, in Andalusia, current syllabi at initial stages of education promote communicative approaches to foreign language learning. As said before, it is only during the final years of secondary education that students and teachers are under the pressure school leaving exams and also pushed to boost the general success of their schools. If the school leaving exam that will determine the future of these students does not have a listening or a speaking component, why should they bother to practice these skills? This is an obvious contradiction which will have to be solved in the coming years. On the downside, practicality makes it difficult to assess listening and speaking in large-scale school leaving exams while reading and writing are easy to mark.

Yet again, the lack of public trust in the soundness of certification criteria in these school-leaving exams prior to higher education led Andalusian universities to create their own language proficiency tests in 2011. There is still a regulatory gap in this respect since no law has officially set, for example, the minimum level of proficiency that university students have to prove to enter or to leave university. In the absence of a more precise regulation, Andalusian universities have followed the recommendations of the DGU► (*Dirección General de Universidades*, General Board of Universities) of Andalucía which, in June 2010 distributed a circular among the rectors of all Andalusian universities (personal communication):

> *Por indicación de Dña. Mª Victoria Román González, Directora General de Universidades, le comunico que en relación con la remisión de los planes de estudio autorizados y verificados para la publicación en el* BOE*, se les recuerda que en todos ellos debe aparecer explícitamente la exigencia del nivel de idiomas que se haya acordado para la titulación (al menos el nivel B1).*

> *En el caso de que se haya producido la publicación de un plan de estudios que no contemple este requisito, deberá ordenarse la publicación de la correspondiente publicación de errores.*[5]

During 2010 and 2011 the different Andalusian universities created their own exams and issued their own certificates too. It was not long before the universities noticed that they would benefit from sharing resources in the production of these language proficiency tests. Since most of them had followed national ACLES standards, as indicated by the CRUE (see section 3.2), the exams were very similar among themselves, which made sharing a lot easier. Thus in 2011, the 9 public Andalusian universities (Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, Pablo de Olavide and Seville) signed an agreement, the *Convenio de Colaboración* (*CC* 2011) to define the standards which would regulate the recognition of foreign language levels in their autonomous community. The *CC* (*ibid*.) was drafted by the AGAE► (*Agencia Andaluza para la Acreditación*, Andalusian Agency for Accreditation), nowadays DEVA► (*Dirección de Evaluación y Acreditación*, Board of Evaluation and Accreditation) and by AAC► (*Agencia Andaluza del Conocimiento*, Andalusian Agency of Knowledge), the last of which would later set the road to certification at Andalusian universities (AAC, 2013). After the *CC* (2011) was signed, the 9 participating universities informally commissioned different representatives to set up a working group that would serve as the driving force effectively linking the directives contained in the *CC* (*ibid*.) and day-to-day practicality concerns at universities. The *CC* (*ibid*.) described different aspects regarding Andalusian exams:

[5] Following the indications from Mrs. Mª Victoria Román González, chair of the Board of Andalusian Universities, I inform you that, in relation to the submission of the syllabi which have been authorized and verified to be published in the *BOE*, all such syllabi must explicitly mention the requirement of the language level which is necessary to finish the degrees for which they were designed (at least B1). / In the event that a syllabus not containing such requisite has been already published, the corresponding erratum will have to be published (my own translation).

1. Objectives and general characteristics of the procedure
2. Characteristics of the language test: contents, structure and assessment criteria
3. Administration of the language tests
4. Procedure for the revision of exam marks
5. Certificates
6. Profile of the examiners
7. Revision, improvement and responsibilities in the procedure

The tests defined by the *CC* (2011) were skill-focused and their number of tasks, items and timing were specified as well. Although it was not explicitly mentioned in any document, the signatory institutions also agreed to have their tests undergo external audits to certify their quality. They decided to apply for 2 different external quality certifications, namely the one issued by DEVA at a regional level and the one issued by ACLES at a national level. Obtaining these 2 external quality certifications would be an accolade for the work developed and the confirmation that the Andalusian tests were on the right track. The Center for Higher Studies in Modern Languages (*Centro de Estudios Avanzados en Lenguas Modernas*, CEALM► henceforward) of the University of Jaén, which was commissioned with designing the tests in its university, was the last to be founded in Andalusia but the first to obtain both certifications. The CEALM became thus a benchmark in the region in terms of quality standards. The certification process established by DEVA (Marcelo *et al*., 2013) was particularly comprehensive and demanding, but has helped to develop as sound protocols as CEALM's *PADLE* (2015).

The *CC* (2011) was particularly important because, for the first time ever, the same regional model of test was defined in Andalusia, with the final goal of mutual recognition. The *CC* (*ibid*.) also included a list of other international certifications which would be recognized by the 9 public universities of Andalusia. Although the *CC* (*ibid*.) left many areas open to interpretation, thanks to the debate on how these aspects should be interpreted, it was possible to start

implementing and fine-tuning the original model of test. In 2013, 2 years after the *CC* (*ibid*.) was signed, a follow-up meeting was organized in Málaga, in which, for the first time, test developers were allowed to participate side by side with institutional representatives. The meeting proved to be a great opportunity to identify the vulnerabilities of the common specifications after 2 years of implementation. From the beginning, it was clear that the 9 signing universities shared certain problems, the foremost being lack of homogeneity in the design of the tests. Since each university had been designing their own tests, work was repeated in some languages (English), while others (German, Russian, French, etc.) were almost unattended since the demand for tests in such languages was much more reduced, along with the number of experts available. There was no centralized source of information and test developers received different messages in different ways. All this led to identify the appropriate path to follow very difficult. The autonomy of each university, which was recognized in the *CC* (*ibid*.), originated differences in the frequency with which tests had to be brought live, as well as in policies regarding the temporal validity of external certificates, exemption criteria for the handicapped, and differences in the criteria regarding the great variety of requests to recognize certificates which were not originally included in the agreement. The *CC* (*ibid*.) was a necessary leap forward but it still had to be tweaked.

The meeting of representatives and test developers in Málaga triggered the constitution of a board of experts (which later would become the Technical Advisory Committee), with at least 1 representative per university, who would ensure compliance with the agreement signed in 2011 through yearly follow-up meetings. The first of such meetings was held in Cádiz (October, 2013). Other meetings have followed in the university of Jaén (July, 2014), Málaga (January, 2015), Seville (May, 2015), Granada (October, 2015 and July, 2016) and Pablo de Olavide (October, 2016).

The frequency of these meetings is a clear example of the commitment of all members of the working group. Each one of the meetings has enabled a

follow-up of the implementation of *CC* (2011) and has also updated important questions such as the recognition and certification of Andalusian processes by ACLES and other independent quality-control bodies.

The board of experts had been working for almost 4 years when, in October 2015, the General Director of Universities in Andalusia summoned in Granada representatives from all Andalusian universities. The meeting was meant to be a follow up of all the actions carried out since the *CC* (2011) was signed. The different tests of the 9 Andalusian universities (whether certified by DEVA, ACLES or not) were discussed. In the meeting, the representatives of the different universities also discussed the funding policies that the Junta de Andalucía (the regional government) had designed to aid Higher Education students of languages in their way to obtain official certifications.

The representatives agreed to recommend the rectors of their corresponding universities to create a technical advisory committee that would take on the work that the board of experts had been developing since 2011. This Technical Advisory Committee would depend on the Follow-up Committee (an intermediate panel to be constituted) which in turn would be depend on vice-chancellors. In practical terms, a three-level decision system would be created in which vice-chancellors (at the top) would regulate linguistic policies based on the advice of the Follow-up Committee (intermediate) as informed by the Technical Advisory Committee (at the base of practicality concerns). The board of experts that had been working since 2011 formed the Technical Advisory Committee, on which language and testing experts served.

In the meeting of October 2015, a revision of equivalence tables for external tests was also suggested. These tables had been first introduced when the *CC* (2011) was signed and had been updated periodically by the board of experts ever since that moment. The tables listed those exams not developed by Andalusian universities which were also recognized (TOEFL, IELTS, the Cambridge suite of exams, Trinity, etc.). The prices of tests were revised in order to balance the existing differences between universities. The board of experts also

discussed in this meeting a proposal of the CRUE which suggested that the students of philology and translation degrees should not be asked to certify their level of competencies in a foreign language. The assumption was that the skills acquired through their degrees equal or surpass the B1 requirement.

All these discussions were recorded and turned into the agenda of a meeting celebrated in February 2016 by the Follow-up Committee. This meeting served to establish the Follow-up Committee officially, to draft updates to the *CC* (2011) and to update its equivalence tables too. Within the newly created structure, the proposals of the Follow-up Committee were then submitted to the vice-chancellors of Andalusian universities who met in March 2016 in the Academic Commission of the Andalusian Council of Universities (*Comisión Académica del Consejo Andaluz de Universidades*). In March 3rd 2016 the vice-chancellors of the Andalusian universities signed an amended version of the 2011 *CC* (*ibid*.) which already acknowledged the Follow-up Committee, recognized a revised version of the equivalence tables of external certificates accepted by Andalusian universities and, finally, revised 2 technicalities related to the characteristics of the exams designed, namely the need of double correction in the productive skills of the Andalusian tests and some details regarding the way in which the marks of the tests had to be released.

Since 2011, the consensus of these 9 Andalusian universities (Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, Pablo de Olavide and Seville) has facilitated mutual recognition of language certificates for the potential +25K students which these universities host in their language programs on a yearly basis. In practical terms, this means that a certificate issued by any Andalusian university is automatically recognized by any other Andalusian university. Out of the +25K yearly students that attend language courses at Andalusian universities, 13,721 sat at least 1 test developed in-house in the academic year 2013-2014, 8,731 did in 2014-2015 and 8,118 in 2015-2016.

Table 3.3.a displays the preferences of students of language at Andalusian universities. The table does not contain data about language certifications

obtained by students outside Andalusian universities. Generally speaking, Andalusian students can access 2 types of language tests. On the one hand, tests developed in-house at universities and, on the other, external exams of well-known brands such as Cambridge, ETS, British Council, Trinity, PLIDA, CELI, HSK, etc. which have signed different agreements with these institutions and which also offer sittings of their exams inside the universities. The importance and number of these external tests varies across Andalusian universities because, as we can see, while universities such as Huelva (UHU in table 3.3.a) rely heavily on external exams, other universities such as Málaga (UMA) or Seville (USE) do not offer any type of external exams. In general, both types of exams coexist.

As we see in the totals of table 3.3.a, once Andalusian students of language enroll in courses at university centers, they tend to choose tests developed in-house rather than the tests of external brands which are also offered inside the universities. These data are particularly relevant for the consequential validity (see sections 2.2.1 and 2.2.3) of our set of rubrics, since they have been developed with the objective of being implemented in in-house tests, which means that its yearly potential candidates are in the range of 8K to 13K candidates.

The decrease in the number of tests developed from 2013 to 2016 is mainly due to the fact that once students certify their degree of competence they do not need to do it again. University students frequently choose tests developed in-house against external ones because the prices of the former are more reduced than those of the latter. In fact, the signatory universities of the *CC* (2011) agreed to limit the maximum price of their tests in order to make them more accessible to all students. Besides this, students at some Andalusian universities do not have to pay for their first attempt at the test. Only if they fail, will they have to pay for a second sitting.

		2013-2014	2014-2015	2015-2016
UAL	Developed	475	222	630
	External	0	0	751
UCA	Developed	1128	1185	1375
	External	24	16	38
UCO	Developed	561	646	539
	External	112	250	681
UGR	Developed	1373	-	-
	External	158	-	-
UHU	Developed	0	98	81
	External	195	376	248
UJA	Developed	707	271	286
	External	56	205	179
UMA	Developed	1584	2428	2336
	External	0	0	0
UPO	Developed	1679	1586	1143
	External	0	0	0
USE	Developed	6214	2295	1728
	External	0	0	0
TOTAL	Developed	13721	8731	8118
	External	545	847	1897

Table 3.3.a. Tests at Andalusian universities (2013-2016)

Going back to the timeline of Andalusian policies, further convergence of the tests designed by the 9 Andalusian universities was enhanced through the design of a joint training course on language testing for test developers of the different universities, many of who were also part of the Technical Advisory Committees of their universities. The course, designed by international experts, regularly brought together 30 of the test designers of these universities who were trained in good practices for test development along 1 year. The contents of the course ranged from the design of specifications to task design or validation procedures through CTT and MTT. This new leap forward in training provided ample opportunities to re-evaluate the weaknesses and strengths of all the tests designed in Andalusia. As a consequence, mutual recognition of test results and certifications was strongly reinforced and re-assessed at 2 different levels: at an institutional level (Follow-up Committee) and at a practical level (Technical Advisory Committee). The Follow-up Committee effectively acts as a link between

the linguistic demands of the stakeholders (analyzed by the Technical Advisory Committee) at universities and the decisions of the different vice-chancellors, balancing the impact of language testing policies in the Autonomous Community, ensuring the practicality of initiatives and maintaining a high quality standard in the tests from the 9 Andalusian public universities.

Main Andalusian laws on plurilingualism	
2005	Law which passes the Plan for the Promotion of Plurilingualism in Andalusia (*BOJA*, 2005)
2011	Order which regulates bilingual teaching in schools of Andalusia (*BOJA*, 2011a)
2011	Order establishing the procedure to authorize bilingual teaching in private centers (*BOJA*, 2011b)
2013	Order that updates previous regulations on language policies (*BOJA*, 2013)
Milestones in the convengence of language proficiency tests of Andalusian universities	
2010	the Board of Andalusian Universities recommends that B1 should be aked in degree studies.
2011	Andalusian universities sign an agreement to design language proficiency tests (*CC* 2011).
2013	Institutional representatives and test designers first meet in Málaga to discuss the 2011 agreement.
2013	First meeting of the Technical Advisory Committee, in Cádiz
2014	Second meeting of the Technical Advisory Committee, in Jaén
2015	Third meeting of the Technical Advisory Committee, in Málaga
2015	Fourth meeting of the Technical Advisory Committee, in Seville
2015	Joint training course on language testing in Córdoba, Granada, Málaga and Seville
2015	Fifth meeting of the Technical Advisory Committee, in Granada
2016	Follow-up meeting organized by the General Director of Universities, in Granada
2016	Meeting of vice-chancellors and the Follow-up Committee, in Granada
2016	Vice-chancellors sign an amended version of the agreement signed in 2011
2016	Sixth meeting of the Technical Advisory Committee, in Granada
2016	Seventh meeting of the Technical Advisory Committee, in Seville (Pablo de Olavide)

Table 3.3.b. Andalusian timeline of regional laws and university meetings

On the question of mutual recognition in Andalusia, the regional government passed in 2015 (*BOJA*, 2015) a law to regulate the procedure through which language proficiency certificates should joint the list of officially recognized titles, in the same fashion in which the Andalusian universities have been updating their tables of external certificates. This regional move has taken place 4 years after the 9 public universities joined forces in 2011, which means that the regional government keeps an eye on the decisions adopted by these

universities. This move is important because it will allow Andalusian universities to include their tests in the list of regionally recognized tests in which, surprisingly enough, they have not been included thus far. It is certainly mystifying that the same government that funds tests of proven quality, such as those by Andalusian universities, does not officially recognize them. The aforementioned law is likely to amend this contradiction.

PART 2

THE EXPERIMENT

Over fifty of the T-14 android as he recalled had made their way by one means or another to Earth, and had not been detected for a period in some cases up to an entire year. But then the Voigt Empathy Test had been devised by the Pavlov Institute working in the Soviet Union. And no T-14 android –insofar, at least, as was known– had managed to pass that particular test.

Dick (1996:29)

CHAPTER 4. DESIGN OF A NEW SET OF RUBRICS

No test is perfect. Even the Voigt Empathy Test (Dick, 1996) had its flaws. One test must serve one particular purpose and is likely not to be useful if used in a context different from the one it was designed for. Language tests are not one exception. The test used to check the English of a group of prospective plane pilots should have some specifications different from those of university entry tests, as discussed in section 2.1.

Tests designed for broad audiences lose specificity. Tests designed for very specific purposes will not be valid for broad audiences. Tests must be designed taking into account what needs to be measured and so must be the assessment criteria linked to such tests. In chapter 4 we will explore the difficulties encountered when designing measurement tools, our rubrics, conceived to be used in 9 different public universities of Andalusia under different types of budgetary and political shortcomings.

In section 4.1 we will focus on the genesis of the project, on how the needs were detected and on how a first attempt to create a common model proved to be inoperative due to the variegation of factors to take into account. In section 4.2, we will describe a scalable protocol to design rubrics which has already proved to be valid. As mentioned in the introduction of this dissertation, this protocol is one of the most relevant contributions of the present work. The protocol provides test experts with a straightforward and usable tool to build rating scales based on the *CEFR* (Council of Europe, 2001). Within the protocol, we will describe how 11 different raters validated in 2 stages a newly created set of rubrics. Section 4.2.4 is particularly relevant because it provides an in-depth statistical validation of the outcomes of the protocol.

4.1 Revision of previous sets of Andalusian rubrics

In the context described in section 3.3, all Andalusian universities had been developing their own rubrics for proficiency tests. The tests served a common purpose across Andalusia and had the same specifications but they were marked through different rubrics. This was an evident problem which had to be solved. In 2011 each Andalusian university created its own rubrics for oral and written production. The contradiction was obvious because these universities had already started to share tasks but were marking them through different rubrics, none of which had been previously analytically validated.

This did not only jeopardize fairness but also the general validity of tests (see sections 2.2.1 and 2.2.3). At that point, the Technical Advisory Committee (see section 3.3) decided to unify rating criteria across universities. Since most of the exams designed by universities were B1 level, this was the level chosen to design the first set of rubrics. If it worked, then the design process could be extended to other levels below and above B1. The Committee also decided to set off by producing rubrics for the oral component of exams since, due to lack of resources and of expertise, it was not possible to tackle the creation of rubrics for speaking and writing at once. Again, if the design proved itself successful at speaking rubrics, it could be extended to writing rubrics *mutatis mutandis*. Since the majority of tests designed in Andalusia were English proficiency tests, that one was the language chosen to the description of the rubrics.

The most challenging aspect in the design and validation of the intended common set of rubrics was, without any doubt, creating a final tool that professionals from different universities accepted as their own. If a set of rubrics is designed without consulting those that will be using it in real tests, rubrics, which are intended to generate consensus, may become a major source of dissent. After the decision to create a new set of rubrics, we were commissioned by Technical Advisory Committee for the task.

To begin our work, we listened to the opinion of experts from the different universities. Rubrics are often a direct way to operationalize the construct of one test. This meant that if the created rubrics were able to contain the proposals of fellow colleagues and even bits of former rubrics used by them, the new yardstick would not be alien to them and it would not generate a negative reactions. After listening to other colleagues we decided to compile and contrast all the existing rubrics to identify the points they shared, their strengths and weaknesses. The assumption was that, after this initial analysis, it would be easy to develop a new set of rubrics taking into account everything learnt from the 9 Andalusian universities. Unfortunately, it was not that simple.

When we compiled the existing rubrics we noticed a high degree of difference among them. Though most of the pre-existing rubrics from Andalusian universities were analytic, some others were holistic, the number of band descriptors ranged from 5 to 10, the linguistic features assessed (adequacy, task achievement, language, pronunciation, etc.) were not always the same and eventually, 2 rubrics which shared 1 feature defined it in different ways. On top of all these differences, none of the pre-existing rubrics was explicitly linked to the *CEFR* (Council of Europe, 2001) although all of them had been developed with it in mind. Finally, and most importantly, none of them had been validated through analytic methods.

It was a major challenge to create a set of rubrics that could compensate for the gaps and which could also be perceived as the common denominator of all the pre-existing ones. To respond to the challenge we envisaged an 8-stage process that would yield (we hoped) the desired result:

Stages of the first attempt to design a common set of rubrics	
Stage 1	Comparison of pre-existing sets of rubrics from Andalusian universities
Stage 2	Definition of the type of scale: analytic vs. holistic
Stage 3	Identification of how many and which linguistic features should be included
Stage 4	Identification of the number of bands and scores to be included
Stage 5	Design of descriptors departing from the pre-existing ones
Stage 6	Qualityative validation
Stage 7	Quantitative validation
Stage 8	Implementation

Table 4.1.a. Stages of the first attempt to design a common set of rubrics

For stage 1 we compiled the B1 rubrics for oral proficiency tests of 8 of the 9 Andalusian universities: the universities of Almería, Cádiz, Córdoba, Granada, Jaén, Málaga, Pablo de Olavide and Seville. All the analyzed rubrics are available for download at <https://goo.gl/UIsLO7>. The University of Huelva had not developed any set of rubrics this far and they used external exams for the certification of their students. Out of the 8 compiled sets, only 7 were usable because the rubrics of the University of Granada were specific for multi-level exams.

The rubrics compiled were generically used across different languages (*i.e.* to assess French, Italian, English, etc.) except in the case of the University of Seville, which used different rubrics for different languages. In the case of the University of Seville, the rubrics used at stage 1 were the ones corresponding to the English area.

Most of the scales were analytic, *i.e.*, rubrics in which the rater assigns a score to each of the linguistic features being assessed in the task (Jonsson and Svingby, 2007:131-132). Two of the sets were holistic (Almería and Córdoba), *i.e.*, clearly aimed at producing overall judgments about the quality of the performance (*ibid.*). There were 2 other sets which included a complementary

table for holistic marks despite being analytic (Cádiz and Pablo de Olavide) and 1 included holistic appraisals as 1 individual linguistic feature (Málaga). Only the set from Córdoba was task-specific, although not consistently. The set from Granada was specific in the design of its outcome but general in the definition of its dimensions.

The most frequent number of bands was 6, which meant that most sets of rubrics distinguished 6 different scores for each linguistic feature measured. The distribution of the numbers of bands was irregular and ranged from 3 (Córdoba) to 11 (Pablo de Olavide, which used half points for bands). Besides, 6 of the 8 sets used numeric labels for the bands (generally from 0 to 6) while 1 used lexical labels (Córdoba: pass, merit, distinction) and another one (Seville) used a mix of numeric and lexical labels. The case of the set from Granada was particularly different because its descriptors were not graded in bands. Instead of graded bands, they displayed achievement dichotomous descriptors (*i.e.* test takers either reached the level of the descriptor or not). In some cases, some bands were described as sharing characteristics from upper and lower levels (Cádiz, Jaén, Málaga).

The set of Granada deserves particular attention. Besides the aforementioned lack of proper bands, it was the only set designed to cover 2 *CEFR* (Council of Europe, 2001) levels (B1 and B2). The set was designed to match the requirements of the exam designed at the University of Granada, which contains 3 different tasks, the first one designed for a B1 level, the second for an intermediate B1/B2 level and the third one for a B2 level. Table 4.1.b reflects the heterogeneity of the sets of rubrics analyzed.

University	Level	General vs. Specific	Analytic vs. Holistic	Number and type of bands	Number of linguistic features	Linguistic features Described
Almería	Unilevel	General	Holistic	5 numeric	4	Grammar and vocabulary Discourse management Pronunciation Interactive communication
Cádiz	Unilevel	General	Holistic Analytic	6 numeric	4 + 1	Pronunciation Fluency and coherence Grammar and vocabulary Interaction Holistic achievement
Córdoba	Unilevel	Specific	Analytic	3 lexical	2	General achievement Task fulfillment
Granada	Bilevel	Specific	Analytic	2 dichotomous	4	Oral fluency Coherence and cohesion Grammar (control and range) Vocabulary (control and range)
Jaén	Unilevel	General	Analytic	6 numeric	4	Grammar and vocabulary Discourse management Pronunciation Interactive communication
Málaga	Unilevel	General	Analytic	6 numeric	5	Grammar and vocabulary Pronunciation and fluency Communication Interaction Global achievement
Pablo Olavide	Unilevel	General	Holistic Analytic	11 numeric	4 + 1	Grammar and vocabulary Discourse management Pronunciation Interactive communication Global mark
Seville	Unilevel	General	Analytic	5 numeric and lexical	5	Interaction - task competion Fluency Pronunciation Accuracy Range of vocabulary - structures

Table 4.1.b. Analysis of pre-existing rubrics

The challenge was obvious, because although some of the sets analyzed were quite similar, let us say, in the dimensions measured (Cádiz, Jaén and Málaga), others were clearly in a league of their own (Córdoba or Granada). It was very difficult to find a common denominator that the participating universities and their test administrators were willing to share. Most consulted test administrators were used to their pre-existing rubrics and early on they proved to be reluctant to a new set that they did not perceive as operational or that they perceived as very different to their previous rubrics. They were not willing to change their system because, at this stage, they did not perceive the new rubrics

as a sound and useful tool which, at the same time, retained part of their previous system. They were pessimistic and thought that the new rubrics would require a time-consuming familiarization process.

Following the analysis displayed in figure 4.1.b there were several obvious decisions to make in the design of the new rubrics, and some others which were not so obvious. At stage 2, for example, it was clear that our rubrics would be analytic, since most of the pre-existing ones were so. Analytic scoring is somewhat less practical than holistic scoring because it takes more time to apply the criteria to each performance to be marked (see East, 2009:91), but they also provide more detailed information about a test taker's performance in different aspects and are for this reason preferred over holistic schemes by many writing specialists (see Weigle, 2002:115). Most raters would agree on the fact that the task of rating is approached in a more consistent way when using a shared set of analytical criteria (East, 2009:91).

Another easy decision was whether to make rubrics general or task-specific. Luckily too, the only task-specific sets of rubrics were those of Córdoba and Granada, which meant that we should design a general type of rubrics. When rubrics are designed with a general character, these can be used in a wider range of tasks. Rubrics which are too specific and designed for one specific task or test type are not useful outside the task and test they were designed for.

The not-so-clear decisions were chiefly related to linguistic features and they arose at stage 3. As table 4.1.b above shows, the rubrics did not always break down the same linguistic features. Finding the common denominator in the existing sets of rubrics was very complex because, for example, some of the universities merged into 1 category features that others analyzed separately. As one example, *Discourse management* in the sets of Almería, Jaén and Pablo de Olavide included aspects of *Coherence and cohesion* and *Fluency*. In turn, *Fluency and coherence* in the rubrics of Cádiz was both linked to *Fluency* and to *Coherence*. At this point, it was important to simplify the variety of linguistic features in order to create operational categories of linguistic features.

Besides the mix of features, the names given to some of them were misleading in many cases and there was no consistency across rubrics in this respect. In this way, for example, the linguistic features of *Grammar* (Granada), *Grammar and Vocabulary* (Almería, Cádiz, Jaén, Málaga and Pablo de Olavide) and *Range of Vocabulary and structures* (Seville) referred pretty much to the same concepts. To clarify how many linguistic features the rubrics had in real fact we analyzed the frequency with which categories of linguistic features appeared across the scales, as displayed in table 4.1.c below.

The grammar of candidates' speeches is still perceived by professionals as an accurate reflection of the knowledge that test takers have of the language, as we learn from its 87.5% prevalence. In the rubrics analyzed grammar is normally used to assess candidates' control over linguistic units such as words and phrases, sentences and clauses. As a result of this view, it is hand in glove with vocabulary and syntactic concerns, in fact, grammar and vocabulary are only differentiated in 1 set of rubrics (Granada). Range of vocabulary and structures (Seville) clearly refered to grammar too.

Pronunciation is on a par with *Grammar* in terms of importance (87.5%). It was present in 7 different sets. It was presented alone on 5 occasions (Almería, Cádiz, Jaén, Pablo de Olavide and Seville) and it was presented together with *Fluency* on 2 (Granada and Málaga).

Category	Linguistic features	Frequency (%)
Grammar	Grammar Grammar and vocabulary Range of vocabulary - structures	87.5
Pronunciation	Pronunciation Pronunciation and fluecy Oral fluency	87.5
Interaction	Interaction Interactive communication Interaction - task completion	75
Coherence and cohesion	Coherence and cohesion Fluency and coherence Communication Discourse management	75
Fluency	Fluency and coherence Oral fluency Pronunciation and fluency Communication Discourse management	75
Holistic assessment	Holistic achievement Global achievement Global mark	37.5
Task fulfillment	Task fulfillment Interaction - task completion	25
Vocabulary	Vocabulary	12.5
Accuracy	Accuracy	12.5

Table 4.1.c. Categories, linguistic features and frequency of pre-existing rubrics

Interaction was found in as many as 6 sets (Almería, Cádiz, Jaén, Málaga, Pablo de Olavide and Seville), which equals 75% of the rubrics analyzed. *Interaction* is, undoubtedly, 1 of the aspects perceived as most important in oral interaction in so far as it is an inherent part of most forms of oral communication. In all universities except in Seville candidates take oral interviews in groups of 2.

Coherence and cohesion was a particularly complex dimension. Globally speaking it was present in 75% of the rubrics analyzed. The concepts of *Coherence* and *Cohesion* only appeared together explicitly in 1 set of rubrics (Granada) and yet many others made reference to either *Coherence* or to *Cohesion* under different headings. In this sense, *Communication* (Málaga) and *Discourse management* (Almería, Jaén and Pablo de Olavide) were, despite their names, clearly referring to *Coherence*, *Cohesion* and to *Fluency* alike. This mix provided a blurred boundary between *Coherence and cohesion* and other dimensions which led us to consider the possibility of unifying this category with the next feature, *Fluency*.

Fluency was also present in many other sets (75%) disguised under different names. In some other cases, the references were explicit, as in the case of *Fluency and coherence* (Cádiz). But also, as we have just discussed, *Discourse management* (Almería and Jaén) as well as *Communication* (Málaga) included references to *Fluency*. Málaga was thus doubling references to *Fluency* since this dimension was also explicitly referred to in the individual linguistic feature of *Pronunciation and fluency*.

Cohesion and coherence and *Fluency* were at this point serious candidates to be merged into 1 category. The justification for this could have been based on 2 aspects. First, theoretically speaking, pragmatic competence can be understood as a function of discourse competences (cohesion and coherence among others), functional competences (where fluency is key) and design competences. This means that *Cohesion and coherence* and *Fluency* are rooted in common grounds (Council of Europe, 2001:123-130; North and Schneider, 1998:227;235). Second, their unification would have produced a more rater-friendly rubric in cognitive

terms thanks to a reduction of linguistic features and descriptors to be born in mind while rating. This approach is followed in tests such as IELTS (UCLES 2012:18) or the Cambridge suite of exams, where "[f]luency and coherence are captured under the Discourse Management criterion" (Khalifa and Ffrench, 2009:13). It will also be the approach followed in the final version of the rubrics presented in this dissertation, as shown in section 4.2.

Despite most sets being analytic, a number of them (37.5%) included *Holistic* appraisals, as we have already mentioned. In 2 cases (Cádiz and Pablo de Olavide) the *Holistic* approach was included as a separate grid of descriptors and only in 1 case (Málaga) was it embedded in the main set of rubrics. However, we felt that a *Holistic assessment* could not be interpreted as a linguistic feature since it is an intrinsic characteristic of the rubric itself, not one of its elements. As a consequence, using a holistic appraisal as a linguistic feature was easily ruled out.

Task fulfillment was only present in 2 sets (Córdoba and Seville) and thus it shows one of the lowest frequencies, 25%. This is consistent with the fact that most sets were general and not task-specific. It was ruled out not only because of its low prevalence but also because having used a task-specific design that could account for all types of oral tasks found across the 8 universities would have been impossible.

At the bottom of the classification we find the cases of *Vocabulary* as well as *Accuracy*, each present in only 1 set (Granada and Seville respectively). The descriptors of *Accuracy* could be matched to *Interaction* or to *Coherence* since they focused on the effectiveness of communication (*i.e.* "A few minor errors which do not impede communication", "Fairly frequent errors which do not prevent communication of the essential message" or "Large number of errors make utterances unintelligible"). This was a conceptual problem since the set which included *Accuracy* as a separate dimension (Seville) already included a specific dimension for *Interaction* and *Task completion*. There was no explicit indication to whether the errors mentioned in the descriptors were grammatical or

errors of any other type. Due to these interpretation problems, *Accuracy* was seen as a complementary feature of other categories.

At this point of stage 3, we decided to limit the number to 4 linguistic features, the most frequent number of features in the analyzed sets. Two of the sets clearly displayed 5 linguistic features (Málaga and Seville), 2 sets (Cádiz and Pablo de Olavide) had 4 features plus an additional holistic appraisal, but there were 3 sets that clearly displayed 4 dimensions (Almería, Granada and Jaén). Thus 4 dimensions was the logical choice.

As to which ones would these features be, the 5 most frequent ones, as shown in table 4.1.c above, were *Grammar*, *Pronunciation, Interaction*, *Coherence and cohesion* and *Fluency*. Since *Fluency* could be easily merged with *Coherence and cohesion*, the final decision was straightforward: by merging 2 of the 5 most frequent features we would be able to come up with 4 categories, namely *Grammar*, *Pronunciation*, *Interaction* and a mix of *Fluency* and *Coherence and cohesion*. This closed stage 3 of our first attempt (see table 4.1.a above).

In stage 4 the objective was to define the number of bands. Despite the frequency analysis (table 4.1.c above) yielded 6 as the most frequent number of bands in previous rubrics (followed by 5) a final decision was made to limit the number of bands to 5 to enhance rater cognition. The assumption was that in a 5-bands scale it would be easy to interpret band 3 as the minimum required level for a given linguistic feature. Band 3 being the pass, set in the middle of the rubric, the rubric is divided into 2 equal halves. In a 6-band set (as the initial frequency analysis suggested), the pass level would be somewhere between bands 3 and 4, which is less intuitive and thus more confusing in terms of rater cognition. This is not to say that a 6-band scheme would have been less reliable, but a reduced number of bands would also reduce the number of descriptors and the cognitive load necessary for raters. In short, the fewer bands, the easier it is for raters to use rubrics.

Having decided the number of bands to use at stage 4, we were ready to move on to stage 5. Stage 5 was at the core of the process because in this stage we would have to fine-tune the pre-existing descriptors, eliminate those which were redundant and we would have to fill in the existing gaps. Stage 5 was dedicated to the creation of the very essence of the rubrics: intelligible, operational and valid descriptors that raters could use swiftly and reliably while rating oral performances.

The first analyses of pre-existing descriptors for stage 5 were far from promising. Unfortunately, we noticed that our analysis had been flawed from the beginning. The pre-existing descriptors did not hold a 1 to 1 comparison. The descriptors from the different universities were not just completely different among them, they also had the problem that they had never been previously validated quantitatively. These had neither been linked to the *CEFR* (Council of Europe, 2001). Because of this, retaining them at all costs did not seem a good idea. Yes, retaining them would have made the new rubric more familiar to raters (our original intention), but at a tremendous cost, at the cost of reliability.

To make de decision we opened an opinion poll and asked 10 of the experts in the Technical Advisory Committee. They, together with other fellow workers, would have to use at some point the new rubrics and, as team leaders, they would have to introduce the new tool to the rest of the raters in their universities. It was paramount to reach consensus at this point. After the survey, 3 experts voted to refurbish the pre-existing descriptors and 7 voted to bin them and start from scratch.

This way, after 5 months of work, we failed to proceed to stage 5 and decided to make a fresh start. Sometimes, a timely retreat is a victory. We had learnt our lesson and time had not been completely wasted. Now we were more conscious than ever of the variegation and limitations of our rubrics. We had also gained awareness on which linguistic features were more important for the teams of raters involved, we had tentatively defined some of these linguistic features and

we knew which was the most operational number of bands to include. Perhaps, we hoped, we should be able to use this knowledge in the future.

4.2 Development of a protocol to design rubrics

Indeed, we had learnt our lesson. When we decided to continue with the design of the rubrics we believed that it was a good idea to revise the literature again to find out if there existed any type of protocol for such purpose. This time we were not only looking for generic directions to create rubrics. We were looking for literature that described how the problem of variegation in descriptors could be solved. We found nothing about this because, generally speaking, rubrics are analyzed in the literature as improvements of previous sets or as sets of descriptors created from scratch. We opted for a fresh start and began to think that it would be interesting to see how the *CEFR* (Council of Europe, 2001) descriptors could be used for the development of rubrics. Again, the literature is not conclusive in this respect and that was the moment in which we envisaged the creation of a protocol that did not only establish the different steps to follow in the design of rubrics, but which could also use the *CEFR* (*ibid*.) for such purpose.

One might expect that, due to the importance that rubrics have in contemporary testing, there should be extensive literature about the way in which they must be designed, validated and implemented. Nevertheless, "there is surprisingly little information on how commonly used rating scales are constructed" (Knoch, 2009:42). Brindley (1998:117) points out that "it is often difficult to find elicit information on how the descriptors used in some high profile rating scales were arrived at". McNamara (1996:182) wrote about the descriptors that make up rubrics the following:

> Given the crucial role that they play in performance assessments, one would expect that this subject would have been intensively researched and discussed. In fact, certainly in the field of language assessment, we find that this is not so; we are frequently simply presented with rating scales as products for consumption and are told too little of their provenance and of their rationale.

Turner (2000:556) makes the same point:

> With the important role that rating scales play in performance evaluation, one would think that the literature would abound with descriptions and procedures for scale construction. But, as we quickly learn, this is not the case.

Generally speaking, this is still true, which is striking since, for the sake of clarity and fairness, many rating scales of high stakes tests are unveiled to candidates but with no reference whatsoever to the way in which they were developed. We wanted our protocol to bridge this gap.

The design of rubrics for writing is more frequent in the literature but, with minimal changes, their insight can be extrapolated to the design of speaking rubrics. To build our protocol we basically took into account Weigle (2002), Knoch (2009), Dean (2012) and, indirectly Turner (2000), Ffrench (2003), Council of Europe (2009) and section 5.1.3 in the *CEFR* (Council of Europe, 2001).

Weigle (2002:109) points that the first decision to make when designing rubrics is to establish the type of rubric that we want to build. She gives the choice of primary trait, holistic or analytic rubrics (*ibid*.) although, as we saw in section 2.1.3, the choice is slightly more complex than this. We chose to develop an analytic scale because these are more reliable (Davies *et al*., 1999:75) and because this was the tendency in most Andalusian universities as we saw in table 4.1.b above.

After this decision, Weigle (2002:122-125) proposes to consider the following ones:

- Who is going to use the rubric (test designers, raters, stakeholders or a combination of them)
- What linguistic features will be assessed by the rubric
- How many bands it will have
- How scores will be reported.

Knoch (2009:38) extends this list by pointing to the need of describing the linguistic features included and by underlining the importance of quantitative validation (*ibid*.:79-100; 193-287). Weigle (2002:134-136) suggests inter- and intra-rater reliability checks to validate the rubrics. Dean (2012:18) only mentions qualitative methods of validation. In terms of validation procedures, the most comprehensive method is that proposed by Knoch (2009), since it uses both CTT and MTT. Knoch's (2009) is, indeed, a very comprehensive method of validation but, in our opinion, it is so comprehensive and complex that it is not practical at early stages of development. There is an additional aspect which is not mentioned by either Weigle (2002), Knoch (2009) or Dean (2012) but which is paramount in the European context of these rubrics. As mentioned above, this aspect is the linkage of the rubrics to the *CEFR* (Council of Europe, 2001), which constitutes the most important contribution of the design protocol that we propose:

Stages		Actions to be carried out
Stage 1	**Previous considerations**	1.a Decide if it will be primary trait, holistic or analytic. 1.b Identify and briefly describe the linguistic features that it will assess. 1.c Decide the number of bands. 1.d Consider the way in which scores will be reported.
Stage 2	**Write the descriptors**	2.a Select the *CEFR* tables that contain relevant descriptors. 2.b Distribute the *CEFR* descriptors as anchor descriptors. 2.c Fill in intermediate and incomplete bands.
Stage 3	**Validation 1 (qualitative)**	3.a Validate the rubric qualitatively by consulting other experts. 3.b Fine-tune the rubric following feedback from 3.a.
Stage 4	**Validation 2 (quantitative)**	4.a Analyze through Facets the scores that 2 raters give to 30 candidates. 4.b Fine-tune the rubric following feedback from 4.a. 4.c Analyze through Facets the scores that 5-8 raters give to 40-50 candidates. 4.d Fine-tune the rubric following feedback from 4.c.
Stage 5	**Implementation**	5.a Conduct rater training and benchmarking sessions. 5.b Use the rubrics in real exam conditions. 5.c Collate data from live administration and draw conclussions.
Stage 6	**Revision**	6.a Set up a cycle of revision. 6.b Collate data from different live administration and draw conclussions. 6.c Fine-tune the rubric if necessary and repeat stage 5.

Table 4.2. Protocol to design a *CEFR*-linked proficiency rating scale

In our opinion, this protocol offers various advantages when compared with other procedures. First and foremost, it describes how to create descriptors which are directly extracted from the *CEFR* (Council of Europe, 2001), which is something that, to our knowledge, has never been done before. On the other hand, it balances validation processes in such a way that it makes them reliable and practical for professionals with little background on statistics. It is neither too complex so that it requires advanced training nor too simplistic so that the final results are not reliable.

Let us see now how this protocol proved to be successful in the development of a rating scale. For this, we will describe its 6 stages and will illustrate them with real examples.

4.2.1 Stage 1. Previous considerations

As already mentioned, some of these previous considerations were already defined thanks to the failed attempt described in section 4.1, mostly those regarding stage 1.a of the protocol. We knew that our scale would have to be analytic to be consistent with the preferences showed by raters and due to reliability concerns.

At stage 1.b we also had an approximate idea of the linguistic features that we wanted our rubrics to have. As shown in the protocol (table 4.2.a above), it is important not only to select the linguistic features, but also to briefly describe them (see table 4.2.1 below) through sub-features so that raters have a clear idea of what they are going to find in each of them. Ideally they should gain this knowledge from standardization sessions and the like. Even so, according to our experience, having a brief description of what each feature describes is very useful.

From table 4.1.c we learnt that the most popular features among raters in Andalusian universities were *Grammar* (which appeared in 87.5% analyzed rubrics), *Pronunciation* (87.5%), *Interaction* (75%), *Cohesion and coherence* (75%) and *Fluency* (75%). We decided to merge *Grammar* with *Vocabulary*

(whose presence was residual in the rubrics (12.5%) but which went hand in glove with the former) to account for the maximum number of aspects.

For this we created the linguistic feature of *Language*, which would be defined through the sub-features *Vocabulary* (range and control), *Grammar* (range and control) and the newly created *Errors*. We decided to include *Errors* in *Language* because they are likely to occur at B1 and they may be a good indicator of the proficiency of candidates.

Secondly, we included *Pronunciation*. We defined this feature as containing the sub-features of *General pronunciation*, *Articulation of sounds* and *Prosody* (stress, rhythm and intonation).

The third linguistic feature would be *Interaction*, which would be defined through the sub-features of *Information exchange*, *Initiate, maintain and end a conversation*, and *Cooperation*.

Finally, as already mentioned, we created a mixed linguistic feature labeled *Discourse* defined by the 3 sub-features *Cohesion*, *Thematic development* (which we understood as a synonym of coherence) and *Fluency*. As anticipated in 4.1, this approach is followed in tests such as IELTS (UCLES 2012:18) or the Cambridge suite of exams, where "[f]luency and coherence are captured under the Discourse Management criterion" (Khalifa and Ffrench, 2009:13). The *CEFR* (Council of Europe, 2001:123-130) also groups *Coherence and cohesion* and *Fluency* as instances of pragmatic competences, which means that they are close with each other.

Summarizing, we identified 4 linguistic features that were in turn briefly defined by the sub-features that built them up. The number of sub-features was limited to 3 with the objective of incorporating 1 descriptor for each sub-feature in the final version of the rubric. The rationale was that if each linguistic feature was built by different components (sub-features), all such components had to be defined (at least) through 1 descriptor across the different bands of the rubric. It is frequent to find rubrics that use descriptors at some levels which later disappear

in the rubric at higher or lower bands. In this sense, for example, as raters, if we understand that the feature of *Language* is built by *Vocabulary*, *Grammar* and *Errors*, we expect to find graded descriptors of these 3 components across all bands. Too many components (sub-features) would have made a greater number of descriptors necessary and this would have made the rubric more difficult to use. All this information is displayed in table 4.2.1:

Main linguistic feature	Sub-features
Language	Vocabulary (range and control) Grammar (range and control) Errors
Pronunciation	General pronunciation Articulation of sounds Prosody (stress, rhythm, intonation)
Interaction	Information exchange initiate, maintain and end a conversation Cooperation
Discourse	Cohesion Thematic development Fluency

Table 4.2.1 Main linguistic features and sub-features of our set of rubrics

As regards section 1.c from the protocol (see table 4.2.a), related to bands, the *CEFR* (Council of Europe, 2001:181) points that there are basically 3 ways in which descriptors can be presented for use as assessment criteria. They can be presented as a scale, as a checklist or as a grid of selected categories. That is precisely what we did (see table 4.2.1 above). According to the *CEFR* (*ibid*.:181-

182) we can arrange descriptors in sub-scales (our linguistic features) to build rating scales

> by selecting or defining a descriptor for each relevant category which describes the desired pass standard or norm for a particular module or examination for that category. That descriptor is the named 'Pass' or '3' and the scale is norm-referenced around that standard (a very weak performance = '1', an excellent performance = '5'). The formulation of '1' and '5' might be other descriptors drawn or adapted from the adjacent levels on the scale from the appropriate section of Chapter 5 or the descriptors may be formulated in relation to the wording of the descriptor defined as '3'.

This is exactly what we did in the development of our new rubric and this is what we suggest to do if the protocol here presented is to be implemented. Using 5 bands is also in agreement with what was customary in most Andalusian rubrics, as we say in section 4.1. As we will also see in section 4.2.2, for the formulation of '1' and '5' we followed the recommendations of the *CEFR* (Council of Europe, 2001) and used the corresponding descriptors from chapters 4 and 5 (*ibid*.:43-130), both of which "present a fairly detailed scheme of categories for the description of language use and the language user". But this will be dealt with in more detail later on. For the time being, according to the protocol, the only important decision was to establish 5 as the number of bands that would be used.

Finally, we considered the way in which scores would be reported (section 1.d of the protocol). Since there might be heterogeneity among the 9 universities, we concluded that the numerical distribution of descriptors would be appropriate for candidates to identify the description of their performance and for raters to convert the scores in a variety of forms. The rubrics would have to be made public and if, for example, one candidate was marked 3-3-4-4 (band 3 for *Language*, band 3 for *Pronunciation*, band 4 for *Interaction* and band 4 for *Discourse*), it would be easy for him or her to check the rubrics and link the numbers with descriptions of performance. At the same time, raters could transform these figures in a variety of ways. Since marks are not weighted in any Andalusian university, if raters should want to report the overall score from the

candidate, in a degree from 1 to 10, just to give one example, this could be done through a simple rule of 3. If the candidate that obtains full marks (5-5-5-5) is marked with 10, then our imaginary candidate above (3-3-4-4) would obtain a mark of 7:

$$\frac{20}{(3+3+4+4)} = \frac{10}{x}$$

4.2.2 Stage 2. Writing the descriptors

Without any doubt, this step is critical in the design of a rubric. Descriptors represent the construct of examinations in a way in which no other aspect of a test does. They are an explicit representation of the sense that test designers, scale developers and raters make of the test not only for their meaning, but also for the way in which they are organized.

At the same time, when we write our descriptors and our rubrics, we must take into account that

> (i)f the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. At any rate, to the extent that the present method of scale construction is affected by the opinions of the readers who help to sort out the original statements into a scale, to that extent the validity or universality of the scale may be challenged.
>
> Thurstone (1928:547-548)

Then, how can we reconcile both ideas, the idea of our rubrics being a representation of test designers view of language and the idea that the rubrics "should not be affected by the opinions of the people who help to construct it"? The answer is that, once the language construct of rubrics designers is integrated in the rubric and proved to be qualitatively valid, it must become a solid measurement tool whose interpretation is clear and not subject to the opinion of

any individual rater. The idea that one set of rubrics can be consistently applied to different candidates by different raters with no change in the final result refers back to the underlying property of specific objectivity of MTT and Rasch models already mentioned in section 2.4.2, which is another argument in favor of the MTT validation process that we have followed in section 4.2.4.

To minimize subjectivity in the definition of our set of rubrics, we used the *CEFR* (Council of Europe, 2001) as an anchoring and authoritative pool of descriptors without having any initial idea of how these would work qualitatively. If we could not recycle the descriptors from the pre-existing Andalusian rubrics, perhaps we could use another set of descriptors which is familiar to most raters in Andalusia, case in point, the *CEFR* (*ibid*.) descriptors.

We have already seen in 4.2.1 what the *CEFR* (Council of Europe, 2001:181-182) suggests about the design of rubrics. Basically, the *CEFR* (*ibid*.) suggests to pick a descriptor, put it at the middle of the scale (at '3' in the case of a 5-band set) and construct the rest from this starting point. The "formulation of '1' and '5' might be other descriptors drawn or adapted from the adjacent levels" (*ibid*.:181). In this way, for example, in a B1 scale we would allocate the *CEFR* (*ibid*.) B1 descriptors at band 3, the *CEFR* (*ibid*.) A2 descriptors at band 1 and the *CEFR* (*ibid*.) B2 descriptors at band 5. Bands 2 and 4 would have to "be formulated in relation to the wording of the descriptor defined as '3'" (*ibid*.:182).

For some experts, A2 or B2 descriptors should never be used in a B1 scale. There are certain rubric developers in Spain and other parts of the world who claim that a B1 scale should be constructed departing solely from B1 descriptors. In short, these scale developers claim that when designing an analytic scale based on the *CEFR* (Council of Europe, 2001), anchoring descriptors should be taken solely from the *CEFR* (*ibid*.) level which the scale is going to measure. Such scale developers are quite opinionated about methods that either do not match the specifications of the *CEFR* (*ibid*.) or have not proved to be more reliable (or even reliable at all) than the method defined by the *CEFR* (*ibid*.) itself.

Following the *CEFR* (Council of Europe, 2001), we constructed our analytic scale using descriptors from different *CEFR* (*ibid*.) levels (type A in figure 4.2.2.a below). The scales built solely with descriptors from the level which they assess, mentioned in the previous paragraph, may feature 2 different construction plans. They either locate the *CEFR* (*ibid*.) B1 descriptors at band 3 and define *ex novo* bands 1, 2, 4 and 5 (type B in figure 4.2.2.a below) or place the *CEFR* (*ibid*.) B1 descriptors at band 5 and formulate descriptors for bands 1, 2, 3 and 4 *ex novo* (type C in figure 4.2.2.a below). Type B structures are not completely opposed to what the *CEFR* (*ibid*.) suggests, but they require the formulation of an unnecessary number of descriptors. Type C structures have simply no correlation with the directions provided by the *CEFR* (*ibid*.).

Languages are not learnt throughout exactly delimited stages. Languages are learnt in a continuum and, from this perspective, it makes sense to think that a very good B1 performer is close to B2. That is the rationale behind the proposal of the *CEFR* (Council of Europe, 2001), the one that we have used. We found no reason to do anything different from what is recommended by the *CEFR* (*ibid*.) since, to our knowledge, there is no scientific proof that not using adjacent descriptors is more reliable than doing so.

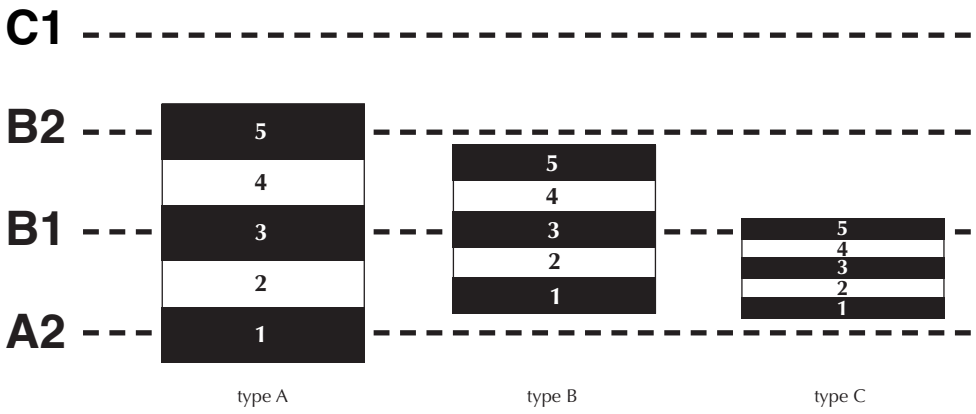

Figure 4.2.2.a. Different approaches to build *CEFR*-linked rubrics

Type A is the structure recommended by the *CEFR* (Council of Europe, 2001:181-182), by ACLES (see section 4.3.3 in ACLES, 2016) and, consequently, the one that we have used. It takes advantage of existing descriptors and uses B1

descriptors to formulate band 3, A2 descriptors to formulate band 1 and B2 descriptors to formulate band 5. An excellent B1 performer who scores band 5 would be close to B2, but could not be awarded a B2 mark if the test and the rubric have been validated exclusively for B1.

Type B is essentially the same structure as type A with the exception that it does not take advantage of pre-existing descriptors. Then, while in type A only 2 bands of descriptors must be defined *ex novo* (bands 2 and 4), in type B we would have to formulate 4 new bands (1, 2, 4 and 5).

Type C is the most controversial. According to this typology, an excellent B1 scorer (band 5) is not necessarily close to level B2 descriptors, which fall completely outside the rubric. In our opinion, this typology breaks the concept of continuity through learning stages which the *CEFR* (Council of Europe, 2001) postulates. Further shortcomings of type-C rating scales are the fact that they require the formulation of 4 bands (against 2 in type A) and that they solely offer 1 linkage point with the *CEFR* (*ibid*.) levels (against 3 in type A). In our opinion, type C is not only contrary to the recommendations of the *CEFR* (*ibid*.), but also far more complex to develop than type A. On top of this, there is no empirical evidence that type-C rubrics are more reliable than type A. Set against this, we have proved type A scales to be reliable, as we will show in sections 4.2.3 and 4.2.4. Type-C rubrics have become relatively frequent across Spain, as it was drawn from the workshops for the development of rubrics that EALTA hosted during their 13th Annual Conference in Valencia in June 2016. Representatives from all over the world attended the workshop, many of which belonged to Spanish institutions within ACLES. During the conversations at the workshop, the great majority of ACLES members acknowledged that, in the absence of a best way to anchor their rubrics to the *CEFR* (*ibid*.), most of them also extracted descriptors from it and placed them in their rubrics according to the 3 typologies above but with no particular criterion or empirical support. In other words, the rubrics of these institutions draw from the *CEFR* (*ibid*.), but due to the lack of a widely accepted system, they follow no common design pattern nor they are, in

most cases, quantitatively validated. The protocol proposed in this dissertation can be a valuable tool for all such institutions, whose characteristics are very similar to those of Andalusian universities.

Another advantage of type A rubrics is their scalability. As we can see in figure 4.2.2.b below, a part of the descriptors of one rubric can be used in the design of the next one. In our case, for example, descriptors in band 3 from the B1 rubric will be the same or very similar to descriptors in band 5 of the A2 rubric. Descriptors in band 1 from the B1 rubric will be at band 3 of the A2 rubric. The same applies for higher levels and descriptors. The descriptors in band 3 of the B1 scale will be at band 1 in the B2 scale, etc.

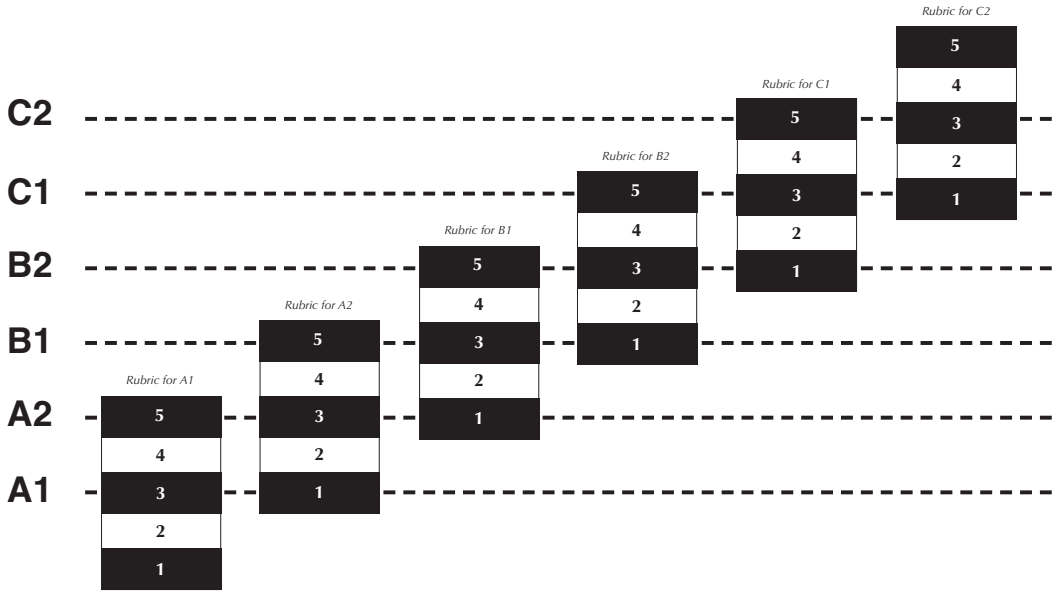

Figure 4.2.2.b. Scalability of the design model proposed in the protocol

This scalability allows saving a tremendous amount of effort since only intermediate bands (2 and 4) must be formulated, and even these can be scaled. More strictly speaking, the only descriptors that would have to be fully formulated are those corresponding to bands 1 and 2 in rubric A1 and bands 4 and 5 in rubric C2, the only ones for which the *CEFR* (Council of Europe, 2001) gives no formulation whatsoever. Let us remember that, occasionally, the *CEFR* (*ibid*.) describes 2 sub-levels in 1 band for different stages of proficiency. That is the

reason why some intermediate bands (2 and 4) from some of the levels are likely to be already formulated in the *CEFR* (*ibid*.).

After this brief justification of the scaffolding system used to build our rubrics, the next step of the protocol is 2.a, *i.e.*, to select the *CEFR* (Council of Europe, 2001) tables that contain relevant descriptors. In our case, we had to choose those *CEFR* (*ibid*.) tables that contained relevant descriptors for the assessment of the linguistic features of oral production described in table 4.2.1. The tables used from chapters 4 and 5 in the *CEFR* (*ibid*.) were:

- Tables from Chapter 4
    - Overall oral production
    - Sustained Monologue: Describing Experience
    - Sustained Monologue: Putting a Case
    - Public Announcements
    - Addressing Audiences
    - Plan
    - Compensating
    - Monitoring and Repair
    - Overall Spoken Interaction
    - Understanding a Native Speaker Interlocutor
    - Conversation
    - Informal discussion with friends
    - Formal discussion and meetings
    - Goal-oriented co-operation
    - Information exchange
    - Interviewing and being interviewed
    - Taking the floor
    - Cooperating
    - Asking for clarification

- Tables from Chapter 5
    - General linguistic range
    - Vocabulary range
    - Vocabulary control
    - Grammatical accuracy
    - Phonological control
    - Sociolinguistic appropriateness
    - Flexibility
    - Turntaking
    - Thematic development
    - Coherence and cohesion
    - Spoken fluency
    - Propositional precision

This amounts to a total of 31 tables, 19 from chapter 4 (all whose descriptors were "can do" statements) and 12 tables from chapter 5 (whose descriptors are formulated as "can", "has" and "shows" statements or as simply describing the aspect of language they refer to). Not all the tables contained in the *CEFR* (Council of Europe, 2001) were used since not all of them contained descriptors relevant to the linguistic features of table 4.2.1.

During the collection of descriptors from these 31 tables, each one was identified with the initials of the table it was taken from, to track its origin more easily. Obviously, some tables contributed more descriptors than others, but since this could not be predicted at the beginning of 2.a, the decision was to go through all the 31 tables comprehensively. Irrelevant descriptors would be ignored. We noticed that the different levels and sublevels are not evenly defined by the *CEFR* (Council of Europe, 2001). Some tables, such as for example *Sustained monologue: describing experience*, provide definitions for most levels and sub levels, while others, such as *Sustained monologue: putting a case* lack most of them, even the ones corresponding to C1 or C2. The reason for this is that the

CEFR (*ibid*.) identifies performances at these levels with performances at lower ones.

The next stage in the protocol, 2.b., was not easy because although some of the descriptors clearly matched 1 of the 4 linguistic features in table 4.2.1, some others did not. Sometimes, 1 descriptor could be placed in different linguistic categories. From the table *Informal discussion*, for example, the descriptor "Can take an active part in informal discussion in familiar contexts, commenting, putting point of view clearly, evaluating alternative proposals and making and responding to hypotheses" could partially match criteria of *Interaction* ("discussion", "proposals", "responding") and criteria of *Discourse* ("clearly", "making and responding hypotheses"). More often than not, since there was a wide variety of descriptors to choose from, these ambivalent formulations did not make it to the final version of the rubric. Other descriptors were duplicated. For example, A2 descriptor "can be made to understand, if the speaker can take the trouble" appears both as an individual sentence in *Understanding a native speaker interlocutor* and as a subordinate sentence completing the meaning of a previous descriptor in the table *Conversation* and *Goal-oriented Co-operation*. The same goes for the B1 descriptor "Can enter unprepared into conversations on familiar topics", which appears in *Conversation* and in *Overall spoken interaction*. Some other times, the choice of one descriptor would yield different results depending on the candidate. In the descriptor below, from table *Public announcements*, the reference to "his/her field" would have different meanings for an engineer, for an air-traffic controller or for an undergraduate student:

> Can deliver short, rehearsed announcements on a topic pertinent to everyday occurrences in his/her filed which, despite possibly very foreign stress and intonation, are nevertheless clearly intelligible.
>
> Council of Europe (2001:60)

By putting each descriptor in the same band of the *CEFR* (Council of Europe, 2001:181-182) it was taken from, our scale was automatically linked to the *CEFR* (*ibid*.) as indicated in figure 4.2.2.b.

After this compilation was made, it was clear that the category most extensively defined was *Interaction*, as most of the descriptors referred to it directly or indirectly regardless the table they were extracted from. The next most extensively defined categories were *Language* (in which, let us remember, we merged vocabulary and grammar) and *Fluency*. On the downside, *Pronunciation* and *Cohesion and Coherence* were scarcely referred to across the *CEFR* (Council of Europe, 2001) descriptors. The amount of descriptors available for *Interaction* and *Fluency* actually mirror the communicative construct of language which underlies the *CEFR* (*ibid*.). On the other hand, the lack of descriptors for *Pronunciation* is a caveat of the *CEFR* (*ibid*.) which is being currently tackled by Brian North himself, who is creating a whole new set of descriptors for *Phonology* and *Pronunciation* (plus others for *Mediation*) which will see the light soon. We had the opportunity to collaborate in phase 2 of North's project to extend the *Phonology* descriptors in the *CEFR* (*ibid*.) and this gave us the clue as to which direction should be follow when defining *Pronunciation*.

Step 2.b in the protocol is pretty much like making a jigsaw puzzle in which the *CEFR* (Council of Europe, 2001) descriptors are pieces that must fit together. The position of each piece in the puzzle is determined by its level and by the linguistic feature it describes. Level and linguistic features are the coordinates that arrange the pieces.

Unfortunately, the *CEFR* (Council of Europe, 2001) puzzle of descriptors has many missing and repeated pieces. The missing pieces are of 2 types. Firstly, most of the missing pieces (but not all) belong to bands 2 and 4 in our scalable scheme, which are intermediate levels. Secondly, there are some missing pieces for a few 1, 3 and 5 bands of our design. The repeated pieces are descriptors that appear 2 or more times across tables.

The final result is an irregular puzzle, more finished in some areas than in others. This is definitely not what one might expect after fitting together more than 250 pieces (our descriptors) at stage 2.b. Due to size constraints, the first draft of our puzzle cannot be reproduced here, but it can be downloaded from <https://goo.gl/Gxa7xP>. We believe it is a taster of the complexity of the puzzle that we had to organize. To have a clearer picture of which areas were more defined and where the missing pieces were, for the next phase of stage 2 we had to use a professional design tool, the software package QuarkXpress (QuarkXPress, 2016). At this stage, all descriptors were labeled with the initials of the table they were taken from to facilitate their traceability.

In stage 2.c, QuarkXpress (QuarkXPress, 2016) (which allows the use of very big word-processing layouts) provided us with a big canvas to work on, as big as 594mm high and 1000mm width (printing A2 size). This yielded the first draft of the rubrics (see the link in the previous paragraph), which was not yet operational but which gave a clear view of which areas had an excess of descriptors and which others had but a few or none. To fill in intermediate and incomplete bands is a key part of our design protocol. While we can remain confident of the reliability of anchor descriptors (*i.e.* those placed in the exact level in which they appear in the original tables), we are walking on thin ice when we define missing descriptors according to adjacent levels (Council of Europe, 2001:181-182). Drawing and defining the correct wording will determine the success of our scale.

After discarding repeated and irrelevant descriptors, we reduced each coordinate of the puzzle to 3 points or definitions. *Language*, for example, was defined in terms of 1) vocabulary range and control, 2) grammar and 3) errors. For the definitions of the remaining linguistic features see table 4.2.1 above and the Appendix. The idea was to define each of the 4 linguistic features according to 3 parameters that could be tracked by raters across bands. This meant that if, for example, *Language* was defined according to *Vocabulary* (range and control), the rubric should contain graded descriptors of *Vocabulary* (range and control) from

bands 1 to 5. The same applies to all other linguistic features. We wanted to provide the rubrics with homogeneity and with a clear display to make it rater-friendly.

There was an additional challenge imposed by the fact that the 3 descriptors linked to these definitions should remain under 65 words. Descriptors were limited to 65 words not only to make them usable in terms of rater cognition, but also because we wanted to fit them in the most common screen resolutions of mobile devices to enable their use through a specific mobile application, which was already under development by that time and that will be referred to in chapter 5.

Summarizing, in our case, stage 2.c involved the identification of the 3 points through which linguistic features would be defined, the elimination of redundant and irrelevant descriptors, the design of new descriptors drawn from the ones that remained after the elimination and, finally the shortening of descriptors to keep them under 65 words per box, without compromising the reliability of the set of rubrics. Easier said than done.

After some months of work and consultation with fellow colleagues we came up with the first draft of our new set of rubrics.

4.2.3 Stage 3. Validation 1 (qualitative)

It is paramount for any set of rubrics to count on the approval of the experts and raters that will have to use it. It is very important to listen to the voices of raters (Turner, 2000) when designing assessment rubrics. Otherwise, a tool that is intended to facilitate understanding and uniformity among raters can be perceived as an intrusion. If a group of raters which has been working with a particular scale for a long time is suddenly asked to change protocols, there is a high degree of probability that they will be reluctant. We do not always like change, particularly when this change involves extra work. New rubrics may meddle in a rater's construct of language for, let us not forget, good rubrics contain very specific

definitions of what language is. As a consequence, it is highly advisable to consult the opinion of those who will implement them later on.

In the context described in sections 3.3 and 4.1, we did not want to tamper with the construct of fellow language raters. To avoid this, we polled their opinions from the very first design stages. In fact, although we present here qualitative validation separated from stage 2 to make clear that 2 types of validation (qualitative and quantitative) are necessary, stages 2 and 3 are best developed in parallel.

Across the design of our set of rubrics we used the same 3 methods that the *CEFR* (Council of Europe, 2001:207-211) establishes for the development of scales. First, through intuitive methods in stages 1 and 2 we interpreted previous experience to provide the rubrics with shape and content. Second, through qualitative methods in stages 2 and 3 we fine-tuned descriptors. Third, through quantitative methods in stage 4 we validated the rubrics. In our case, qualitative methods involved working with experienced informants, fellow workers and raters in general. They participated to different extents in test design or marking in their home universities, where the new set of rubrics would be later implemented. Due to time and work constraints, only raters from 4 universities (Cádiz, Huelva, Jaén and Seville) helped actively. All in all, their implication contributed to start the consensus ball rolling.

Over several iterations of the initial draft, after countless Skype videoconferences and several months of work, these colleagues helped to prepare a first operational version of the new set of rubrics (see the Appendix) which would be validated quantitatively in different scenarios as described in section 4.2.4.

4.2.4 Stage 4. Validation 2 (quantitative)

The present section is critical to demonstrate that the rubrics that we designed are valid and draws on the statistical basis already described in sections 2.4.2, 2.4.3 and 2.4.4. In these sections we mentioned that the multi-faceted Rasch

measurement model is an extended version of the mathematical Rasch model (McNamara, 1996:249) and we discussed how its advantages outweigh CTT methods (Knoch, 2009:200-201; Hambleton and Jones, 1993:38).

Once our rubrics were qualitatively validated at stage 3, we carried out a first quantitative validation trial with 2 raters and 84 candidates. This was stage 4.a of the protocol. Sample size in psychometric studies is not an area of consensus for most researchers (Anthoine *et al.*, 2014:8). According to Linacre (personal communication) "(a)t least 30 candidates (more is better) and at least 2 raters (more is better)" are necessary for this type of validation, and that is the minimum that should be used at early validation stages.

Right after the first trial, we fine-tuned the rubrics at stage 4.b, following the feedback obtained during the marking sessions. Despite the fact that the initial results were promising, to make sure that our claim of validity was grounded on sound statistical arguments, we carried out a second quantitative validation trial with 9 raters and 44 candidates. This was stage 4.c of the protocol, thanks to which we came up with an improved version of the rubrics at 4.d. This second trial contained more data lines than the first (1,584 vs. 672) and was closer to the parameters of other relevant validation analyses found in the literature (Deygers and Van Gorp, 2015; Ffrench, 2003; Bruce and Hamp-Lyons, 2015 or the seminal McNamara and Adams, 1994 to quote but a few).

Before the first trial, raters 1 and 2 attended a standardization session. Before the second, raters 1-4 and 8 also attended standardization sessions while raters 5-7 and 9 did not. The candidates rated during both trials were participants in the B1 English proficiency exams that took place at the University of Jaén in February and June 2016. They were recorded on an Olympus WS-6505 device. Each interview of 2 candidates is 10-12 minutes long. As a consequence, each one of the first 2 raters devoted between 7 and 8.4 hours to rate candidates distributed in 2 sessions. For the second trial, the number of candidates was reduced to 44, which meant that each of the 9 raters at this stage devoted between 3.6 and 4.4 hours to mark candidates in 2 sessions too.

The following paragraphs present the outcome of both trials in parallel and do not explicitly break down the sequence described in table 4.2 above. However, it must be understood that such sequence was followed and that the rubrics were slightly tweaked after the first trial, in which several typos and minimal errors where found.

Our claim of validity in sections 4.2.4.1 to 4.2.4.4 will be supported by 1) aspects of data fit, 2) by an analysis of Facets vertical ruler, 3) by an analysis of rating scale utility as well as 4) by other aspects.

4.2.4.1 Data fit

Facets (Linacre, 2014) is a very powerful tool able to yield profuse psychometric data. However, for our data to be interpretable by Facets (*ibid*.) these must *fit* the mathematical model that underlies the program. Consider for example the table below, which contains scores of 6 different candidates who took a seven-item test in which questions were ordered according to their difficulty. Numbers 1 stands for one correct answer and numbers 0 for a wrong answer:

	Items							
Person	Easy **1**	**2**	**3**	**4**	**5**	**6**	Hard **7**	**Person Score**
1	1	1	1	1	1	1	0	**6**
2	1	1	1	0	1	0	0	**4**
3	1	1	0	1	0	0	0	**3**
4	1	1	0	0	0	0	0	**2**
5	1	1	1	0	0	0	0	**3**
6	0	0	0	1	1	0	1	**3**
Item Score	**5**	**5**	**3**	**3**	**3**	**1**	**1**	

Table 4.2.4.1.a. Example of data misfit (candidate 6)

We can observe from the table that items are indeed ordered according to their difficulty. At the bottom row of the table, we see that 5 people answered item 1 correctly while only 1 answered item 7 correctly. We assume that item 1 is easier than item 5 because more candidates got it right. In probabilistic formulation we would say that item scores are accurate predictors of the difficulty of items. Facets (Linacre, 2014) analyzes our data according to this assumption.

Are person scores in table 4.2.4.1.a accurate predictors of the ability of candidates? For Facets (Linacre, 2014) to build a valid analysis of our data it must be so, *i*.e., we need to assume that person scores are also accurate predictors of the ability of candidates. If we take person 1, for example, we see that she displays a believable answer pattern because she answers the first 6 items correctly and fails the last one, which is supposed to be the most difficult one. The same believable patterns are seen in candidates 2-5, who consistently tend to answer easy items correctly and answer difficult ones wrongly. When we get to person 6, however, we notice that he has answered 3 of the most difficult items correctly while he has failed easy ones. This pattern seems to contradict the ordering of the items as defined by the majority of other people. The question at this point is, can the person score of candidate 6 accurately predict this person's ability? The answer is that, probably, it cannot. In such cases Facets (*ibid*.) red-flags those candidates, raters, etc. which show unpredictable behavior. We can say that Facets (*ibid*.) signals misfitting elements.

When we feed our data into Facets (Linacre, 2014), the program goes through multiple calculations of this type to find out if the data provided are enough and have the necessary characteristics to build a probabilistic matrix. In other words, the program looks for those data that fit and do not fit its mathematical model. There are 2 ways in which we must check if our data are adequate for Facets (*ibid*.) and thus for the analysis. First, we must check the "general picture", that is to say, if the data as a whole can build a meaningful source of information. Second, even if the overall data are good, we must check if 1 or various individual elements are misfitting. Very frequently the general picture

pops up as expected but still some individual data misfit expected values. These misfitting elements point to discrepancies between predicted and observed data in our analysis. When an element of our analysis is misfitting, the individual responses will be the opposite of predicted beyond assumed randomness levels, and this generates distortion, "noise" in the interpretation of results. As it is logic to conclude, the fewer individual misfitting elements, the more reliable our results will be.

It is very easy to check for the general picture. If the data as a whole conform to the mathematical requirements, Facets (Linacre, 2014) will display a number 1 appended to the output file (Green, 2013:209) in which it presents the results. If the data do not fit or are insufficient, number 1 will not be displayed. Fortunately, in the 2 validation trials that we carried out, our data fitted the model. This means that the appraisals made from such data are valid according to multi-faceted Rasch modeling.

The second check, in which individual elements must be scrutinized, is not so straightforward because we must go individually through the different facets that constitute our data matrix. This means that, for example, in one analysis like ours which is built by the interaction of raters, test takers, linguistic features and the bands of our rubric, we must analyze all elements that compound each group individually. How do we check all this? We must analyze 2 indicators that Facets (Linacre, 2014) provides. These indicators are called infit and outfit mean-squares (McNamara 1996:137-138;169-179; Green, 2013:167; Deygers and Van Gorp, 2015:528). The optimal values for infit and outfit mean squares proposed by Linacre (2013:266) are those in the range of 0.5 and 1.5, meaning that any value within such range provides meaningful data for the reconstruction of the Rasch model.

The data of our 2 trials are provided in table 4.2.4.1.b. Notice that only 2 raters participated in validation trial 1. Raters 1 and 2 in validation trial 1 are not the same ones as raters 1 and 2 in validation trial 2 but their data have been displayed together for the sake of clarity.

The first thing we appreciate in table 4.2.4.1.b is that all values are in the expected range between 0.5 and 1.5 (Linacre, 2013:266). We also notice that many of them are very close to 0, which is good because these represent optimal values for Facets (Linacre, 2014) to operate. Particularly interesting are the very good results of the linguistic features of the second trial, which show a homogeneous and very stable behavior, much more homogeneous than in the first trial.

		Validation trial 1		Validation trial 2	
Facet analyzed		Infit mean-square	Outfit mean-square	Infit mean-square	Outfit mean-square
Raters	**1**	0.86	0.82	0.81	0.82
	2	1.18	1.16	1.44	1.43
	3	-	-	0.86	0.85
	4	-	-	0.71	0.68
	5	-	-	0.94	0.97
	6	-	-	0.95	0.92
	7	-	-	1.01	1.00
	8	-	-	0.80	0.80
	9	-	-	1.31	1.27
Linguistic feature	**Language**	1.1	0.99	0.85	0.84
	Pronunciation	0.93	0.92	1.04	1.05
	Interaction	0.76	0.72	1.00	1.00
	Discourse	1.34	1.32	1.02	1.00

Table 4.2.4.1.b. Infit and outfit mean squares of the 2 validation trials carried out

These data are indicating that we can be sure that our linguistic features are sorting candidates as they are expected to do according to the probabilistic matrix generated. As we see, our rubrics have successfully passed the first important tests, the tests that confirm that the data obtained are adequate (as a whole and individually) to build a probabilistic matrix.

4.2.4.2 Vertical ruler

The vertical ruler in Facets (Linacre, 2014) is a graphic representation of the way in which the different aspects (or facets) of our analysis interact. In the vertical ruler, the different sources of variability are already compensated (remember the "hawk vs. dove" example in 2.4.3) and the elements that interact in the rubric (raters, candidates, linguistic features and bands) are ordered lineally according to a logit scale.

The vertical rulers obtained after validation trial 1 and validation trial 2 are displayed in figures 4.2.4.2.a and 4.2.4.2.b below, and they include the different components of the analyses arranged in columns (from left to right the units of measure or logits, the raters, the test-takers, the linguistic features and the bands of the rubrics). Two raters and 84 candidates participated in the first trial and 9 raters and 44 participated in the second.

How do we interpret these 2 vertical rulers of? Let us go column by column. Column 1 (Measr) establishes the logits scale. The program has compensated for all the sources of variation and has created a unique scale for our elements, the middle point of which is at 0. Point 0 only reflects the average of all measures, an intermediate point.

Column 2 (Rater) orders the different raters that participated in the validation trial according to their severity or leniency. From figure 4.2.4.2.a we find that rater 1 was averagely severe (or lenient) because he is aligned with a value of 0 logits. Rater 2 in figure 4.2.4.2.a, on the contrary, is slightly below -1 logits, which means that he was more severe than rater 1. While the average measure of severity is not very representative when only 2 raters are analyzed, what we can already see is that, in validation trial 1, 1 of the raters (2) was indeed more severe than the other.

Figure 4.2.4.2.a. Vertical ruler for validation trial 1 (2 raters and 84 test takers)

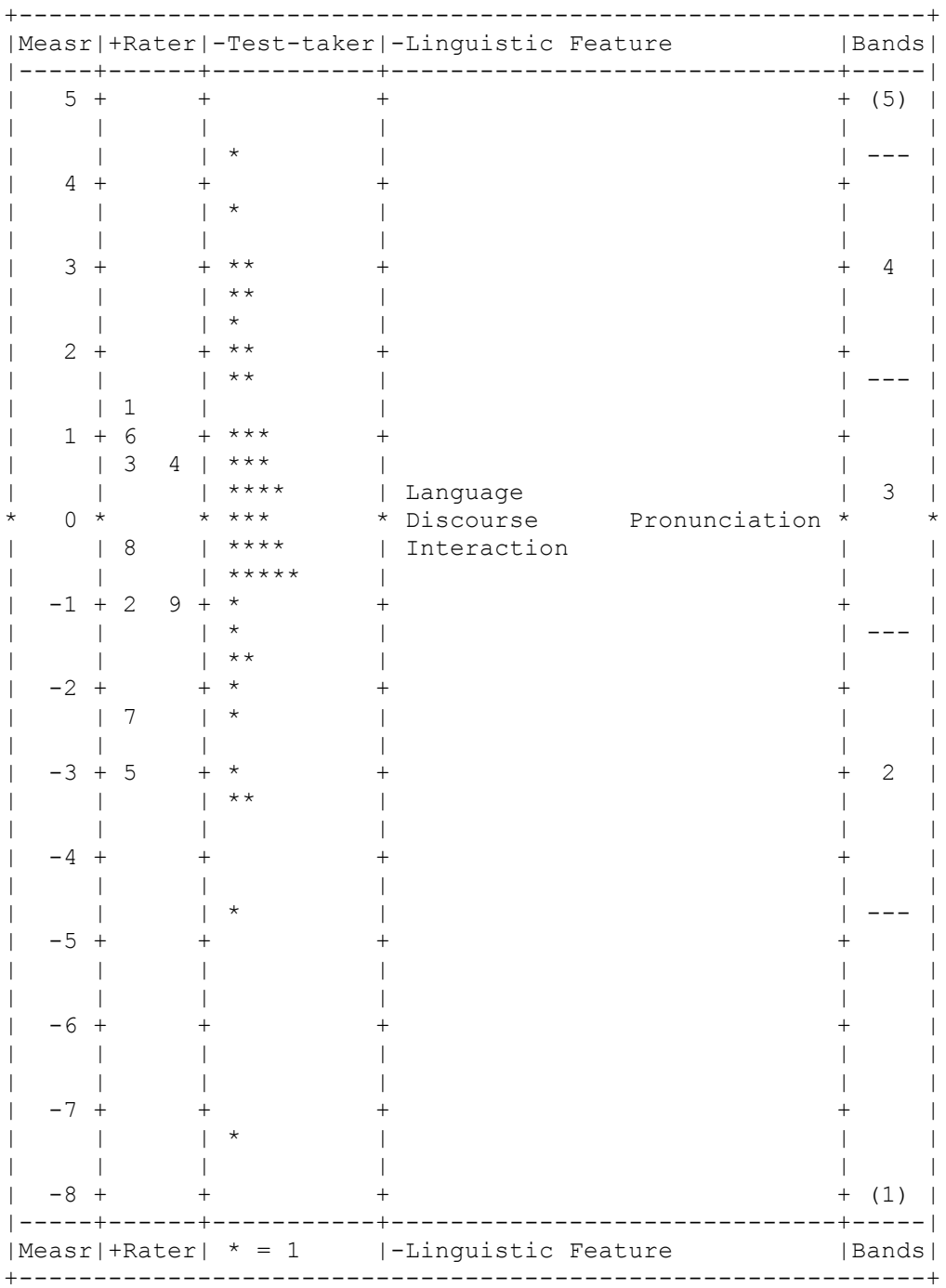

Figure 4.2.4.2.b Vertical ruler for validation trial 2 (9 raters and 44 test takers)

Validation trial 2 was more representative in this respect. In column 2 from figure 4.2.4.2.b we see that there are 9 different raters. Raters 1, 3, 4 and 6 can be considered as the group of "lenient" raters because they are in positive logit values. Raters 2, 5, 7, 8 and 9 can be considered the "severe" raters because they are in negative logit values. Remember that the great advantage of Facets (Linacre, 2014) is that in the analysis of the rest of components of our rubrics, this variability has already been compensated, which means that if one rater is more or less severe it does not affect the other data. In figure 4.2.4.2.b, raters which show negative logit values are also more disperse than the raters in positive values. Notice that, at the lower extreme, rater 5 is -3 logits far from the average 0 point. At the top extreme, rater 1 is only a bit further than 1 from logit 0. This suggests that severe raters have a tendency to perform more randomly than lenient ones and that, as a consequence, they are more disperse in the ruler.

As to what makes a rater more severe than others, we noticed that raters 7 and 9 where native speakers of English while 2, 8 and 5 were not. This might suggest that being a native speaker of the language assessed is a predictor of the severity with which our rubrics will be applied (although the contrary can also be argued). This observation is consistent with what we see in figure 4.2.4.2.a, in which rater 2 was also a native speaker of English. However, further analyses with native speakers would be necessary to confirm this hypothesis.

It is worth mentioning that rater 5 in figure 4.2.4.2.b, the one with most extreme values (-3 logits), is the only rater who works with high school students and not with undergraduate students. This might suggest that our rubrics are subject to more strict interpretations by raters outside tertiary education. Although further analyses would be necessary to prove this assumption, it is consistent with our experience in PAU (*Prueba de Acceso a la Universidad*, University Entry Tests) in Andalusia. In this respect, PAU English tests are corrected indistinctly by secondary teachers or university lecturers and the former show a general tendency to be stricter than the latter.

The last interesting aspect regarding column 2 of validation trial 2 is related to standardization. Raters 1-4 and 8 (who were considered as a sort of control group) attended standardization sessions while 5-7 and 9 did not. This suggests that, as expected, standardization sessions have a positive impact on the interpretation of rubrics. If we consider for example raters 1-4 and 8, we notice that they are closer to 0 logit values than raters 5, 7 and 9. This suggests that after standardization sessions raters tend to act more predictably when they have to interpret the rubrics. The case of rater 6 is one exception because he acted quite predictably despite not having attended a standardization session.

Let us now move on to column 3 (Test-taker). In this column Facets (Linacre, 2014) orders the test takers that were interviewed lineally. Remember that whether they were rated by severe or lenient raters does not affect this arrangement because the software has already compensated this source of variability. Thus we can be quite sure that this arrangement is an accurate predictor of the ability of our candidates.

When there are too many test takers, Facets (Linacre, 2014) uses an asterisk (*) to indicate that 2 raters occupy this position in the chart while a point (.) indicates 1 candidate. When this happens, it is denoted at the bottom of column 2. Compare the bottom of columns 2 in 4.2.4.2.a and in 4.2.4.2.b. The first thing that we appreciate in column 2 is that our rubric has been able to distribute our candidates normally. What does this mean? If we turn column 2 in 4.2.4.2.a and in 4.2.4.2.b 90 degrees left and imagine that they are in horizontal position, we will notice that the distribution of candidates resembles a Gaussian function►. This is what we mean by normal distribution►. This outcome is consistent across the 2 validation trials since the shape of the function is similar in both.

With the ruler back to vertical position, we see that the majority of test takers are clustered around band 3 from column 5 (Bands). If we go to band 3 in column 5 and draw a horizontal line straight across to the first column, we will see that the line crosses column 3 very close to where most test takers concentrate and we will also notice that this imaginary line is very close to 0 logits in column

1. This is precisely what one might expect from the candidates to this type of tests. The samples rated were taken from real English B1 proficiency tests of the University of Jaén. In such tests most candidates are expected to be at intermediate (*i.e.* pass) B1 levels. In other words, we expect that most of the candidates that sit the exam are actually prepared for the test. If, for example, the majority of candidates in column 2 were distributed in front of bands 4 and 5, the rubrics would have to be considered too lenient because most candidates would have obtained scores higher than expected, higher than the pass level that they were sitting.

Since this normal distribution is consistent across both validation trials, we have our first argument to claim that the descriptors that we designed departing from the *CEFR* (Council of Europe, 2001) were well calibrated and that the bands that we designed are also well distributed.

This ties in with what we see in column 4 (Linguistic feature) from both trials. There, the 4 linguistic features are also crossed by the imaginary line that we have drawn from band 3 and, most importantly, they are very close to each other. It is very important for the different linguistic features to occupy similar positions in column 4. If, for example, one of the features were at -3 logits while the other 3 remained close to 0, this would mean that this particular dimension at -3 was designed in a more severe fashion than the others, and this would not be good for our rubrics. We want all our features to measure different aspects of speech but at the same level, at a B1 level, which is exactly what we find in the figures above.

Consequently, since the linguistic features are grouped across the 2 validation trials, we have 1 more argument to claim that our rubrics are valid. While this outcome was more or less predictable in the case of *Language* and *Interaction*, it was not so in the case of *Pronunciation* and *Discourse*. The descriptors for the first 2 were mostly taken from specific coordinates in the *CEFR* (Council of Europe, 2001) but, in the case of *Pronunciation*, descriptors had to be designed from scratch and, in the case of *Discourse*, we had combined 2 other

linguistic features. The results above are the first confirmation of the hypothesis that our design had been carried out correctly.

4.2.4.3 Rating scale category utility

A rating scale that functions correctly should establish adequate scale step functionality. "A rating scale is made up of a number of different band levels. It is important for each level to function appropriately for the entire scale to perform efficiently" (Knoch, 2009:2004). In this respect, Linacre (1999) lists a series of statistics that should be analyzed when a rating scale functionality of interest. All such statistics are provided by Facets (Linacre, 2014) in a single output table, which we reproduce below for our 2 validation trials.

Linacre (1999) lists 8 different observations (or "guidelines") that must be taken into account when analyzing these tables. To validate our rubrics 1) we must have at least 10 observations of each category, 2) there must be regular observation distribution, 3) average measures must advance monotonically with each category, 4) outfit mean-squares must be less than 2.0, 5) step calibrations must advance, 6) ICC curves must be stable, 7) step difficulties must advance by at least 1 logit and 8) by less than 5.0 logits. Let us analyze if our 2 trials meet these requirements.

DATA				QUALITY CONTROL		RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts		Cum.		Avge	Exp. OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	Meas	Meas MnSq	Measure	S.E.	Category -0.5		from	Thresholds	Prob
1	55	55	8%	8%	-3.15	-3.07 .9			(-4.90)		low	low	100%
2	232	232	35%	43%	-1.57	-1.60 1.1	-3.79	.17	-2.43	-3.92	-3.79	-3.84	66%
3	281	281	42%	85%	-.19	-.15 1.0	-1.06	.10	.31	-1.04	-1.06	-1.06	66%
4	81	81	12%	97%	1.35	1.18 .7	1.77	.13	2.46	1.50	1.77	1.62	49%
5	23	23	3%	100%	2.21	2.48 1.4	3.08	.25	(4.32)	3.50	3.08	3.27	100%
								(Mean)		(Modal)	(Median)		

Table 4.2.4.3.a. Scale category statistics for validation trial 1

DATA				QUALITY CONTROL		RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts		Cum.		Avge	Exp. OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	Meas	Meas MnSq	Measure	S.E.	Category -0.5		from	Thresholds	Prob
1	125	125	8%	8%	-4.21	-4.34 1.1			(-5.73)		low	low	100%
2	500	500	32%	39%	-2.18	-2.18 1.1	-4.65	.12	-2.97	-4.72	-4.65	-4.68	72%
3	619	619	39%	79%	-.15	-.06 1.0	-1.32	.07	.25	-1.31	-1.32	-1.32	71%
4	241	241	15%	94%	2.10	1.91 .8	1.86	.09	2.99	1.74	1.86	1.79	60%
5	99	99	6%	100%	4.83	4.85 1.0	4.11	.16	(5.26)	4.33	4.11	4.19	100%
								(Mean)		(Modal)	(Median)		

Table 4.2.4.3.b Scale category statistics for validation trial 2

First, to check if we have at least 10 valid observations of each category (in order to avoid imprecise step calibration) we must go to column 3 (Counts used) in tables 4.2.4.3.a and 4.2.4.3.b, and check if the numbers there are 10 or bigger. In our case, we see that the lowest count in the first table is 23 and in the second 99, both well above the minimum 10 observations.

Still in column 3, to check if the observation distribution is regular (*i.e.* more or less normally distributed) we must see how the counts used advance. In column 3 of trial 1 and 2 we see that most of the observations are concentrated in a pivot-point established at band 3 (281 observations in trial 1 and 619 in trial 2) with a decreasing number of observations as we approach the extremes of the continuum. In statistics this is called a unimodal distribution, *i.e.*, a distribution in which there is only one single highest value, which is precisely what we should expect according to Linacre (*ibid.*). The fact that the apex of the unimodal distribution is located at band 3 is, again, good news for the same reasons that we mentioned in the previous section: we expect most candidates who sit this exam to hit scores around the "pass" level, located at band 3. Consequently, the majority of observations will be also concentrated at this band.

The next step is to check that average measures (column 6) advance monotonically with our categories. Monotonous advance means to advance according to a particular order, according to steps. Imagining a person that climbs stairs can be useful at this point. Lower treads correspond to lower logit averages. Higher treads correspond to higher logit averages. In our rubrics we expect the average logits to increase steadily as we climb the stairs. "In general, observations in higher categories must be produced by higher measures (or else we don't (*sic*) know what a "higher" measure implies)" (Linacre, 1999:111-112). The measures in column 6 show the average logit value of the candidates at each band level (Knoch, 2009:205). Here we expect the average mark of candidates marked with band 1 (Score 1 in column 1) to be lower than the average mark of candidates who were marked with band 5 (Score 5 in column 1). Again, this is exactly what we find in our 2 trials. In trial 1 we see that the average logit score of candidates

that received at least a band 1 mark is -3.14, and this advances monotonically all the way up to 2.21 logits for candidates in band 5. In trial 2 we find a similar pattern and with -4.21 average logits for candidates at band 1 and a monotonous progression up to 4.83 in candidates of band 5. In other words, as candidates were rated with higher bands, their average score also increased. The same goes for lower bands.

The next check is straightforward. We must simply corroborate that outfit mean-squares are less than 2. Why? Basically because the Rasch model already accounts for a reasonably uniform level of randomness throughout the data. Data with too little randomness (*i.e.* too predictable) or with too much randomness (too unpredictable) distort results by causing what Linacre (1999:113) calls "noise". For this type of analyses, mean-square fit statistics are set at 1 (*ibid*.). This means that the closer our outfit mean-squares are to 1, the better they fit the model, 2 being the threshold beyond which data are not able to provide accurate predictions. As we see, all our mean-squares in trials 1 and 2 (column 8) flow smoothly along 1, band 5 in trial 1 being the most distant value (1.4) but still far from the dangerous limit of 2.

Now we have to analyze if step calibrations advance as they are supposed to do. What we are going to check at this point is that the different bands that we have designed cover the same range of the continuum of linguistic features. In other words, we expect to find the same degree of difficulty across bands, it must be as difficult (or easy) to move from band 1 to band 2 as it must be to move from band 2 to band 3. It would not make sense to develop a scale in which lower bands are very easy to hit but in which higher bands are difficult to reach. "Failure of these parameters to advance monotonically is referred to as 'step disordering'. Step disordering does not imply that the substantive definitions of the categories are disordered, only that their step calibrations are" (Linacre, 1999:114). Luckily for us, the steps of our rubrics (column 9) are neatly organized both in trial 1 and 2. This means that we have been able to calibrate properly the

threshold of ability which is necessary to move from band to band across all the linguistic features.

We also want to scrutinize ICC curves according to the guidelines proposed by Linacre (1999:115), as described in section 2.4.2 of this dissertation. From figures 4.2.4.3.a and 4.2.4.3.b below we learn that 1 and 5 are, as expected, the extreme categories in our data. The fact that the slope of the curves is steep indicates that the different bands are good at discriminating among candidates (the steeper the slope, the more they discriminate). In general, these can be considered as very stable ICC curves, which is good. But, how do we interpret the relationship between the axes and the curves? Let us look at one example. For this we must go to the 0 logit value on the bottom axis and draw a line straight upwards. If we read off the highest number we find that it is 3 in both trials. Now we must draw another line from that 3 straight across to the probability axis. Again, in both trials the line crosses the probability axis two-thirds of the way up. This tells us that the chances of a candidate with a 0 logit difficulty measure being given a 3 is approximately 66%. This percentage can be fine-tuned is the probability axis is divided more precisely. In such case we would find that in trial 1 the probability is 65% and in trial 2 it is 68%.

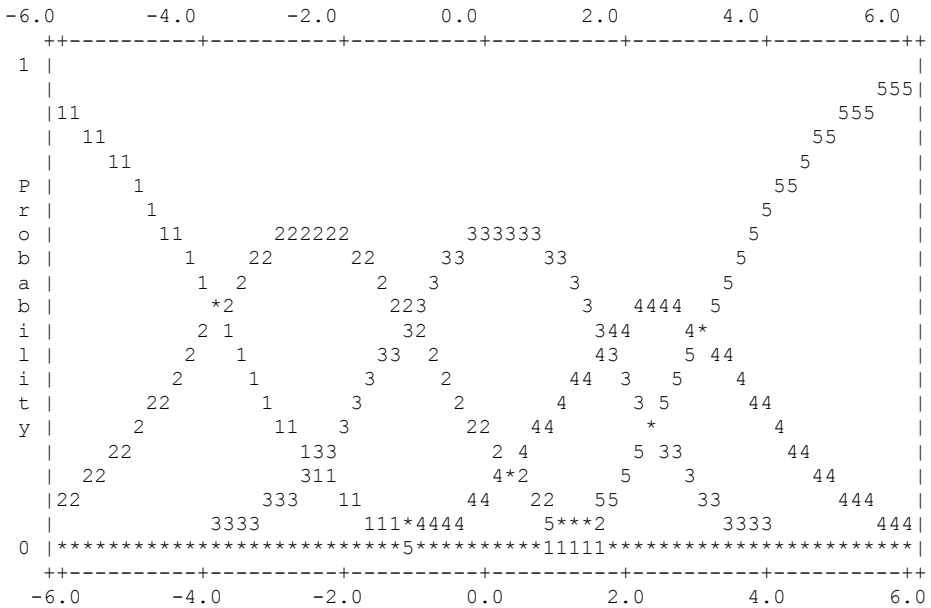

Figure 4.2.4.3.a. ICC curves for trial 1

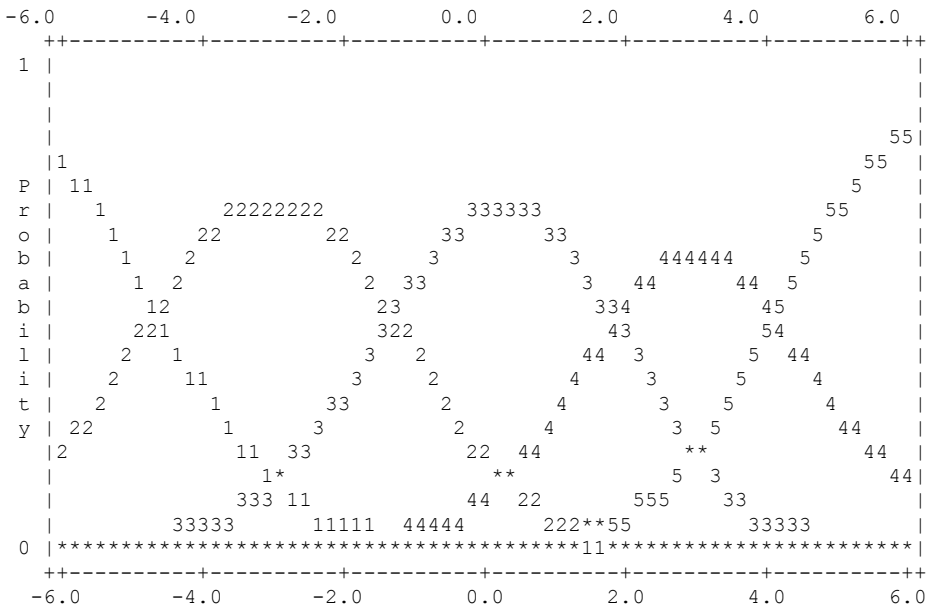

Figure 4.2.4.3.b. ICC curves for trial 2

The final 2 checks are, again, easy to carry out and take us back to column 9. According to Linacre (1999:117-120), these step calibration measures must advance by at least 1 logits (for a 5-band scale) and by less than 5 logits, a condition which is met by our data and which guarantees that the distinctions made by our bands are dichotomous.

As we can see, the 8 checks described above are easy to carry out through simple Facets (Linacre, 2014) analyses and can thus be incorporated in any protocol to design rubrics as a powerful quality check.

4.2.5 Stage 5. Implementation

After rubrics are validated, these must be brought live. After the successful validation, the next big step in this sense will be to report the data obtained to fellow universities to meet an agreement on how these will be implemented.

The rubrics discussed here (see the Appendix) and the designed protocol were originally meant to be implemented before the *viva voce* of this dissertation. However, getting 9 universities to work at the same time can be difficult. Some of these universities are already working at different stages of the protocol (Cádiz, Huelva, Jaén or Seville) while some others have committed to do so during

academic year 2016-2017. Moreover, to obtain a comprehensive view of how the rubrics are actually working at Andalusian universities, we need to collate the data corresponding to 1 academic year, counting from the moment at which these are implemented across the different institutions.

Thus, once the rubrics are finally implemented in all (or most) universities by the end of 2016-2017, we will still have to collate data during another academic year. This work is still to be done (see section 6.2.2) and will be done in due time, but we chose not to delay the end of the dissertation until academic year 2017-2018 since the main validation process of the rubrics has already proved successful and since the mobile application that will be used in the process is also finished. In fact, even if we presented the aforementioned data, the protocol has been designed to be in continual improvement, as we are going to see later on. This means that there is never going to be any particular point of time at which the protocol will be truly finished. What follows in sections 4.2.5 and 4.2.6 is the description of the plan already developed to implement and revise the protocol.

At stage 5.a of our protocol, for example, standardization and training sessions will be necessary. At the moment of finishing this dissertation, the plan is to select a group of team leaders from all the universities that will use the rubrics. These team leaders will attend standardization sessions in which they will be explained how the rubrics were built and how these must be used. The participation of the colleagues that collaborated in the design of the rubrics will be important in these sessions. The message will be more easily conveyed if it comes through co-workers who had an active role in the design of the measurement tool. Remember that we do not want to impose the rubrics, we want raters to be convinced of their utility.

These standardization sessions should take place at least 2 times every academic year, more frequently if the frequency of tests is higher too. The sessions should be attended by at least 1 team leader from each one of the

universities participating in the implementation of the rubrics. The sessions should be organized following different steps:

Step 1. Prior to the sessions, a set of 5 benchmarked► recordings will be sent by email to the raters. These samples will have to contain meaningful examples of the different points of the rubric. Raters will be asked to order the scripts and try to justify their ordering by using the rubric at home. The recordings are accompanied by a "gold standard"► recording containing a list of functions a candidate at the required level should be able to carry out. The "gold standard" recording will be taken from the specimens that the Council of Europe has made available (CIEP, 2008). Also at home, raters will have to complete a *CEFR* (Council of Europe, 2001) familiarization exercise.

Step 2. The training of the standardization sessions will start with the answers that raters prepared at home for the *CEFR* (Council of Europe, 2001) familiarization exercise. These answers will be shared among the participants commenting on the descriptors that exemplify the speaking ability of a B1 candidate. Participants will be then asked to share the order of the set of recordings marked at home and comment on the qualities that motivated their decisions based on the rubrics. The opportunity is taken to clarify ambiguous terms in the scale.

Step 3. A set of previously benchmarked problematic scripts (3 to 4) is then distributed and raters are asked to rate them individually. Once they have finished rating, the ratings will be collected and shown anonymized for public discussion. The discussion should be used to exemplify typically problematic issues, explain procedures, level of discrepancy allowed, etc. The benchmark for the problematic scripts will be shown together with the comments made by rater participants explaining their decisions.

Step 4. Questions and answers will be encouraged at this final phase to help raters understand the benchmark but agreement will not be forced. Those raters who deviate from the score by more than 1 point out of 5 will be called back again individually to discuss their reasons.

After these standardization sessions, the representatives from the participating universities that attend them should be able to carry out similar procedures taking the role of conductors at their home universities.

The results obtained in these sessions will have to be analyzed statistically to check levels of inter-rater reliability. Besides the general data about infit and outfit mean squares of raters previously displayed in table 4.2.4.1.b, Facets (Linacre, 2014) also offers more specific information about inter-rater reliability which can be used in standardization sessions at stage 5.a of our protocol. Facets (*ibid*.) offers statistics of exact observed agreement among raters and of expected observed agreement between raters. The exact agreement observed reports what percent of the ratings by one particular rater agree exactly with the ratings made by other raters. On the other hand, the exact agreement expected reports the agreement that would be expected if the data of one particular rater fitted the Rasch model perfectly.

In general, the observed (real) agreement tends to be slightly higher than the expected (theoretical) agreement because raters tend to be "agreeable" with each other. Levels of observed agreement considerably higher than expected agreement are normally a red flag indicating problems. Facets (Linacre, 2014) models raters to be independent experts in which a proportion of agreement is expected as well as a proportion of disagreement. However, degrees of exact agreement close to 100% show that raters are behaving the same as rating machines and that they have become a part of the data-collection mechanism. In this case, they are no longer a facet of our measurement situation and, as a consequence, if we ever spotted such levels of exact agreement, a different type of approach should be used for the study. The data of our 2 validation trials are displayed in table 4.2.5.a below. As in the case of table 4.2.4.1.b, raters 1 and 2 in validation trial 1 are not the same ones as raters 1 and 2 in validation trial 2 but their data have been displayed together for the sake of clarity:

	Validation trial 1		Validation trial 2	
Rater	Exact agreement observed	Exact agreement expected	Exact agreement observed	Exact agreement expected
1	41.4	41.2	40.4	35.7
2	41.4	41.2	39.9	42.1
3	-	-	41.9	40.3
4	-	-	43.2	40.7
5	-	-	26.8	28.4
6	-	-	42.6	38.3
7	-	-	31.6	34.4
8	-	-	44.2	43.3
9	-	-	41.3	41.9

Table 4.2.5.a. Exact agreement observed and expected among raters

From this table we learn that Facets (Linacre, 2014) could provide little meaningful data for validation trial 1, in which only 2 raters participated. Yet, the results of inter-rater reliability for validation trial 1 are slightly higher in the exact agreement observed than in the expected exact agreement, as we anticipated. The results for the second validation trial, on the other hand, provide interesting data about the behavior of which raters should be revised. Raters 2, 5, 7 and, to a lesser extent 9 show less observed exact agreement than expected. They are performing "less agreeably" than the rest.

We have already mentioned that Facets (Linacre, 2014) is able to compensate for these differences by applying the probabilistic Rasch model and, as a consequence, these variances do not affect the data described in 4.2.4. However, for the sake of uniformity, it might be relevant to zoom in those features and bands in which these "less agreeable" raters are behaving differently. If we are able to spot such bands, we will be able as well to devote more work to clarify the definitions in them.

Unfortunately, Facets (Linacre, 2014) does not provide any such statistic but, fortunately, there are some possible ways to obtain it. A good candidate to check inter-rater reliability at specific linguistic features and bands is κ statistics. This CTT method and its mathematic rationale are easy to understand and to apply, which provides a good opportunity for developers untrained in statistics to take a first step into mathematical standardization. It is very frequently used in clinic medicine to check agreement among doctors when they have to diagnose patients. In medicine, giving the right diagnosis may entail the difference between life and death. Thus, if κ statistics is reliable enough in such delicate contexts, it may well be valid for our purposes too.

The data that we can obtain from 2 raters using our rubric qualify for an inter-rater reliability study based on a multinomial distribution► model, which is characterized by the following facts:

- The experiment consists of *n* repeated trials.
- Each trial has a discrete number of possible outcomes (ordinal categories, our bands).
- The trials are independent in the sense that the outcome of one trial does not affect the outcome of other trials.

Let us see how this works through 2 examples, a theoretical one and a real one. First, for the theoretical example, we will imagine that we want to investigate how one linguistic feature behaves. Since κ statistics only allows comparing raters in groups of 2, we will have to choose 2 raters. One of these raters could (but does not have to) be a rater that has proved adequate balance of observed and expected exact agreement and the other rater could be a misfitting rater.

If we put in a 2-axis table the bands that each rater endorsed to the different candidates in this one linguistic feature, we should expect something as table 4.2.5.b below. Our rubric has only 5 bands, but this analysis can be carried out with an indefinite (*n*) number of them. On the left side of the table we see 1, 2, 3, 4, …, and *n*, which represent the bands for rater 1. At the top of the table we see the same for rater 2. At the bottom and at the right of the table we see the totals.

Rater 1 \ Rater 2	1	2	3	4	…	n	Total
1	a_{11}	a_{12}	a_{13}	a_{14}	…	a_{1n}	b_1
2	a_{21}	a_{22}	a_{23}	a_{24}	…	a_{2n}	b_2
3	a_{31}	a_{32}	a_{33}	a_{34}	…	a_{3n}	b_3
4	a_{41}	a_{42}	a_{43}	a_{44}	…	a_{4n}	b_4
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
n	a_{n1}	a_{n2}	a_{n3}	a_{n4}	…	a_{nn}	b_n
Total	c_1	c_2	c_3	c_4	…	c_n	N

Table 4.2.5.b. Theoretical distribution of ratings used to obtain κ coefficient

Now, let us think about imaginary candidates and remember that we are looking at a particular linguistic feature. If both rater 1 and rater 2 endorsed our imaginary candidates band 1 in this one feature, these candidates should occupy coordinate a_{11} in our table, 11 being the corresponding Y and X values (or bands) that the candidates were awarded by raters. Similarly, candidates in position a_{12} would have been endorsed band 1 by rater 1 and band 2 by rater 2, candidates in position a_{21} would have been endorsed band 2 by rater 1 and band 1 by rater 2, etc. If our raters are consistent in the use of the bands of this particular linguistic feature, they will award every single candidate the same band. As a consequence, agreement between raters should be found across the diagonal of the table, identified in red, which is where the ratings of both raters coincide. Everything outside this diagonal represents candidates which rated differently by the 2 raters.

The κ coefficient measures the overall percentage of agreement by means of the expression

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

Where P_o is the proportion of observed agreements and P_c is the proportion of agreement expected by chance (Sim and Wright, 2005:258), which are calculated through

$$P_o = \frac{1}{N}\sum_{i=1}^{n} a_{ii}.$$

$$P_c = \frac{1}{N^2}\sum_{i=1}^{n} b_i c_i$$

and hence

$$P_o = \frac{1}{N}\sum_{i=1}^{n} a_{ii} = \frac{a_{11} + a_{22} + a_{33} + \ldots + a_{nn}}{N}$$

$$P_c = \frac{1}{N^2}\sum_{i=1}^{n} b_i c_i = \frac{b_1 c_1 + b_2 c_2 + b_3 c_3 + \ldots + b_n c_n}{N^2}$$

Although we have used colors for the sake of clarity, let us see how this works in a real case. Let us consider now table 4.2.5.c below, in which we include some data taken from a real standardization session

Rater 1 \ Rater 2	1	2	3	4	5	Total
1	1	0	0	0	0	1
2	0	3	1	0	0	4
3	0	1	25	1	0	27
4	0	0	0	7	0	7
5	0	0	0	0	1	1
Total	1	4	26	8	1	40

Table 4.2.5.c. Real data used to obtain κ coefficient

In table 4.2.5.c we see that raters 1 and 2 endorsed the same 25 candidates band 3. They also agreed on awarding band 1 to one candidate, band 2 to 3 candidates, band 4 to 7 candidates and band 5 to one. This means that the different bands of this linguistic feature are apparently clear and they are used consistently because, for example, no candidate was endorsed with band 1 and 5 by raters 1 and 2. This is what 0s mean in all the coordinates in which they appear. They mean that there is no candidate awarded with the values where Y X cross at this particular position. So, while in the example above there are 0 candidates awarded with band 2 by rater 1 and with band 1 by rater 2, 0 candidates awarded with band 3 by rater 1 and with band 1 by rater 2, etc., there are some 1s in the table. There is, for example, one case in which one candidate was endorsed band 3 by rater 1 and band 2 by rater 2, one candidate who was endorsed band 2 by rater 1 and band 3 by rater 2, and one final candidate who was endorsed band 3 by rater 1 and band 4 by rater 2. This is still within the expected range of values because the difference in the scores is located in

adjacent bands and the number of such differences is reduced (there are only 3 cases).

But how can we transform these data into objective, interpretable figures? The answer is in the formulas above for κ coefficient. If we go back to these formulas, by substitution we find that

$$P_o = \frac{1}{40}\sum_{i=1}^{5} a_{ii} = \frac{1+3+25+7+1}{40} = 0.925$$

$$P_c = \frac{1}{40^2}\sum_{i=1}^{5} b_i c_i = \frac{1\cdot 1 + 4\cdot 4 + 27\cdot 26 + 7\cdot 8 + 1\cdot 1}{40^2} = 0.485$$

$$\kappa = \frac{P_o - P_c}{1 - P_c} = \frac{0.925 - 0.485}{1 - 0.485} = 0.864$$

Thus, in the case proposed in table 4.2.5.c, κ=0.864. Since when total coincidence occurs then κ=1, the closer we get to 1, the better. The data obtained after several standardization sessions at stage 5.a can be collated and used to establish the minimum accepted thresholds of reliability, which is one of the objectives of stage 5.c. These data can also be compared with the data obtained at 5.b.

As mentioned at the beginning of section 4.2.5, once the second semester of academic year 2016-2017 begins, the rubrics will start to be used in real exam conditions at stage 5.b in most Andalusian universities. The first piloting will take place in 2 different exams of the University of Jaén, which will be preceded by standardization sessions. We expect the rubrics to be used at most other Andalusian universities as of the beginning of academic year 2017-2018. The data of standardization sessions and the extent to which they mirror data obtained during the administration of exams will mark a new milestone in the protocol and

will provide us with the opportunity to analyze the consequential validity of the protocol.

4.2.6 Stage 6. Revision

The implementation of our newly designed rubrics should be revised yearly. For the revision the data from standardization sessions and from live administration of exams will be used. As stated before, these data will enable a revision of the rubrics and will be a quality check of the way in which it is being implemented. Besides, feedback from stakeholders (raters and candidates) will be useful to establish the consequential validity of the instrument.

Several Andalusian universities have already suggested that, after the validation of our B1 oral scale, the process should be scaled and B2 and C1 oral scales should be designed through the same protocol. This will also be another interesting opportunity to set the protocol to work. Besides, these universities have also mentioned that the same should be done with B1, B2 and C1 writing scales (see section 6.2 for further discussion).

CHAPTER 5. DESIGN OF A MOBILE APPLICATION: RUBRIK©

A reliable assessment of candidates to oral exams is not only limited by the reliability of the rubrics used but also by rater cognition, a factor of the sense that raters make of the rubrics and their working memory limitations (see section 2.2.2). From the beginning of this research study we believed that we could go beyond the design and validation of a set rubrics. We thought that we could create an innovative system that could make the work of language raters easier.

In long evaluation sessions it is very difficult to have a clear picture of all the information that your rubrics contain. In long evaluation sessions raters are also limited by ordinary working memory constraints as well as by fatigue.

Generally speaking, when marking oral performance through rubrics in language proficiency tests, raters tend to focus on intermediate bands and then compare candidates' performance to such intermediate bands. If the performance of the candidate is above intermediate bands, they check higher bands by eyeballing their physical copy of the rubric or by retrieving such bands from their memory span. The same goes for performances below intermediate bands. This process of judging-comparing-judging is repeated continuously during marking sessions and may become dull for raters, who are at risk of ending up marking impressionistically due to fatigue. We thought that presenting our rubrics in an intuitive, rater-friendly way would partially alleviate the fatigue produced by the judging-comparing-judging process.

The present chapter describes how a mobile application to implement rubrics through tablets was envisaged and the different concerns and steps that were taken into account along the way towards its final build[6]. The mobile application is named Rubrik© and can be downloaded from Google Play (Play

[6] I am very much obliged to engineer Raul Pérez Fuentes for his priceless advice and constant help in the development and programming of this mobile application.

Store) on Android tablets. A CD can also be found attached to this dissertation which contains the apk file of the application. Please, notice that the application is not optimized for smartphones and its content will not display properly unless it is installed in a tablet. We recommend to use devices with a minimum of 9″ screens.

We strongly recommend to install Rubrik© directly from Google Play (Play Store). This is the easiest way to download the application and the best form to keep Rubrik© updated. To download the application, simply launch the Google Play (Play Store) app on your Android tablet and search for the application by typing "Rubrik" in the browser of the application. Notice that the application may not appear in the first place and you may have to scroll down to find the app icon to download it (see the icon displayed in figure 5.1.4.b). Once you find the icon of Rubrik©, launch it and follow the instructions on the screen of your device.

In the event that you have trouble downloading Rubrik© from Google Play (Play Store), we have attached a CD to this dissertation which contains the first gold version of the application. Notice, however, that the last updated version of the application will only be accessible through Google Play (Play Store). If you choose to use the CD, you will have to copy its content on to a desktop computer. Once you have the apk file of the attached CD on your desktop computer, there are several ways in which you can install the application. You can, for example, send the apk file through email to one account that you have configured on your tablet. Once you have sent the email with the apk file to your own account, open it on your tablet and simply execute the apk file from there. From your desktop computer you can also upload the apk file on to any cloud storage account that you have installed on your device (Google Drive, DropBox, etc.). Once you have uploaded the file to your cloud storage account, open the cloud storage account on your mobile device, synchronize it and execute the application from there.

5.1 Design concerns

Our objective thus was to create a mobile app that could provide raters with visual aid to identify band descriptors and that, at the same time, provided other necessary functionalities like voice recording or an integrated stopwatch to control the timing of interviews. The app had to be built in such a way that it could be used in different contexts and according to different criteria, not only in the context for which our rubrics were designed. The more versatile, the more likely the application was to be adopted by other colleagues.

5.1.1 Functionalities

Early on, based on our experience as raters of oral language proficiency exams, we established a list of functionalities that the application should contain. It should be able to:

1. Evaluate 4 or more linguistic features
2. Divide each linguistic feature into a minimum of 5 bands
3. Preload specific descriptions
4. Preload the ID of test takers before marking sessions
5. Record the performance of test takers
6. Mark each test taker individually
7. Export the data generated

Our rubrics, the rubrics designed in chapter 4, contained 4 linguistic features, but other analytic scales may contain up to 6. Because of this, we wanted to make the application match our needs but we also wanted to make it possible for other professionals to add further categories. The same happened with bands. Our rubrics were broken down into 5 bands but there are others which may contain more. We too wanted to make it possible for other professionals to choose the number of bands that matches their scales.

The possibility of uploading custom-made descriptors was crucial if the application was to be used by other professionals. Again, we would preload the descriptors that we designed and validated through section 4.2, but these are not necessarily the descriptors that other raters want to use in contexts different to ours. As a consequence, we had to envisage a system to upload other descriptors on the application.

The data generated by the app should be linked to particular candidates. Consequently, the ID of these candidates has to be retrieved at some point of interviews. We thought that typing the ID of candidates individually during sessions should be an option, but that the application should also enable the possibility of pre-uploading batches of data. Before marking sessions, raters should be able to type or directly import from Excel files (or the like) the ID of large numbers of candidates to save time during the marking session.

At the early stages of design, when functionalities were being revised, the possibility of recording the performance of candidates seemed very attractive. However, it posed 2 problems. On the one hand, not all oral examinations are recorded as ours is. Cambridge oral exams, for example, are not. In some countries it is simply illegal to record candidates during their oral exams. On the other hand, there were also technical issues. Storing locally all the recorded files would require a varying amount of bytes which is not available in all types of devices. Likewise, the microphone for the recording must remain close to candidates. We experimented with different types of Bluetooth connections and devices only to reach the conclusion that the amount of work which was necessary to implement this functionality did not pay off. Should recording be necessary, digital voice recorders would be more versatile. As a consequence, this functionality was soon discarded.

Although there are many tests in which candidates are interviewed individually, the tendency is nowadays to interview them in groups of 2. We chose to develop Rubrik© to evaluate candidates in groups of 2 because most of the Andalusian universities involved in the project interview their candidates in

pairs. This favors peer-to-peer interaction and alleviates the pressure that candidates suffer when they have to interact exclusively with their examiners. We wanted our application to reflect this fact and wanted it to allow the possibility of rating at least 2 candidates in parallel without losing track of which marks are awarded to each one of them. As a consequence, the marking screen for both candidates should be quickly and easily accessible at any time during the interview.

Of course, all the data collated during the marking session would have to be exportable in a popular format for their processing.

To implement all these functionalities, aspects such as programming, user experience (UX►), user interface (UI►) and production had to be previously planned.

5.1.2 Programming

After determining the functionalities, the next big step was to find a versatile programing environment to define aspects as data storage (*i.e.* how data will be stored in the application), the operative systems in which the application would be installed or the use of Internet connection.

In this last respect, for example, it was clear that, due to the conditions in which oral tests normally take place, it would be difficult to rely on an Internet connection for data storage. Tests may take place in locations unfamiliar to raters with no Internet connection. Thus, the most logical conclusion was to create an offline data storage model based on a type of JavaScript► library known as WebSQL, which would allow local storage of data. WebSQL is, then, a web-based application development environment which allows the storage of databases locally. The use of WebSQL was at the same time necessary and a decision that marked the course of programming. We needed WebSQL because it allowed us to do without Internet connection, but at the same time it forced us to work in a web-based environment. In other words, if we wanted to use WebSQL (and we needed to do so), we would have to develop our application through a

web browser (such as Chrome, Explorer, Safari or Firefox) and then export in formats that different operative systems could recognize. This is the equivalent of creating a Word document and then exporting it in pdf, txt or doc format. Finally, we envisaged the following integrated development environment:

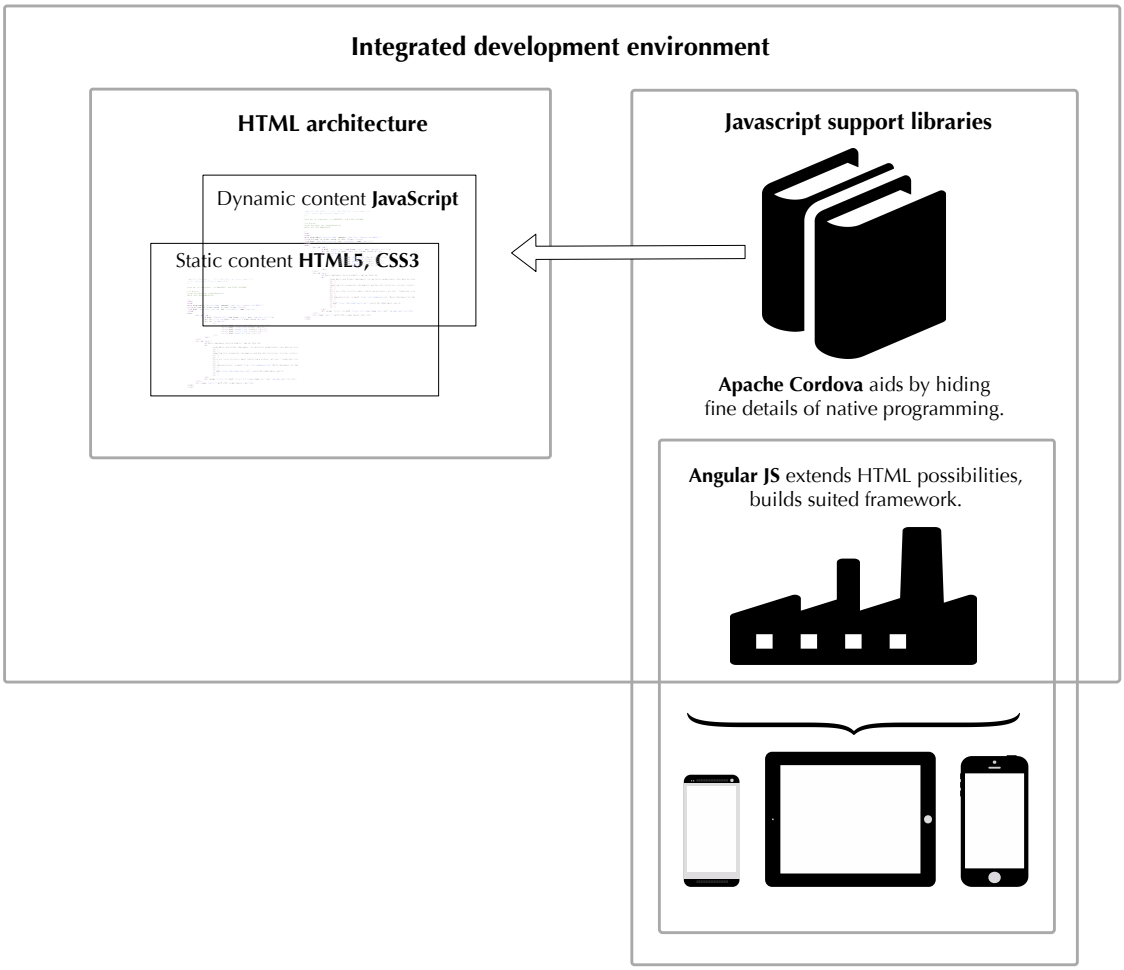

Figure 5.1.2. Integrated development environment

In this integrated development environment there were 2 main layers, first the HTML structure and, second, the JavaScript support libraries. The HTML structure is built through the 3 programming languages that compound virtually all the content that we see on web pages and most applications, HTML5 itself, CSS3► and JavaScript. The HTML layer can be understood as the basis of the application. If we had to compare the process of programming an application with cooking, HTML5, CSS3 and JavaScript would be the ingredients of a recipe. The second layer, which contains JavaScript support libraries was created through

Apache Cordova (*Apache Cordova*, 2016), an open-source mobile development framework. Apache Cordova (*ibid*.) allows using standard, common web technologies such as HTML5, CSS3 and JavaScript for cross-platform development. Cordova (*ibid*.) is a type of software that hides some details of programming and which, at the same time, provides extra dynamic ready-made JavaScript libraries. To continue with the comparison of our application and cooking, Apache Cordova (*ibid*.) would be like a stand-alone ready-made sauce that you can buy in the shop and add to the basic ingredients of your recipe. If you are preparing paella, you can buy the cooking base (*i.e*. Apache Cordova, *ibid*.) but you still have to boil the rice (HTML architecture) in it. Within Apache Cordova (*ibid*.), a second layer was coded in the early build of the application through AngularJS (*AngularJS*, 2016). AngularJS (*ibid*.) is a toolset for building the framework that allows us to virtualize mobile applications on a web browser without necessarily installing them on a physical mobile device. Thanks to AngularJS (*ibid*.) we could test the early versions of our application through Firefox, Safari or Chrome without installing them in actual mobile devices.

The great advantage of using Apache Cordova (*Apache Cordova*, 2016) and AngularJS (*AngularJS*, 2016) combined is that they allow a cross-platform approach. This means that you can program through them without defining if the application will be used in Android or iOS► because they allow us the possibility to export code for both operative systems. This mixed approach was important because we wanted to leave the door open at this early stage to possible changes in the platform in which it would be published. At the same time, this cross-platform approach could also be an advantage even if it was successfully published for only 1 such operative system. If it worked in Android, for example, it would be very easy to make it available for iOS or Linux. Thanks to Apache Cordova (*Apache Cordova*, 2016), we would be able to easily compile the code of the application and export it on to other platforms such as Android or Linux. In Apache Cordova (*ibid*.) applications execute within wrappers targeted to each platform, and rely on standards-compilant bindings to access each device's

capabilities such as sensors, data, network status, etc. (Cordova Overview, 2016) without necessarily designing the software as a native application.

5.1.3 UX analysis

In computer sciences, UX refers to the process of enhancing user satisfaction by improving usability, accessibility and interaction patterns between users and software.

We wanted to create an application with a clean interface that could help raters in their work. Traditional paper-based methods have long proved to be usable and effective and, as a consequence, we were aware of the fact that any proposed change introduced by our application would have to improve the traditional system. Otherwise, if the use of the application becomes difficult, users will abandon it and will go back to traditional methods.

As a consequence, the application was built on the principle of simplicity. In building our application we followed the motto *less is more* in the sense that van der Rohe made of it, that is to say, as an aesthetic tactic of arranging the necessary components of the piece of software so as to create an impression of extreme simplicity that is at the same time at the service of the user and at the service of functionality. The application did not only have to make the work of raters easier, it had to give a very good impression from the beginning.

We were also aware of the fact that we had to build a modular piece of software equipped with the abovementioned functionalities and, at the same time, leave room to implement new ones. We had to build a modular product with a great margin for customization.

Another key aspect was the use of Internet connection as mentioned in section 5.1.2. Since marking sessions are often carried out in non-familiar premises with limited resources, we wanted to make an Internet-independent application or, at least, one that would not require the use of an Internet connection to work properly. Our assumption was that the Internet should only be necessary for those functionalities that were not required during the marking

session, as for example candidate's data pre-load or descriptors editing, which could be planned and carried out in a friendly and controlled environment (*i.e.* at raters' office, for example) before the test.

A smooth experience during the use of the application depended on the way in which the functionalities presented in 5.1.1 were implemented with the aforementioned simplicity through an attractive interface. The main objective in this experience was to make the different bands and descriptors quickly and easily accessible during marking sessions.

For this we chose to use a cover flow interface. Cover flow are animated, three-dimensional graphical user interface that are frequently integrated within Apple devices. Cover flow gestures allow users to flip through photographs, bookmarks or, as in our case, through descriptors by sliding fingers across the touch screen. Different variations of this system are nowadays included in virtually all mobile applications, which provides Rubrik© with an intuitive baseline of gestures.

5.1.4 UI design

UI design is closely related to UX experience since both must marry to produce a good and user-friendly final product. Again, UI was built on the principle of *less is more*. Simplicity would also help to avoid complex programming, which might make the app crash during a marking session with the consequent loss of data.

For this we used several graphic mock-ups (*i.e.* prototypes of the application) that contained executable basic functionalities and that were easily accessible through regular browsers thanks to Apache Cordova (*Apache Cordova*, 2016) and AngularJS (*AngularJS*, 2016). Figure 5.1.4 below shows an early mock-up of the application running on Apple's Safari through a local host.

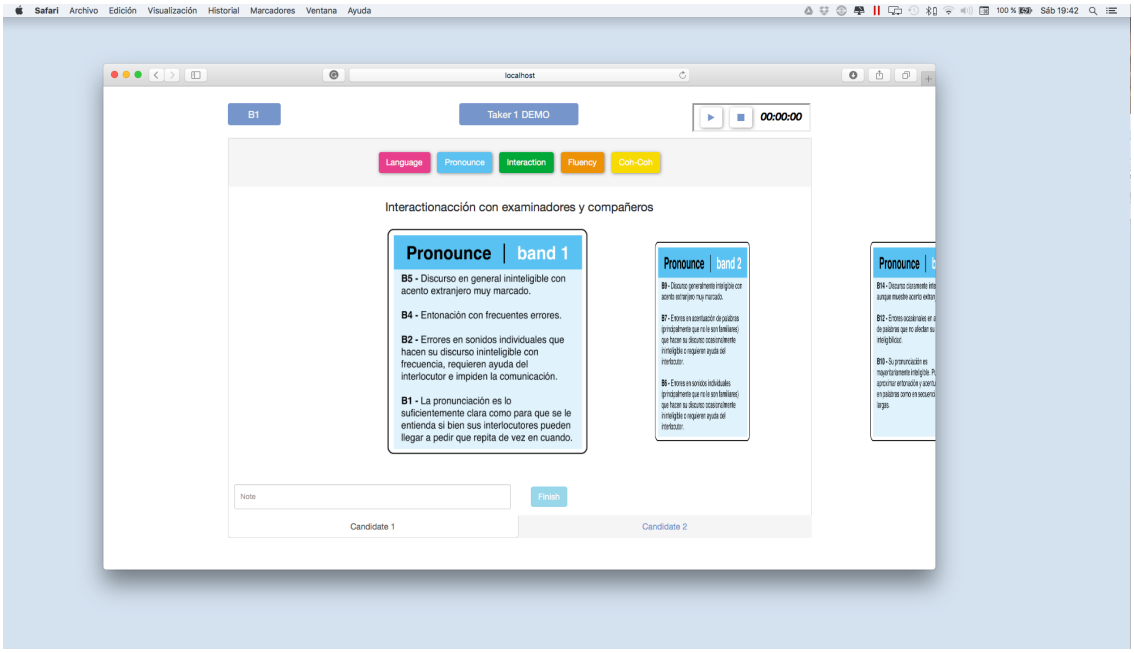

Figure 5.1.4.a. Early mock-up of the application with a bug

These mock-ups were circulated among colleagues to obtain feedback prior to the final codification phase. It helped to find and solve bugs as the one displayed in figure 5.1.4 above, which generated descriptors in wrong places (notice the blue box at the right margin of the white area).

Thanks to the previous mock-ups, different menus and screens were created for the application, and a recognizable icon/logo was designed as well, the one shown in figure 5.4.1.b below. The application was also named during this phase of design. It would be called Rubrik©, a name that played with the idea of associating the mechanics of the application to those of the Rubik cube toy. The famous toy associates different colors to the different sides of a cube and Rubrik© links each linguistic feature to a color. At the same time, the way of turning Rubik cube's pieces to solve it is similar to the cover flow system described in section 5.1.3, which is used in Rubrik©.

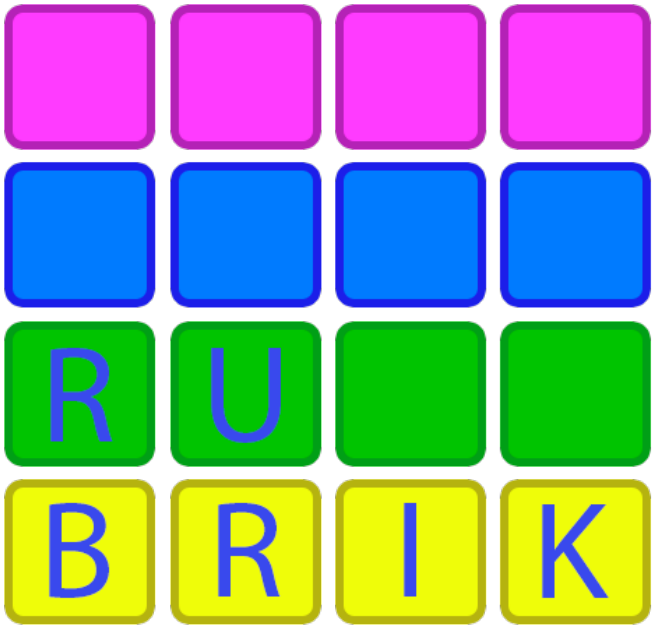

Figure 5.1.4.b. The logo/icon designed for Rubrik©

Along with the icon/logo, that would be displayed in the main screen of the application, the main screens and menus of the application were built during this stage of UI design.

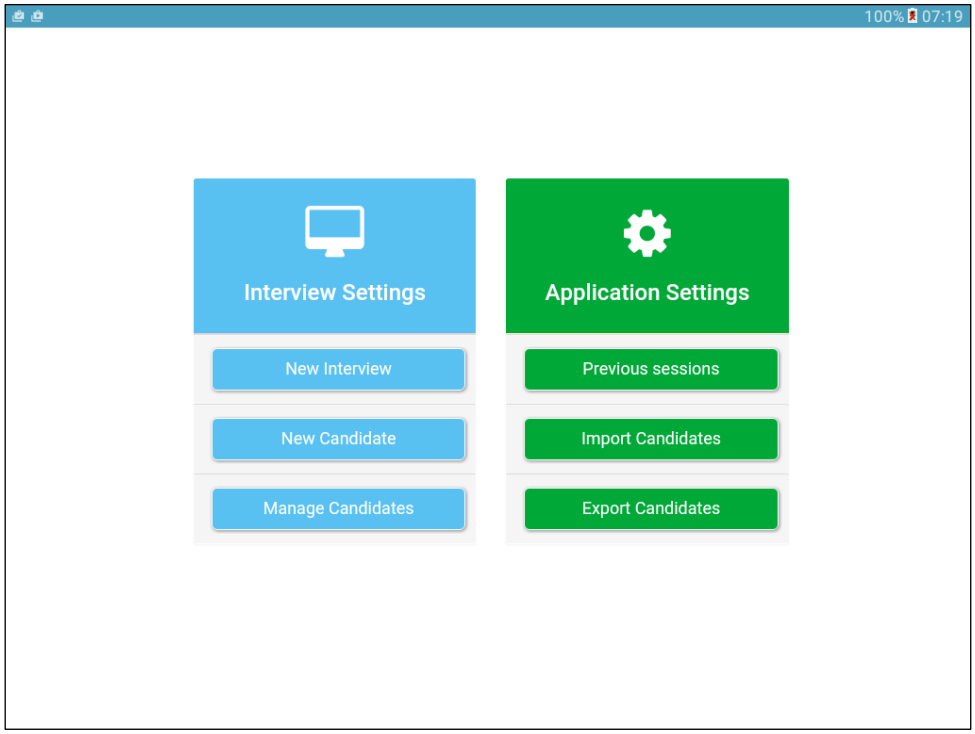

Figure 5.1.4.c. Main menu screen with settings

The main menu screen had to contain all the basic features of the application that made it to the final stage of design (notice that some of the initially intended functionalities, such as voice recording, had already been discarded at this point). The main menu screen (see figure 5.4.1.c) contained 2 different submenus. On the one hand we have *Interview Settings*, containing the basic functions related to the configuration of oral interviews and, on the other, *Application Settings*, where we find the different options that Rubrik© offers to manage the data generated during interviews. In the following paragraphs we are going to describe the functionalities that each button offers.

In the first submenu, *Interview Settings* (see figure 5.4.1.c), buttons are displayed in blue. Here, the *New Interview* button generates a second level screen (figure 5.4.1.d below) through which raters can select the name of the candidates to be interviewed as well as the language and the level. In the current version of the application only English language and level B1 are available. We have chosen to show the other languages and levels, although they are not usable, to give a taster of the potential scalability of the application, which can be used for different languages and different levels at the same time.

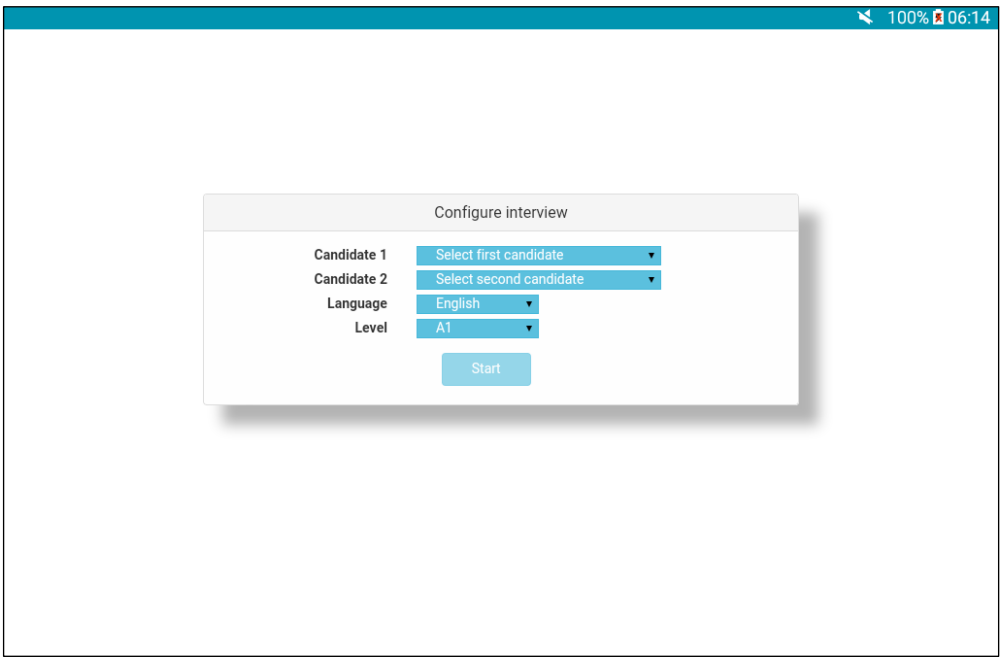

Figure 5.1.4.d. Screen for new interviews

The second button in the *Interview Settings* menu (figure 5.4.1.c) was *New Candidate* (figure 5.4.1.e below). This screen allows the possibility of introducing the name of new candidates manually for those cases in which the ID of all of them was not previously imported. For this purpose, a pop-up virtual keyboard was integrated in the screen.

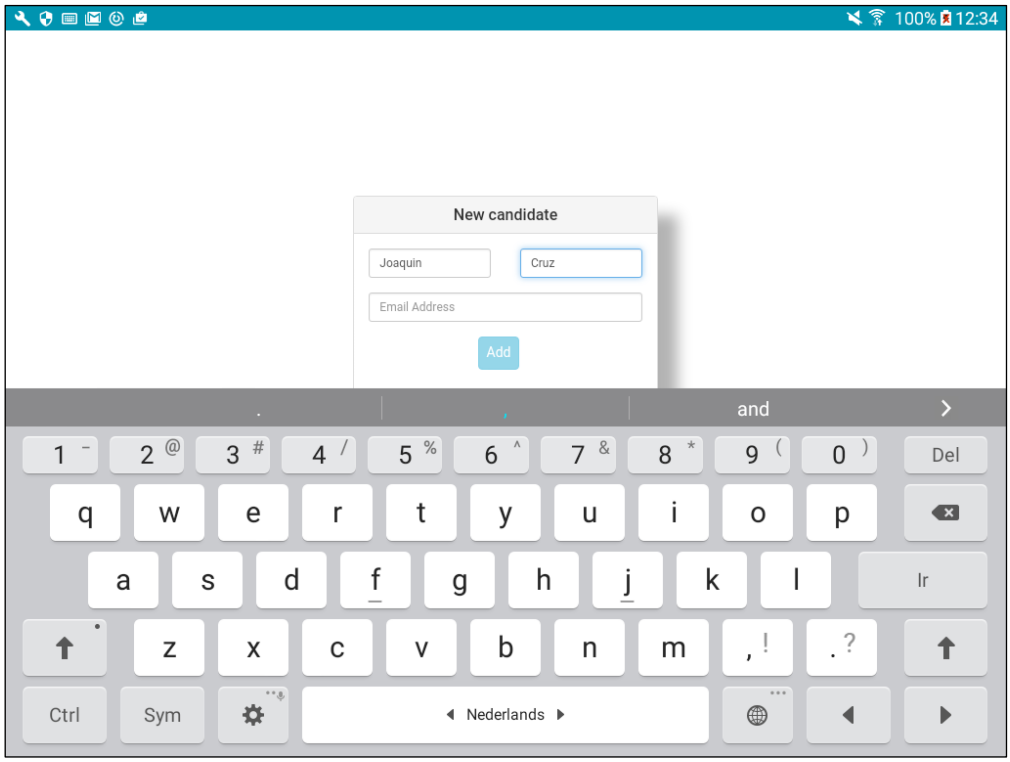

Figure 5.1.4.e. Screen to introduce new candidates in the application

The third and last button in the *Interview Settings* menu (figure 5.4.1.c) is *Manage Candidates* (figure 5.4.1.f). This functionality of Rubrik© was not originally thought of. However, after several tests we became aware of the importance of being able to retrieve the results of different marking sessions over time. This functionality of the application allows us to manage such data from previous sessions. It offers the possibility of browsing through different search criteria to find one particular candidate and edit its details (*i.e.* changing email address, editing mistakes in personal data, etc.).

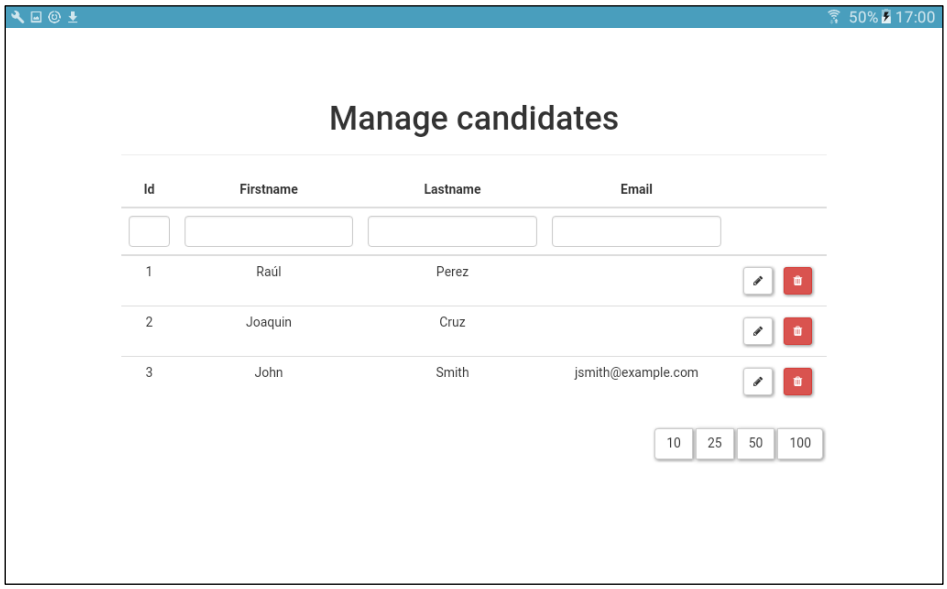

Figure 5.1.4.f. Screen generated by the *Manage Candidates* button

On the right side of the main menu (figure 5.4.1.c) we find the *Application Settings* submenu, whose buttons are displayed in green. The first button in this submenu is *Export Candidates* (figure 5.4.1.g below), which, as its name indicates, can be used to browse through previous sessions to export the data generated after each oral interview. These data can be sorted by date, by candidate name, by the level of the test and by the language in which it was taken.

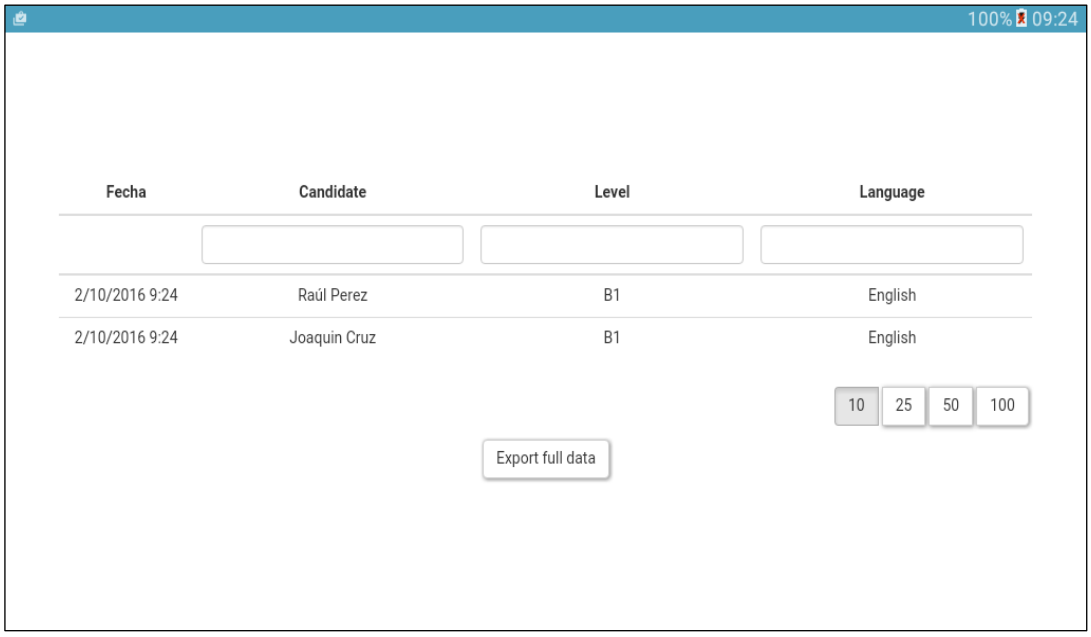

Figure 5.1.4.g. Screen generated by the *Export Candidates* button

The *Export Candidates* button allows us to select which candidates will be exported to a CSV (Comma-separated Values) file. CSV files are, as their name suggests, plain text files in which different values are separated by commas. These files are readable by all word processors and, most importantly, they are readable by spreadsheet software (as Excel) and by statistical analysis packages (such as IBM SPSS (IBM Inc., 2016), Facets (Linacre, 2014) or Winsteps (Linacre, 2016)). This allows the possibility of selecting the data generated during one oral exam to export them to a desktop computer to manage them through Excel or any of the above software packages. This functionality is among the most important contributions of Rubrik© to the work of language testers.

The next button, *Import Candidates* (figure 5.1.4.h), opens one of the most interesting functionalities of Rubrik©, the possibility of preloading big numbers of candidates' names and IDs. The idea that underlies this functionality is for raters to be able to gather and organize the data of candidates prior to interviews and for them to upload such data on the application before the marking session begins. This avoids a considerable amount of time since raters are not obliged to go through the *New Candidate* button (in the blue menu *Interview Settings*, figure 5.4.1.c) before every interview begins.

To import batches of data, Rubrik© offers the possibility of accessing files stored locally in the mobile device. In practical terms, this means that the data of candidates can be first stored locally in the device (through a synchronized cloud storage system, for example) to be used by the application, as shown in figure 5.4.1.h, Rubrik© includes a built-in tutorial explaining how to import batches of data, which can be downloaded from the *Import Candidates* window. In the following 2 figures we show first the *Import Candidates* window (5.1.4.h) and the built-in browsing system (5.1.4.i) that Rubrik© uses to import batches of data previously stored in the mobile device.

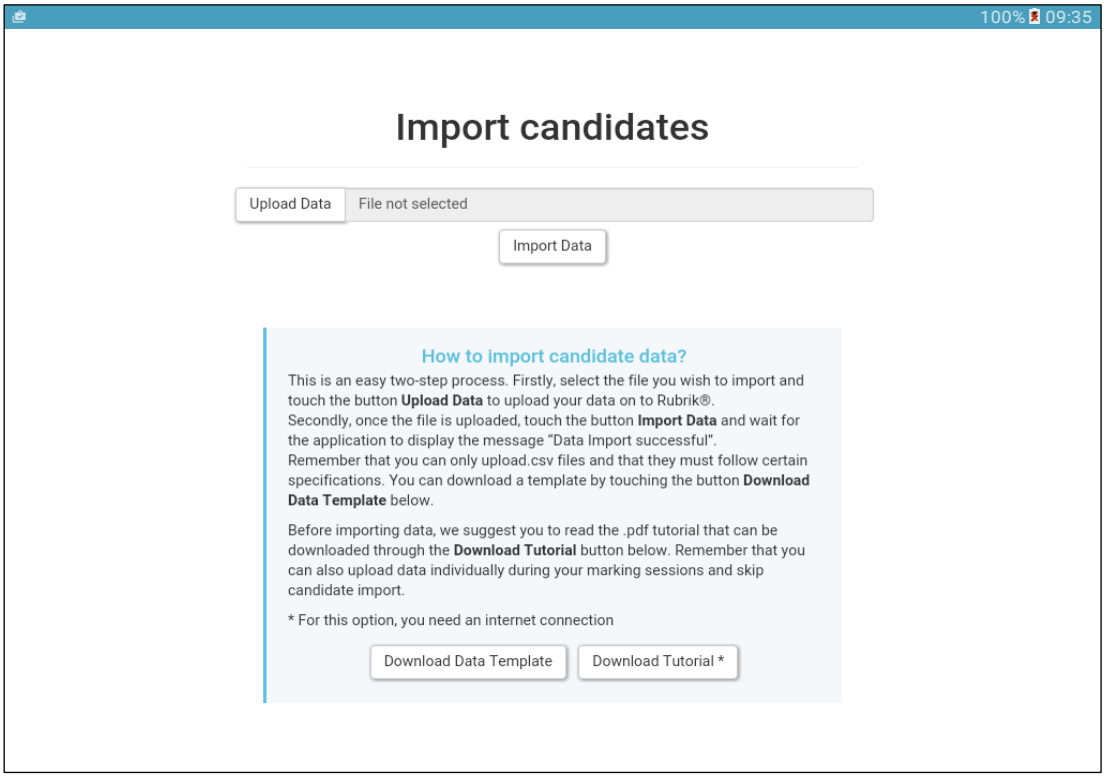

Figure 5.1.4.h. Screen for the *Import Candidates* button

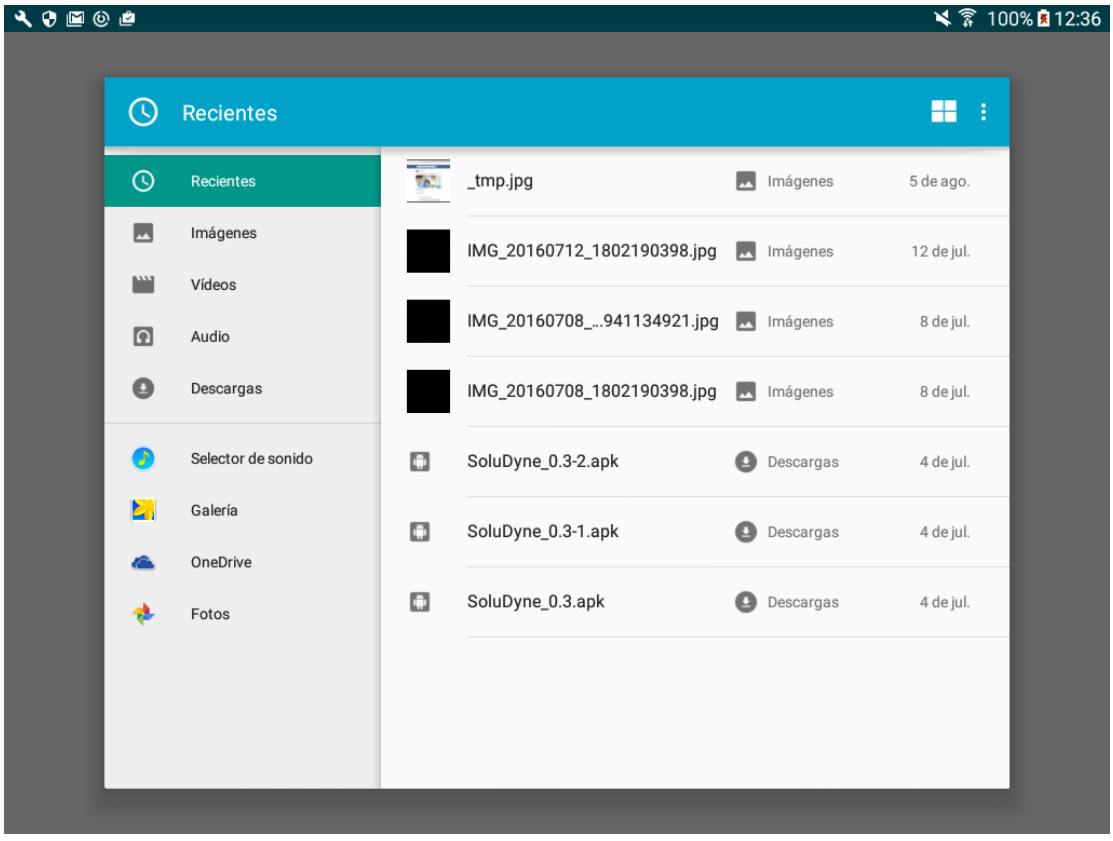

Figure 5.1.4.i. Browsing system to collect data stored locally

Finally, the last button in the *Application Settings* menu (figure 5.4.1.c) is *About*. This button contains general information about the application, about its functionalities and about its development team.

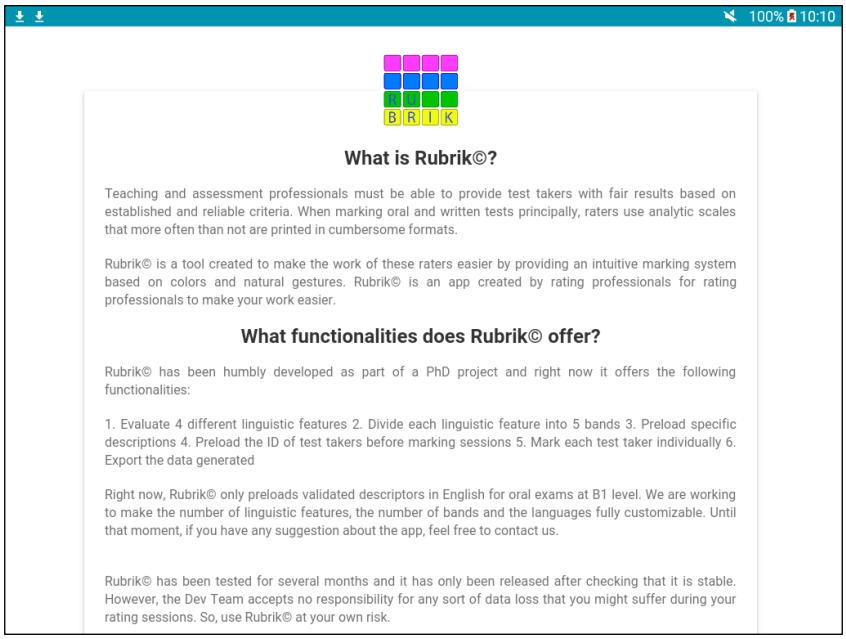

Figure 5.1.4.j. Screen generated by the *About* button

All in all, these UI screens and menus are not the most innovative part of Rubrik©. The most interesting aspect of Rubrik© is the way in which it allows us to manage descriptors and marks during interviews. The main innovation of our application is the way in which descriptors are presented through a cover flow system as showcased in figure 5.1.4.k below.

Figures 5.1.4.k and 5.1.4.l contain captions from early builds of Rubrik©. They still display 5 linguistic features (which, as mentioned in 4.1, was one of the possibilities considered during the first stages of design) against the 4 features that made it to the final build. Along with this, descriptors are not entirely translated into English, as they are in the final build of Rubrik©. Notwithstanding this, they are a representative taster of the way in which the application works. We have chosen to include captions of different development stages to depict more comprehensively the different phases that it has gone through.

Figure 5.1.4.k. Interview screen

At the top of the interview screen in figures 5.1.4.k and 5.1.4.l we see (left to right) in blue the level of the descriptors displayed (A1 in this case), the name of the candidate (Joaquín Cruz) and the timer which controls the interview (stopped at 00:58 seconds in the caption). Right below these we find the heading of the different linguistic features that can be marked. Each of them is linked to a color, which makes it easier for raters to identify which feature they are rating. Each linguistic feature is briefly described to help raters remember the most important aspects that define this particular feature. In figure 4.1.4.k, for example, *Language* is described as "Inteligibilidad, entonación, acento de palabra y sonidos individuales". These descriptions will be customizable in the coming versions of Rubrik©. Right below the definitions we have the different descriptor boxes, with their corresponding colors and which can be visualized back and forth by swiping with one's finger. Below the descriptor boxes we find a field devoted to taking notes about candidates' performance through a virtual keyboard. Last, at the bottom of the screen, we find the 2 tabs that allow instantly swapping between candidate 1 and candidate 2.

There are some hidden functionalities in this screen which are intended to make rating easier. First, we can swap candidates any time by simply touching their corresponding tab at the bottom of the page and their name will be displayed at the top. Second, when any of these candidates is marked for one particular linguistic feature, the chosen band is highlighted by a green frame and the linguistic feature turns from its regular color to dark blue (the same color or the bar with the name of the candidate), to indicate the rater which linguistic features have already been rated. All this can be seen in figure 5.1.4.l below.

Figure 5.1.4.l. Interview screen with band 2 selected

All these functionalities constitute an intuitive and neat UI which makes the task of rating far less tedious.

5.2 Production

The release life cycle of software is commonly divided into 4 stages, the pre-alpha, alpha, beta and gold phases. Over these stages the software undergoes different robustness and stability tests which range from in-house testing by developers during pre-alpha tests to open betas in which the software is made available to wide numbers of testers who will give feedback on its (mal)functioning. Each stage is accompanied by a revision of the code which

leads to an improved build of the software until it reaches the gold phase, the last one, in which the application is made available to the customer.

In the case of Rubrik©, the pre-alpha build was ready 6 months after the first UI design was drafted. The pre-alpha version of the application was initially tested on 1024x768 pixels screens and other common resolutions to check the behavior and stability of the functionalities. As mentioned in section 4.1.4, mock-ups of the application running on common web browsers were distributed among colleagues to obtain initial feedback. After 1 month of testing, different amendments were made in the code and major bugs were corrected to prepare the application for its alpha phase.

During the alpha phase of testing, software is still unstable and can cause crashes and data loss. Set against this, during the alpha phase our application already contained most of the features that were planned for its final build and that was the reason why we decided to show it publicly for the first time. Rubrik© was showcased for the first time ever in the *CEFR* SIG (Special Interest Group) Meeting celebrated by EALTA in May 2016 in Valencia, Spain. In this meeting, a group of experts with a focus on *CEFR* (Council of Europe, 2001) concerns is summoned by EALTA periodically to discuss aspects related to the *CEFR* (*ibid*.). The meeting is attended by experts from all around the world. In the meetings, issues concerning the design, application and implementation of the *CEFR* (*ibid*.) are discussed and that is the reason why we thought that it would be a great opportunity to show Rubrik© for the first time among expert colleagues. This particular meeting took place during the 13th Annual Conference of EALTA, and was attended by +30 testing experts from around the world. It was chaired by Sauli Takala and Neus Figueras, 2 of the researches that collaborated in the development of the *CEFR* (*ibid*.) and who co-authored with Brian North the *Manual for Relating Examinations to the CEFR* (Council of Europe, 2001). They also collaborated in the DIALANG project together with John de Jong, who was also present in the meeting.

Some of the experts present in the meeting, as Mr. Takala, showed particular interest for the application. An anonymous feedback survey was handed out after a 15 minutes' presentation of the application. In the presentation, the main functionalities of the application were explained and its alpha version was shown to the audience.

The survey included questions about the usefulness of the application in different contexts (for entry tests, placement tests, etc.) and about the functionalities shown. The survey also included open questions about prospective improvements for the application and offered the participants to sign up for a testing beta phase.

In the questions about the usefulness of Rubrik©, the participants were asked to mark in a five-level Likert scale how handy they thought it might be in different contexts. The results as to which applications of Rubrik© were seen as more useful are displayed in table 5.2.a below.

Context	Average mark (out of 5)
In general terms	4
For entry tests	3.4
For placement tests	3.5
For end-of-course tests	3.9
For proficiency tests	4.1

Table 5.2. Feedback for Rubrik© from the EALTA *CEFR* SIG Meeting

One of the participants who answered the survey suggested that it could also be used for diagnostic purposes. The results of the table above proved at that stage that, according to experts, the application had potential in different contexts. The application was generally well perceived by this group of international experts for whom it would be most useful in proficiency tests, the type of tests it was actually designed for. The fact that it was perceived as useful for achievement end-of-course tests was also interesting since it showed potential for a wider range of users, lecturers or teachers who are not necessarily enrolled in high-stakes test

development but who need to assess the achievement of their students at the end of a course. Again, this was very good news since it linked the tool to usefulness and practicality, 2 of the main reasons that triggered its design.

The functionalities of the application were evaluated in the survey in a different form. The participants were asked to choose among the functionalities that they considered most helpful. The idea was to establish a rank of priorities in the development of the application, that is to say, to establish which aspects were perceived as more important and thus which aspects should be devoted more time. The rank of functionalities was:

1. Export the data generated for spreadsheet software
2. Evaluate up to 5 dimensions of speech
3. Customizable descriptors
4. Preload ID of test takers
5. Mark test takers individually
6. Up to 5 bands per dimension
7. Quick selection

Finally, in the open-ended questions which the survey also included, the participants made some interesting suggestions, as for example the implementation of a functionality to test candidates remotely through recording or live video using VPN► protocols. Another suggestion was that raters using it could be assigned an ID to keep track of their markings, which could be used in standardization sessions, to check inter and intra-rater reliability or for standard setting. Several participants proposed to include a built-in set of tasks and their scripts that could help raters to conduct interviews. A very interesting proposal was related to diagnostic assessment. One of the participants advocated for the app to offer the possibility of linking the marks of candidates to the specific descriptors that they were marked with, which would yield the possibility of a very interesting type of feedback. Another participant suggested that the

application should include the possibility of recording interviews. This last one was a functionality that had to be discarded at the beginning of the design, as previously mentioned, since it requires internal storage in the device in which the app is installed, which is a shortcoming because it would require an amount of internal memory with which not all mobile devices are built in.

On the downside, 2 of the participants found a tutorial to the application missing and claimed that the ID of the candidate being tested should be more clearly displayed. Both aspects, together with some others, will be discussed as further work in section 6.2.1. All in all, the general perception was that the application clearly had potential and that it could become a useful tool. One of the most striking comments of the answers provided in the survey was that of a colleague who wrote "I did not know that I needed such a tool until I saw it today. I want it for my job". Another expert wrote "I can imagine that high-stakes testing organisations would value such a tool, to simplify and streamline assessment and record keeping". This is precisely the leading idea behind Rubrik©.

It goes without saying that such a positive response from a group of qualified experts was a morale booster. After almost 3 more months of work on the alpha version, Rubrik© went beta on July 27th 2016, the day on which it was uploaded on to Google Play, Google's online platform for the distribution of Android-based applications. The beta release is intended to make a still incomplete version of the application available to the public. Through beta versions, users outside the development team have the opportunity to get to work with the application in real contexts. After these stress tests take place, valuable feedback is obtained to debug the application before a more complete gold version is released. The beta phase lasted 2 months after which different stability issues were corrected.

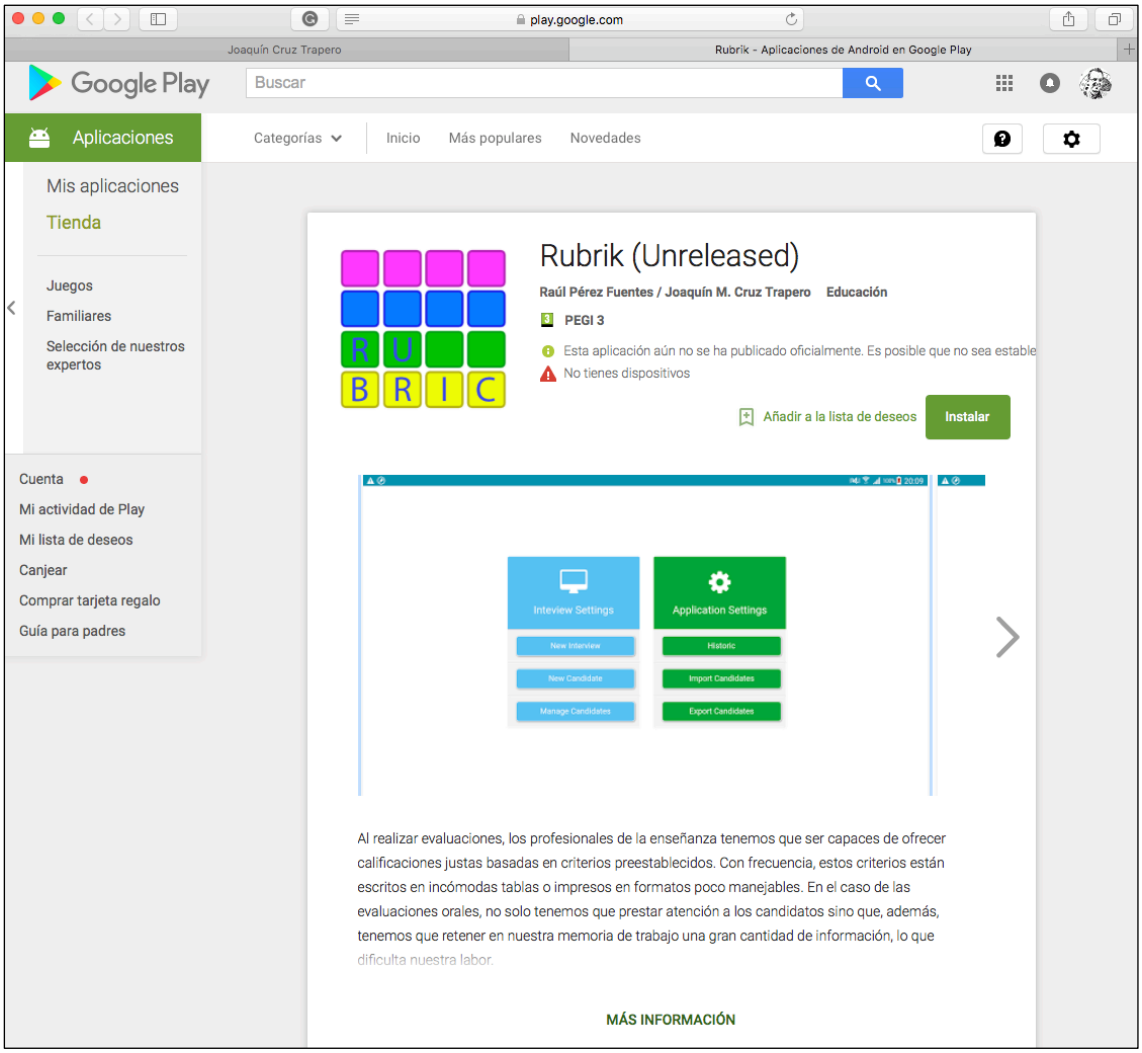

Figure 5.2. Rubrik© on its beta release, July 27th 2016, with typo in the icon

Last, but not least, the application went gold in October 2016. Gold versions are considered to be the final, fully usable build of a piece of software. Gold versions are, nonetheless, subject to revisions and updates. The gold version of the application was fully functional at the moment of printing this dissertation. However, since constant feedback is obtained, new updates of the application are made available regularly. We strongly recommend to update the application frequently.

PART 3

CODA

The game of science is, in principle, without end. He who decides one day that scientific statements do not call for any further test, and that they can be regarded as finally verified, retires from the game.

Popper (2002)

CHAPTER 6. CONCLUSIONS

Chapter 6 ends this dissertation by providing a summary of the previous chapters, by pointing to prospective lines of work and, finally, by describing a series of methodological implications.

In this respect, section 6.1 is a summary of the main points made in the dissertation. It is primarily intended to answer the research questions which were posed in the introduction. Section 6.2 describes some possibilities of further work on different topics. In this section we first go back to theoretical tenets and, once again, we remark the possibilities that biolinguistics offers to embed traditional linguistics in the core of natural sciences. Section 6.2 also points out some improvements from which the protocol and the rubrics created can benefit and, finally, lists a series of improvements for Rubrik© which can extend its usability beyond the Andalusian context. Finally, section 6.3 discusses different methodological implications derived from the present work.

6.1 Concluding remarks

We are aware of the fact that this dissertation is different from the most frequent ones in the field of applied linguistics because it encompasses 3 distant disciplines of research, namely linguistics, statistics and computer sciences. Usually, the 2 halves of a dialogue have more words in common than 2 monologues on the same subject (Skinner, 1957:56) and for this reason we have tried to establish a multidisciplinary conversation to reach the necessary conclusions. Across the pages of this dissertation we have tried to provide a scientific answer to different questions that we consider important regarding the research questions with which we started our work.

First we presented a different approach to linguistics based on the biological nature of human beings. This view, the biolinguistic approach, is not new but yet, it has never been used, up to this day, to define a construct of

language that can be operationalized in language testing. The approach proposed does not provide all the desirable answers but helps in the advance of investigation by pointing to some relevant considerations as, for example, the extent to which the cortical and psychological functions that operate in second languages differ from those of foreign languages. We have also presented the most important challenge that linguistics has to face, the challenge of unification. The more our discipline scatters throughout fragmented branches, the more structure we feel obliged to bring to it. Unification may entail a reduction in the number of disciplines but at the same time pursues, by its own nature, a more usable construct of language. The human faculty of speech starts and ends in the human brain and it is then the human brain that we must turn our eyes to. As linguists we would be doing little good to ourselves if we averted our gaze from this evidence. Thus, in the answer to research question 1, we have not only found the group of disciplines that can help to define the construct of language (see section 1.1), but we have also started to walk the way towards the unification of our discipline through the path of biolinguistics.

We have also presented a construct of testing that ties in with scientific methods by turning observable psychometric abilities into probabilistic mathematical models, which answers research question 2 as to whether there is a scientific method capable of assessing oral levels of proficiency in foreign languages. The Spanish tradition of testing has recently started to move towards psychometrics but there is still a long way ahead to catch up with our European and American colleagues who have been using such methods for decades. We are convinced that twenty years from now psychometrics will be a core part of syllabi at Spanish universities with a serious interest in teaching and testing. It is the best method that we know of to objectivize language teaching and assessment methods and to support our hypotheses. If future proves us wrong and psychometrics does not become central to these syllabi, Spanish universities will have fallen behind in this challenge or they will have discovered more powerful scientific methods.

Leaving aside whatever the future holds, we have proven how psychometrics can be used to measure a tool designed in turn to measure language. We are not the first to validate an analytic scale, but we may be the first to have designed a protocol to create and validate through multi-faceted Rasch models one analytic scale that draws directly from the *CEFR* (Council of Europe, 2001). It is with this protocol that we identify the "scientific method" referred to in research questions 2 and 3. This protocol may become a valuable tool for those testing professionals who seek to define their own rubrics and validate them through well-defined stages. The protocol is, at the same time, scalable, helpful and easy to implement. All this, and more precisely the scalability of the protocol, answers research question 3. Unlike the protocol itself, understanding statistics is not something necessarily easy. That is why we have tried to present the different validation statistics in an accessible way likely to be understood by inexperienced linguists. We have tried to do it straightforwardly to help readers to follow the analysis, and not to distract them with excessively complex formulations.

In an age in which smartphones and tables are ubiquitous, we have integrated our rubrics with an innovative application that allows raters to mark candidates more easily. Again, the characteristics of this application make it scalable and likely to be used in contexts different from the one in which it was envisaged. Rubrik© is the answer to research question 4, an innovative tool that can be used to implement the outcome of the protocol to design rubrics.

There is an underlying achievement in this dissertation which we hope that we have been able to transmit. We have described, almost in real time, the biggest effort that Andalusian universities have made to date in the field of testing. We have described how 9 different institutions whose decisions affect thousands of stakeholders decided to set out for a journey of no return in their quest to unify testing standards and criteria. We have pulled together an informative amount of data which, we hope, will be considered one day as the first log of this journey, once the goal is reached or, who knows, perhaps even earlier.

6.2 Further work

As we learn from the quote by Popper at the beginning of part 3 of this dissertation, science is an enterprise that never ends. On many occasions, relevant research is not only that which provides unequivocal answers to important matters. In fact, this type of research is scarce and, more often than not, relevant advances are rooted in adequate questions rather than in answers.

All in all, in the present dissertation we have provided our share of answers. Yet, some questions remain to be answered and more work remains to be done. To finish, we would like to indicate some pathways for future research and work which we consider relevant. There are 3 areas which we would like to consider in this respect, biolinguistics, additional work on our protocol and extended features for our mobile application.

6.2.1 Biolinguistic concerns

As we mentioned in 1.1, biolinguistics is considered as a young discipline in the making, still in exploratory phase. A truly integrated view of biolinguistics which favors some directions over others is a matter of placing one's bets, a necessary part in any scientific inquiry, and "(i)t is more a matter of gut feelings than anything else" (Boeckx, 2013:320). This perspective is exciting because it leaves great room for innovation, but at the same time makes us feel uneasy as it leaves many important questions unsolved.

In our opinion, defining if, from the biolinguistic perspective, learning and acquisition of languages are the same thing is paramount. The pursue of the paradigm that finally separates or unifies learning and acquisition will not only determine what to assess and how but also will shed light on the ultimate biological foundations of language, which must be the objective of linguistics.

In our opinion, since the biolinguistic mainstream is nowadays centered on the evo-devo distinction, linguists must have their say in the future of their discipline by grounding all their research on biological principles. No linguistic

phenomenon occurs in isolation from the biological nature of humans and, as a consequence, this biological nature is the first to be revisited.

Such revision requires a major change in the paradigm from which most linguistic studies have been traditionally developed. These studies have focused on analyzing the outcomes of the way in which linguistic phenomena materialize. We should devote more time to the study of how human brain generates grammar patterns than to the study of the patterns themselves. The search for an adequate conceptualization of language performance and acquisition is not easy, but the need to broaden the discussion of the origin of both is pressing. A linguist may be a language teacher, but a language teacher is not necessarily a linguist. It is the role of language teachers to explore language patterns as much as it is the role of linguists to explore brain patterns.

6.2.1 Additional improvements to the protocol

The protocol that we have created is already a fully functional tool that has proved its usefulness. However, there are certain stages of the protocol that would benefit from further work.

For example, we have proposed inter-rater reliability tests based on κ statistics but we think that such studies are likely to benefit from studies of intraclass correlation coefficients (Cyr *et al*. 2014:20).

The set of rubrics validated in chapter 4 is now being introduced in most Andalusian universities, which are likely to embrace it quickly due to the fact that many of them have participated in its development and because ACLES' standards have recently swayed to vouch for analytic scales designed and validated in exactly the same way in which our set was created (see section 4.3.3 in ACLES, 2016). Since most Andalusian universities already belong to ACLES (and must thus comply with its quality standards) and the rest are working to join the association, it is just a matter of time before the new set of rubrics penetrates Andalusian higher education language assessment centers.

This penetration will provide extensive qualitative and quantitative data about the set of rubrics. Hopefully, this will help not only to improve the rubrics already created but also to extend them to the *CEFR* (Council of Europe, 2001) levels beyond B1, as we are going to describe in section 6.3.

6.2.1 Extra functionalities for Rubrik©

Rubrik© has been designed with very limited resources. Even so, it has reached a high degree of functionality and has definitely met the goals that triggered its design, making the job of raters of oral exams in language proficiency tests easier. However, as it is presented for this dissertation, the application is only usable in a particular context, the very same Andalusian context for which our rubrics were designed.

Rubrik© is modular from its inception and, consequently, new functionalities are easy to implement. We are currently developing a system to preload remotely customized descriptors. This will definitely extend the number of potential users since the application will be usable in many other contexts in which any user would be able to adapt the descriptors. We are planning to implement this functionality through an FTP (File Transfer Protocol) system that will allow users to modify descriptors from desktop computers.

Along with this, it is our plan to make the number of descriptors and bands customizable so that users can choose how many to include.

Rubrik© has been developed for Android devices but it will surely gain more impact the moment it becomes available for iOS mobile systems too, but this entails additional requirements since developing software for iOS devices is more complex due to the restrictions that Apple applies.

During the last stages of development we also realized that the application was likely to fit in smartphones as well. However, for this, UI must be reinterpreted and programmed in such a way that all the necessary information is displayed in screens of small size. The advantages of this adaptation would also be likely to boost the usability of Rubrik©.

Last, but not least, the feedback obtained during the *CEFR* SIG EALTA Meeting in Valencia (see section 5.2) should also be taken into account to extend the possibilities of Rubrik©.

6.3 Methodological implications

In our opinion, there are 3 main outcomes of the present dissertation which may have important implications: the protocol designed, the first set of rubrics derived from it, and Rubrik©.

On the one hand, the protocol designed in chapter 4 is applicable to virtually any context in which it is necessary to design an analytic scale for the assessment of oral proficiency in languages. Moreover, the protocol and its validation procedures are particularly suited for those contexts in which the *CEFR* (Council of Europe, 2001) has to be the starting point. Thus this easy-to-implement protocol is relevant not only in the Andalusian context but might also be in the Spanish and European ones.

The first set of rubrics derived from the protocol is already being implemented in some Andalusian universities and is expected to be implemented in most others along the academic year 2016-2017. This will be the next major step into the process towards mutual recognition that Andalusian universities started in 2011. As we saw in section 3.3, the number of candidates likely to be tested through the new rubrics is considerable.

Since the protocol is suited to design rubrics for different levels of the *CEFR* (Council of Europe, 2001) we foresee that the current B1 rubrics will be soon followed by rubrics in, at least, B2, C1 and C2 levels (see figure 4.2.2.b). This will create a sound, continuous and comprehensive group of rubrics suitable to assess the most frequently tested *CEFR* (*ibid.*) levels in the Andalusian context. Some fellow workers have even suggested that, with certain adjustments, the protocol might also be used to design rubrics for written production. Of course, this entails rethinking some of the steps of the protocol.

Finally, Rubrik© is perhaps the most conspicuous outcome of the present dissertation. Since it is going to be available for free for a wide audience, it is also likely to carry the most noticeable implications in the mid run. As it is now, it is already fully functional at B1 tests in the Andalusian context for which it was designed. However, if the additional improvements listed in 6.2.2 are finally implemented in further stages of development, Rubrik© will be useful in many other assessment contexts.

Rubrik© was born to meet the needs of a very specific type of user in a simple, fast and intuitive way, and we think that we have created an instrument likely to simplify the (at times) tough job of oral raters in an efficient way.

PART 4

SECTIONS IN SPANISH

TÍTULO

PROTOCOLO PARA DISEÑAR UNA ESCALA ANALÍTICA DE PRODUCCIÓN ORAL BASADA EN EL *MCER* Y SU IMPLEMENTACIÓN A TRAVÉS DE DISPOSITIVOS MÓVILES

ÍNDICE

Agradecimientos iii

Introducción ix

PARTE 1 – ASPECTOS TEÓRICOS

Capítulo 1. EL CONSTRUCTO DEL LENGUAJE 1

1.1. Una definición contemporánea del lenguaje: biolingüística 3

1.1.1. Biología 9
1.1.2. Psicología 11
1.1.3. Neurolingüística 17
1.1.4. Informática 20
1.1.5. Lingüística 23

1.2. Una teoría sobre el dominio de la lengua 30

1.2.1. Adquisición y aprendizaje 31
1.2.2. Competencia y actuación 33
1.2.3. El *MCER* (Council of Europe, 2001) y sus niveles 40

Capítulo 2. EL CONSTRUCTO DE LOS TESTS 45

2.1. Medir, examinar y evaluar 46

2.2. Una visión moderna de los tests 48

2.2.1. Validez 49
2.2.2. Fiabilidad 53
2.2.3. Equidad 56
2.2.4. Sentido práctico 59

2.3. Diferentes tipos de tests 62

2.3.1. Métodos directos, indirectos, analíticos y holísticos 64
2.3.2. Rúbricas: historia y definición 69

2.4. Psicometría 75

2.4.1. Teoría clásica y dependencia de las muestras 79
2.4.2. Teoría moderna y modelos probabilísticos 84
2.4.3. La familia de modelos logísticos de un parámetro y el modelo Rasch de múltiples facetas 91
2.4.4. Facets (Linacre, 2014) 94

Capítulo 3. EL CONTEXTO 101

3.1. Políticas europeas 104

3.1.1. El Consejo de Europa 105
3.1.2. El plan Bolonia 107

3.2. Políticas españolas 117

3.3. Políticas andaluzas 122

PARTE 2 – EL EXPERIMENTO

Capítulo 4. **DISEÑO DE UN NUEVO CONJUNTO DE RÚBRICAS** 139

4.1. Revisión de las rúbricas andaluzas preexistentes 140

4.2. Desarrollo de un protocolo para diseñar rúbricas 152

4.2.1. Fase 1. Consideraciones previas 155
4.2.2. Fase 2. Redactar los descriptores 159
4.2.3. Fase 3. Validación 1 (cualitativa) 169
4.2.4. Fase 4. Validación 2 (cuantitativa) 170
4.2.4.1. Idoneidad de los datos 172
4.2.4.2. *Vertical ruler* 176
4.2.4.3. Utilidad de las categorías de la escala 182
4.2.5. Fase 5. Implementación 186
4.2.6. Fase 6. Revisión 196

Capítulo 5. **APLICACIÓN PARA DISPOSITIVOS MÓVILES: RUBRIK©** 197

5.1. Aspectos de diseño 199

5.1.1. Funcionalidades 199
5.1.2. Programación 201
5.1.3. Análisis de la experiencia de usuario 204
5.1.4. Diseño de la interfaz 205

5.2. Producción 215

PARTE 3 – CODA

Capítulo 6. **CONCLUSIONES** 225

6.1. Observaciones finales 225

6.2. Trabajo futuro 228

6.2.1. Aspectos biolingüísticos 228
6.2.2. Mejoras adicionales al protocolo 229
6.2.3. Funcionalidades adicionales para Rubrik© 230

6.3. Repercusiones metodológicas 231

PARTE 4 – SECCIONES EN ESPAÑOL

Título 235

Índice 237

Introducción 241

Resumen 253

Conclusiones 257

GLOSARIO 259
REFERENCIAS 289
APÉNDICE 311

INTRODUCCIÓN

El matemático Alan Turing, nacido en Londres en 1912, vivió largos periodos de su infancia en la India debido a las comisiones de servicio que su padre, un funcionario británico, se veía obligado a realizar. Tras su vuelta definitiva a Inglaterra Turing estudió en el King's College y se especializó en matemáticas. Se estima que su contribución a descifrar los mensajes encriptados de las Potencias del Eje contribuyó a salvar millones de vidas y a acortar la Segunda Guerra Mundial entre 2 y 4 años.

Cuatro años después del final de la Segunda Guerra Mundial, una fría tarde de invierno, Turing se encontró con el químico y filósofo Michael Polanyi y con el zoólogo y médico en fisiología J. Z. Young en uno de los seminarios de filosofía de la Universidad de Manchester. El objetivo de aquel encuentro era debatir sobre el futuro de la inteligencia artificial, un tema que había suscitado el interés del gran público en occidente tras los éxitos científicos cosechados en este terreno durante la Segunda Guerra Mundial. La conversación que estos 3 investigadores mantuvieron entonces acabaría convirtiéndose en el famoso artículo "Computing Machinery and Intelligence" publicado en 1950 en la revista *Mind* (Turing, 1950). En términos generales, en este artículo, Turing proponía diseñar un test que pudiera responder la pregunta de si las máquinas son capaces de pensar o, más concretamente, un test capaz de dilucidar si una máquina sería capaz de imitar el comportamiento humano sin ser descubierta. La película sobre la vida de Turing, titulada en inglés *The Imitation Game* (Tyldum, 2014) de alguna manera refleja esta idea. El test de Turing y sus implicaciones se encuentran aún hoy entre las premisas más influyentes dentro del ámbito de la investigación en inteligencia artificial.

El de Turing es, no obstante, tan solo un ejemplo moderno de uno de los ejes articuladores de esta tesis doctoral, los tests. Precisamente los tests y otro de los ejes articuladores de nuestra investigación, el lenguaje, se mezclan en el siguiente pasaje bíblico extraído del Libro de los Jueces (12:4-6):

> Jefté reunió a todos los hombres de Galaad y atacó a Efraím. Y los de Galaad derrotaron a los efraimitas, que decían despectivamente: "Vosotros, los de Galaad, sois fugitivos de Efraím, en medio de Manasés". Galaad ocupó los vados del Jordán para cortarle el paso a los Efraimitas. Y cuando un fugitivo de Efraím intentaba pasar, los hombres de Galaad le preguntaban: "¿Tú eres de Efraím?" Si él respondía que no, le obligaban a pronunciar la palabra *Shibolet*. Pero él decía *Sibolet*, porque no podía pronunciar correctamente. Entonces lo tomaban y lo degollaban junto a los vados del Jordán. En aquella ocasión, murieron cuarenta y dos mil hombres de Efraím.

La validez de este test improvisado es, como mínimo, cuestionable. Su impacto, dramático. La equidad de sus resultados está, además, seriamente comprometida por la poca fiabilidad que los jueces, esto es, los galaaditas que decidían quiénes morían y quiénes vivían.

Este puede parecer un caso extremo si bien, como veremos más adelante, hay ejemplos contemporáneos de similar impacto. El 23 de marzo de 2016, por ejemplo, el Tribunal Superior de Justicia del Reino Unido anuló una orden de extradición dictada por Gobierno británico en virtud de la cual se habían deportado previamente a 48.000 personas que habían participado en varias convocatorias de un examen de idiomas concreto. Estos 48.000 estudiantes habían sido, según el Tribunal Superior de Justicia Británico, injustamente detenidos y expulsados del Reino Unido tras haber sido acusados de hacer trampas en uno de los exámenes de lengua que los inmigrantes han de aprobar para lograr sus permisos de residencia en el Reino Unido (Menon, 2016). El escándalo estalló después de que un documental de la BBC asegurase haber destapado actividades fraudulentas en un centro examinador del este de Londres en el que estudiantes extranjeros realizaban el *Test of English for International Communication* (TOEIC) (Ali, 2016). Según el criterio del Tribunal Superior de Justicia del Reino Unido, la investigación que siguió al escándalo tuvo fallos y sus resultados fueron injustamente extrapolados a muchos candidatos, la mayoría

indios, que habrían obtenido sus calificaciones de forma legal. Tal es el impacto que pueden llegar a tener los exámenes de idiomas.

Desde los primeros exámenes de oposición chinos surgidos durante la época de la dinastía Han (201 a. C. – 8 d. C.) (Spolsky, 1995:16) hasta el auge de la moderna industria de los tests de idiomas, pasando por las disputas orales medievales, los tests han adoptado diferentes formas y han sido utilizados para una gran variedad de propósitos. Los tests no solo se usan para comprobar si existen máquinas capaces de pensar o para saber si determinadas personas tienen un nivel concreto en una lengua extranjera. A diario acercamos nuestros labios a la cuchara de sopa para comprobar su temperatura o pesamos naranjas en el supermercado, y si hacemos todo esto es porque observamos que hay cierta información de la que carecemos, porque queremos comparar una cosa con otra o por simple curiosidad, que es al fin y al cabo el origen de toda ciencia.

La lengua, que surgió entre los seres humanos como un método de comunicación basado en ciertos atributos biomecánicos preexistentes, también será, como hemos anticipado, un aspecto fundamental de los próximos capítulos.

Nuestra visión de la lengua combina los mencionados atributos biomecánicos, la observación, la aleatoriedad y la interacción. Desde nuestro punto de vista, el lenguaje ha de ser considerado como un producto exclusivo del ser humano y, por lo tanto, algo que no puede ser concebido sin el ser humano. Nosotros vemos la lengua, al igual que Berwick y Chomsky (2011:20), como un órgano más del cuerpo humano. El fenómeno de la adquisición del lenguaje, el principal interés de la biolingüística, se repite de forma milagrosa generación tras generación y, a pesar de ello, es aún muy poco lo que conocemos sobre cómo se desarrolla la capacidad del habla en nuestros primeros años de vida. Observemos por ejemplo el comportamiento de los recién nacidos. Mientras éstos son amamantados, sus órganos fonatorios descansan en una posición tal que el aire que fluye a través de su laringe es articulado de forma involuntaria como una suerte de vocal intermedia central (similar a la de las voces inglesas “about” o “but”) debido a la posición concreta de la boca y la lengua del bebé. De la misma

manera, los labios del recién nacido se abren y cierran al succionar la leche del pecho materno produciendo accidentalmente la articulación de consonantes nasales y bilabiales (como en "mamá" o "papá"). Aunque no hay forma de corroborarlo, algunos científicos han barajado la hipótesis de que el origen de palabras como "mamá" y "papá" (Jakobson, 1962) está precisamente en este movimiento biomecánico, lo que podría explicar por qué las palabras para "padre" y "madre" son tan similares en algunas comunidades lingüísticas alejadas y sin relación entre sí. Tal vez la primera vez que una madre escucha la combinación accidental de ambos sonidos, consonante y vocal juntas, o cualquiera de las variaciones de "ma" o "pa", está predispuesta a interpretar tales sílabas como una apelación voluntaria de su bebé. De esta manera, sonido y significado se unen aleatoriamente y el comportamiento humano (el de la madre) se ve condicionado por el comportamiento innato del bebé (el sonido que hace al succionar la leche materna) en un ciclo que ha sido perfeccionado a través de la historia de la humanidad hasta llegar a nuestros días. Este es un ejemplo maravilloso de interacción entre los mencionados atributos biomecánicos preexistentes, la observación y la aleatoriedad que, combinados, crean un ejemplo primitivo de lenguaje.

La verdadera naturaleza y complejidad del lenguaje humano están todavía lejos de ser entendidas. Este hecho es de gran importancia para quienes enseñamos lenguas o para quienes estamos al cargo de evaluar cómo progresan los estudiantes en su dominio de dichas lenguas. En este sentido, los expertos han propuesto definiciones elegantes y complejas para describir la manera en que funcionan las lenguas. En los últimos 100 años la medicina y la informática han hecho que sea posible observar *in vivo* lo que ocurre en el cerebro humano mientras se habla. Y, a pesar de ello, a pesar de todos estos avances sin precedentes, aún carecemos de una teoría unificada capaz de dar respuesta al porqué de todo lo relacionado con el lenguaje humano, desde la adquisición del mismo hasta la manera en que la gramática se articula. Ante tal situación es lógico preguntarse si es posible valorar o medir algo, el lenguaje humano, cuya

naturaleza última desconocemos. La respuesta es que *sí* es posible, si bien no conocer la naturaleza última del lenguaje humano nos fuerza a trabajar con suposiciones más que con reglas bien fundamentadas o con procedimientos exactos. Quienes investigan el lenguaje humano con frecuencia trabajan a ciegas y no poseen reglas tales como $g=9,8m/s^2$ que les pudieran permitir valorar qué capacidad lingüística tienen sus alumnos. Sería mucho más sencillo si nuestro conocimiento de una lengua pudiese medirse en kilos pero, desgraciadamente, no es así.

En 1957, B. F. Skinner reflexionaba en su conocido trabajo, *Verbal Behavior* (Skinner, 1957) acerca de por qué en el seno de la ciencia se había negado al comportamiento verbal del ser humano el lugar que este se merece y que, según Skinner pensaba, emana de la psicología. Skinner argumentaba que el origen de esta negación eran ciertos argumentos ficticios que la psicología había tardado en desentrañar (Skinner, 1957:5). De alguna manera, lo mismo se puede decir de las aproximaciones científicas al estudio de la lengua que no emanan directamente de la psicología. Los descriptores y las escalas analíticas mencionadas en la cita de más abajo (Knoch, 2009:12) serán definidos profusamente en los capítulos venideros, pero hemos decidido incluir esta cita en este punto para ejemplificar la sensación de inseguridad que muchos sentimos ocasionalmente al enseñar o evaluar lenguas:

> I often found that the descriptors provided me with very little guidance. On what basis was I meant to, for example, decide that a student uses cohesive devices '*appropriately*' rather than '*adequately*' or that the style of a writing script '*is not appropriate to the task*' rather than displaying '*no apparent understanding of style*'? […] This lack of guidance by the rating scale often forced me to return to a more holistic form of marking where the choice of the different analytic categories was mostly informed by my first impression of a writing script […]. I often felt that this was not a legitimate way to rate and that important information might be lost in this process.

Describir o analizar el lenguaje es algo complejo, y aun lo es más evaluar

sobre la base de abstracciones. Si deseamos sentirnos legitimados para hacer estas valoraciones, en primer lugar hemos de desarrollar una teoría del lenguaje.

Los primeros intentos modernos de desarrollar un método científico aplicado al estudio de la lengua más allá de la gramática datan del primer cuarto del siglo XX. Algunos encuentran estos primeros intentos en el *Cours de Linguistique Générale* (Saussure, 1995) de Saussure y otros lo hacen en el positivismo lógico y en los trabajos de Skinner. En cualquiera de los casos, la imagen era entonces y sigue siendo ahora parcial. Sin lugar a dudas, los mencionados trabajos supusieron un impulso considerable al estudio científico de la lengua que se ha visto intensificado en tiempos recientes. De hecho, no es una exageración decir que hemos aprendido más sobre la lengua en los últimos 25 años que en los varios milenios anteriores (Berwick y Chomsky, 2011:29).

La visión conductual de Skinner ha sido ampliamente superada en nuestros días, pero bien puede ser considerada como el punto de partida gracias al cual la lingüística ha progresado hasta el punto en que hoy se encuentra. Esta progresión ha ayudado a definir los principales intereses de la lingüística y a diferenciar, por ejemplo, la lingüística aplicada de la teórica. También ha dado origen a la psicología cognitiva y a la teoría de principios y parámetros e incluso ha hecho posible aplicar modelos matemáticos al estudio de fenómenos lingüísticos a través de la psicometría. Con todo, a pesar de todos estos avances, hoy en día siguen estando vigentes las dudas manifestadas por Chomsky y otros intelectuales durante la década de 1960 acerca de si la psicología o la lingüística habían alcanzado un nivel de conocimiento teórico que les permitiese construir una "tecnología" de la enseñanza de los idiomas (Lawler y Selinker, 1971:28; Savard y Laforge, 1981:74).

Resulta sorprendente que no exista aún una teoría unificada del estudio de la lengua. Cada vez que un lingüista rasca la superficie se topa con las mismas preguntas que, desde hace siglos, siguen sin respuesta. Lo mismo se aplica a las disciplinas derivadas de la lingüística. En ocasiones el lingüista tiene la sensación

de que dichas preguntas tan solo pueden ser respondidas mediante la creación de una nueva teoría, mediante la creación de una nueva rama de investigación, o aun peor, mediante la creación de una nueva rama de investigación sobre una teoría. Tal heterogeneidad lastra la evolución de una disciplina, la lingüística, cuyo destino más lógico es el de ser estudiada en el seno de las ciencias naturales si atendemos al hecho de que las lenguas son un producto humano y, más concretamente, un producto de la mente humana.

Si hubiera algo parecido a una teoría universal sobre el funcionamiento del lenguaje humano, todos los investigadores y docentes la estarían usando de la misma manera en que el químico usa la tabla periódica. Pero no la hay, lo que hace de nuestra labor una tarea compleja. No obstante, esta complejidad no debería ser utilizada como excusa en los tiempos de avances científicos sin precedentes que ahora vivimos. El futuro normalmente nos da pistas y es hora de que los lingüistas comencemos a seguirlas para anticipar una teoría unificada que explique las bases del funcionamiento del lenguaje humano. El mapeo de los genes de la mosca de la fruta que Sturtevant realizó en 1911 es al descifrado completo del genoma humano lo que el vuelo de los hermanos Wright al programa Apolo que llevó a la humanidad a la luna (NHGRI, 2016). Después de revisar los trabajos de Saussure y Skinner o incluso los más recientes de Chomsky, uno podría pensar que la lingüística apenas ha realizado su primer vuelo de 12 segundos. Es hora de despegar. Parafraseando a Schrödinger (2013), podríamos decir que la lingüística debe recorrer un camino plagado de incertidumbres con la única certeza de que su incapacidad actual como disciplina para resolver preguntas fundamentales no es razón para pensar que tales preguntas no puedan ser resueltas en el futuro.

Este paso adelante en el campo de la lingüística, que supondrá una teoría unificada, impulsará y unificará en algún momento el progreso de todas las ramas de la lingüística, la evaluación entre ellas, y nos ayudará a distinguir las teorías útiles de aquellas que no lo son. Esta tesis doctoral ha sido escrita con el firme convencimiento de que el avance que la lingüística necesita no es solo una

necesidad científica sino también una necesidad social. Nuestro mundo ya no es una realidad estática. Las fronteras culturales están hoy más desdibujadas que nunca. La movilidad internacional presenta retos no solo para los científicos que desean comunicarse de manera eficiente en todo el mundo sino también para quienes buscan encontrar trabajo en el diferentes países o estudiar en universidades extranjeras. Todo ello ha dado lugar a una demanda sin precedentes en los ámbitos de la formación y la acreditación de idiomas en todo el mundo. Estas son las pistas antes mencionadas que hemos de seguir los lingüistas. Quienes concurren a exámenes han dejado de ser exclusivamente estudiantes de lenguas. Los examinandos son ahora ejecutivos de empresas que quieren firmar contratos con empresas extranjeras o personas que desean obtener permisos de trabajo y de inmigración. En este contexto se espera de quienes corregimos y diseñamos exámenes de dominio de lengua que seamos permeables a teorías derivadas de campos tan distintos como la lingüística (que describe el fenómeno del lenguaje) y la psicometría (capaz de medir ciertos atributos psicológicos humanos). El trabajo de correctores y redactores de pruebas de dominio está a la misma vez dirigido y delimitado por necesidades institucionales, económicas, sociales e incluso políticas, frecuentemente antagónicas entre sí (Spolsky, 1955:4).

Desgraciadamente, nuestra tesis doctoral no ofrece respuesta a todos estos problemas. Al contrario de lo que Leonardo Grassi escribió al Duque de Urbino sobre el libro *Hypnerotomachia Poliphili* (Colonna, 2013:66), los capítulos que siguen no encierran la sabiduría de todos los manuscritos antiguos ni contienen la respuesta a los misterios de la naturaleza. Muy al contrario, esta tesis doctoral parte de una ambición mucho más humilde, y para ello se basa en propuestas científicas que, al menos a nuestro parecer, al tiempo que humildes pueden también ser consideradas innovadoras e interesantes. Las páginas que siguen tratarán de dar respuesta a diferentes preguntas:

1. ¿Qué disciplinas pueden ayudar a definir el fenómeno del lenguaje en el contexto de una teoría lingüística no unificada?

2. ¿Existe un método científico que pueda ayudarnos a evaluar un nivel concreto de dominio oral de lenguas?
3. Si el método científico mencionado en la pregunta 2 existe y puede ser validado ¿puede este convertirse en la base de un sistema para evaluar varios niveles de dominio oral de lenguas?
4. Si el método científico descrito en la pregunta 2 existe y puede ser validado ¿puede ser implementado de una manera que trascienda los sistemas tradicionales de evaluación para ayudar así a evaluadores y evaluados?

Es en la respuesta a estas preguntas donde, pensamos, hemos realizado algunas contribuciones relevantes. Como se verá más adelante, en primer lugar hemos contribuido a diseñar un protocolo efectivo para el diseño de escalas analíticas a través de una serie de pasos sencillos. En segundo lugar hemos descrito en tiempo real el proyecto que 9 universidades públicas de la comunidad autónoma de Andalucía, en el sur de España, están llevando a cabo con el objeto de converger en el reconocimiento de sus pruebas de dominio de idiomas. Por último, hemos diseñado un sistema innovador para implementar exámenes orales que deseamos sea útil para la creciente comunidad de evaluadores de lenguas de todo el mundo. Todo lo anterior se articula en 4 partes.

La parte 1 contiene los fundamentos teóricos en los que basamos nuestra respuesta a las preguntas 1 y 2, así como una descripción del contexto en que el objeto de nuestro estudio está inserto. Dentro de esta primera parte, el capítulo 1 está dedicado a profundizar en la idea del constructo de lenguaje mientras que el capítulo 2 define el constructo de evaluación. El objetivo de esta primera parte es definir aquello que pretendemos medir, la habilidad del lenguaje, así como los medios que queremos usar para tal medición. Dado que ni la lengua ni su evaluación son independientes del contexto en el que ocurren, dicho contexto será también definido en el capítulo 3 de esta primera parte.

La parte 2 contiene el experimento que hemos diseñado para responder a las preguntas 3 y 4. El capítulo 4 describe el proceso seguido para diseñar y

validar estadísticamente un nuevo conjunto de rúbricas de producción oral. El capítulo 5 utiliza las rúbricas diseñadas en el 4 para crear una innovadora aplicación digital para dispositivos móviles.

La parte 3 contiene el capítulo 6, en el que se extraen las conclusiones más importantes del presente trabajo, se abordan respuestas concretas a las preguntas planteadas, se plantea el futuro de los resultados obtenidos en la presente investigación y se debaten algunas de las implicaciones metodológicas derivadas de la misma.

La parte 4 no añade contenido nuevo. Esta parte incluye exclusivamente algunas traducciones de secciones de la tesis que, por requisito oficial, han de ser incluidas en el presente trabajo.

Para concluir con la introducción nos gustaría realizar algunas aclaraciones sobre cuestiones formales y sobre las convenciones que se han seguido en la redacción de los distintos capítulos. Obsérvese que los acrónimos tan solo se desarrollan de forma completa la primera vez que aparecen. Además de aquellos casos donde es habitual, la *cursiva* se ha usado para destacar el nombre de los aspectos lingüísticos que nuestras rúbricas evalúan, para el nombre de las tablas del *Marco Común Europeo de Referencia* (Council of Europe, 2001), cuando éstas aparecen en el cuerpo del texto, así como para el nombre de los botones y pantallas de la aplicación móvil. La cursiva también se ha usado para marcar las palabras de lenguas distintas al inglés (inclusive en abreviaturas tales como *ibid*.). Esto no se aplica a la palabra "etcétera". El acrónimo inglés del *Marco Común Europeo de Referencia* (*CEFR*, *Common European Framework of Reference*) ha sido tratado como sustantivo y se ha usado como núcleo de sintagmas nominales o como complemento del nombre del núcleo de otros sintagmas nominales. Dado que en algunos puntos de la tesis el uso de cifras se vuelve muy frecuente, hemos preferido escribir dichas cifras con número en lugar de con letra (así pues, hemos optado por escribir "2 exámenes" y no "dos exámenes"), con la excepción de los casos en que los números aparecen al

comienzo de una frase o como parte de una palabra compuesta (es decir, al escribir en inglés hemos preferido mantener "three-parameter model" en lugar de "3-parameter model"). Los casos ambiguos en que el numeral inglés "one" puede ser considerado como un equivalente enfático del artículo indefinido "a" (Quirk *et al*. 1985:273-274) se han tratado de manera diferente. En estas ocasiones se ha mantenido la forma numérica solo en los casos en los que el sintagma en que se encontraba hacía referencia clara a una cantidad. Las ilustraciones y las tablas están numeradas de acuerdo con el capítulo y sección en las que aparecen, lo que ayuda a ubicarlas más rápidamente. De manera adicional, como se observa en la introducción, esta tesis doctoral contiene un glosario de términos técnicos usados en su redacción que utiliza el mismo corpus de referencias bibliográficas que el resto de la tesis. Dicho glosario se encuentra ubicado después de la parte 4, justo antes de la bibliografía, que está desarrollada según el estilo Chicago. En el glosario se incluyen todos los términos marcados con el símbolo ► a lo largo de los capítulos, un símbolo que, al igual que los acrónimos, tan solo aparecerá la primera vez que sea necesario. El objetivo de este glosario es evitar notas a pie de página que puedan entorpecer el flujo natural de la lectura. Si el lector está familiarizado con alguno de los términos marcados con ►, no es necesario que compruebe su significado, y podrá así establecer su propio ritmo de lectura en función de la familiaridad que tenga con la terminología usada.

RESUMEN

Esta tesis doctoral surge en el seno del Centro de Estudios Avanzados en Lenguas Modernas de la Universidad de Jaén después de que en 2011 los rectores de las 9 universidades públicas andaluzas decidieran firmar un acuerdo para hacer converger sus políticas lingüísticas en materia de acreditación lingüística (*CC*, 2011).

En su proceso de adaptación al Espacio Europeo de Educación Superior, las universidades andaluzas comenzaron alrededor de 2010 a aprobar Memorias de Grado en las que se exigía a los alumnos de la práctica totalidad de sus titulaciones que acreditasen un nivel mínimo de dominio en una lengua extranjera. Hasta el momento en que se firma el mencionado acuerdo, las universidades públicas andaluzas habían desarrollado distintos criterios y pruebas de acreditación de idiomas para sus estudiantes. Con la firma del acuerdo, los rectores de las universidades públicas andaluzas se convierten en pioneros en España al perseguir, antes que el resto, la convergencia y el reconocimiento mutuo de sus certificaciones de dominio de lengua extranjera.

La Universidad de Jaén, por su parte, comienza a desarrollar sus exámenes de acreditación en ese mismo año, en 2011, y pronto es consciente de las limitaciones que se han de superar para que dichos exámenes sean pruebas fiables y equitativas. Como era de esperar, la gran mayoría de las restantes universidades públicas andaluzas se encuentra en una situación similar a la de la Universidad de Jaén, de manera que todas ellas deciden redoblar esfuerzos en su proceso de convergencia. Con esta filosofía se crea un grupo de trabajo compuesto por miembros de las 9 universidades públicas andaluzas, cuya trayectoria es descrita en esta tesis doctoral, que velará por que las diferentes universidades andaluzas se acerquen cada vez más a la deseada convergencia.

En este contexto la Universidad de Jaén detecta una carencia fundamental dentro del proceso de evaluación de dominio de lenguas extranjeras. En su día, los criterios usados para corregir las pruebas de producción oral de estos

exámenes de dominio fueron diseñados de forma intuitiva y jamás fueron validados de forma objetiva, por lo que se podía considerar que, hasta cierto punto, dichos criterios reflejaban exclusivamente la opinión de un grupo de expertos.

Con el objeto de suplir esta carencia, desde la Universidad de Jaén se propone al resto de universidades andaluzas diseñar un conjunto de escalas analíticas (o rúbricas) para las pruebas de producción oral que recojan parte del contenido de las escalas preexistentes en cada una de las universidades andaluzas y que, a la vez, actualicen su contenido y puedan ser validadas de manera científica.

Nada más comenzar a comparar las rúbricas preexistentes se hace evidente que la mayoría de ellas no tienen nada que ver entre sí. En ese momento se llega a la conclusión de que, dado que ninguna de estas otras escalas había sido validada científicamente, combinarlas para más tarde validarlas podría ser una tarea costosa y muy probablemente infructuosa. Por ese motivo se decide abandonar la idea de combinar las rúbricas preexistentes y se comienza a diseñar un nuevo instrumento de medida, unas nuevas rúbricas, partiendo de los descriptores del *Marco Común Europeo de Referencia* (Council of Europe, 2001), que sí se encuentran validados.

En el proceso de diseño de estas nuevas rúbricas, la presente tesis primero analiza el estado en que se encuentra hoy día la lingüística, la disciplina en cuyo seno se estudian las lenguas que las rúbricas han de medir. Además del objeto de estudio, la lengua, también analizamos los instrumentos de medida con los que se pretenden validar las rúbricas que se generen más adelante para llegar así al terreno de la psicometría, la ciencia que estudia cómo medir determinadas características intelectuales y cognitivas del ser humano.

Descritos tanto el objeto de la medida como los sistemas de medición, pasamos a crear un protocolo que, a la postre, permitirá generar rúbricas para todos los niveles del *Marco Común Europeo de Referencia* (Council of Europe, 2001). Este protocolo incluye sistemas psicométricos que permiten validar

científica y objetivamente las rúbricas creadas y convertirlas así en un instrumento fiable. En la validación de las rúbricas se lleva a cabo un experimento en dos fases, en el que las rúbricas creadas son sometidas a dos pruebas distintas. En primera instancia, las rúbricas son usadas en una sesión de evaluación real en el Centro de Estudios Avanzados en Lenguas Modernas de la Universidad de Jaén en las que 2 evaluadores las utilizan para calificar a 84 candidatos. Los resultados de esta primera prueba de validación son positivos y, con el objetivo de confirmarlos, se lleva a cabo una segunda validación en la que intervienen 9 evaluadores expertos de 4 de las 9 universidades públicas andaluzas, que califican a 44 candidatos.

Finalmente, con el objetivo de facilitar la tarea de los profesionales que se dedican a evaluar en pruebas orales mediante rúbricas, presentamos una aplicación programada para dispositivos digitales móviles que permite utilizar dichas rúbricas de una forma sencilla, intuitiva y útil.

CONCLUSIONES

Las conclusiones extraídas de esta tesis doctoral se encuentran estrechamente ligadas a las preguntas que se plantean como punto de partida en la introducción.

En primer lugar, hemos llegado a la conclusión de que la lingüística, como disciplina, se encuentra hoy en día fragmentada en multitud de subdisciplinas que, en su mayoría, se centran en analizar las manifestaciones propias del lenguaje humano y no en analizar aquellos fenómenos que las producen. En nuestra opinión, los fundamentos generales de la lingüística han de reorientarse hacia la biolingüística. La biolingüística, tal y como esta se concibe hoy día, está enfocada al estudio de la adquisición de la lengua dentro del paradigma de la evolución y el desarrollo de la misma. A pesar de esta orientación ostensiblemente distinta a la de la lingüística tradicional, la biolingüística es la única disciplina que ha realizado intentos serios de analizar lenguaje humano en el seno de las ciencias naturales. A nuestro entender, esta última es una perspectiva fundamental y un punto de partida inequívoco: puesto que nuestro lenguaje es un producto exclusivo del ser humano, ha de ser estudiado tomando la biología del mismo como punto de partida.

Posteriormente, en esta tesis doctoral se observa que la psicometría, la ciencia que estudia la manera de medir características intelectuales y cognitivas del ser humano, ofrece las herramientas adecuadas para poder medir de manera objetiva distintos niveles de dominio en la producción oral de lenguas extranjeras aprendidas. Si bien esta afirmación no es una conclusión en sí misma dado que la psicometría no es una ciencia nueva, ni siquiera reciente, sí que se llega a la conclusión de que ciertos análisis psicométricos muy fáciles de aplicar pueden ser integrados en un sencillo protocolo que nosotros hemos usado para la creación de escalas analíticas que permitan medir el mencionado nivel de dominio en una lengua extranjera. En virtud de este protocolo hemos creado una escala analítica que hemos validado a través los resultados obtenidos en 2 ensayos distintos. En el primer ensayo, 2 expertos utilizaron las rúbricas para

calificar a 84 candidatos. Al observar que los resultados de este primer ensayo fueron positivos, decidimos llevar a cabo un segundo ensayo en el que participaron 9 evaluadores de 4 de las 9 universidades andaluzas, que a su vez calificaron a 44 candidatos. Tras obtener resultados positivos en ambos ensayos hemos llegado a la conclusión de que el protocolo creado y los análisis psicométricos que este incluye son adecuados para el desarrollo de rúbricas basadas en descriptores del *Marco Común Europeo de Referencia* (Council of Europe, 2001).

Por último se llega a la conclusión de que es posible crear una herramienta de base digital, Rubrik©, mediante la que las mencionadas rúbricas se aplican de una manera distinta a la tradicional, con el objeto de hacer más sencillo el trabajo de los calificadores de exámenes orales.

GLOSSARY

AAC (*Agencia Andaluza del Conocimiento*, Andalusian Agency of Knowledge): This institution, formerly known as DEVA (see entry) holds the competences of evaluation and accreditation of all activities linked to universities in Andalusia. It is in charge of promoting and managing the evaluation and accreditation of research, development and innovation across Andalusian universities. This agency is also in charge of managing and implementing different programs linked to advanced teaching and plans that promote innovation or teaching in other regions and countries. It promotes technological innovation in Andalusia through participation in European Research and Development and Innovation plans. This agency is particularly relevant for the present PhD because the official exams for which the rubrics were designed had to undergo AAC's quality certification processes.

ACLES (*Asociación de Centros de Lenguas en la Enseñanza Superior*, Association of Language Centers in Higher Education): This association is oriented to promote the learning and knowledge of different languages in Spanish higher education programs and institutions. It is currently developing a process of standardization and mutual recognition of certifications linked to the *CEFR* (Council of Europe, 2001) across Spanish universities. In practical terms, this means that any language certification which seeks to be recognized as valid in Spain must first be certified by ACLES. Together with AAC's (see entry) certification processes, the tests of Andalusian universities also aim at abiding by ACLES standards.

AGAE (*Agencia Andaluza para la Acreditación*, Andalusian Agency for Accreditation): Former name of the AAC (see entry), which is the regional organism in charge of issuing quality certifications.

Analytic scoring (see *holistic scoring*): This is "a method of subjective scoring often used in the assessment of speaking and writing skills, where a separate score is awarded for each of a number of features of a task, as opposed to one global score" (Davies *et al*., 1999:7). In general, it is agreed that analytic scoring allows for more exact diagnostic analysis of performance and "leads to greater reliability as each candidate is awarded a number of scores" (*ibid*.) instead of single, global score. Analytic scoring is criticized on the contention that focusing on specified aspects (or linguistic features, as we label them) may divert raters' attention from an overall consideration of performance.

Android (see *iOS*): It refers to a mobile operating system developed by multinational technology company Google and based on the Linux kernel, another operating system. Android operating system was primarily designed for touchscreen and mobile devices and its user interface is based on direct manipulation through touch gestures. Android has the largest installed base of all operating systems and has outsold iOS, Apple's operating system for mobile devices. Android is very popular among mobile app developers because it allows for greater integration and scalability than iOS and because Google is less restrictive than Apple in the validation processes that mobile applications must undergo before they are uploaded to distribution platforms.

AngularJS (*AngularJS*, 2016): This is an open-code programming framework based on JavaScript (see entry) which is mainly used in the development of web applications (*i.e*., software applications in which the client runs in a web browser such as webmail, online retail sales and auction applications, wikis, etc.). Although AngularJS (*ibid*.) is not primarily designed to develop mobile applications, it helps to alleviate a variety of shortcomings frequently encountered during web development because it provides libraries for database access, templating frameworks and session management. In practical terms, it is used to create early builds of applications which can be run and tested on web browsers.

These early builds can be later compiled, packaged and turned into executable applications very easily.

Apache Cordova (*Apache Cordova*, 2016) (also *PhoneGap*): Apache Cordova is a very popular mobile application development framework. Mobile application development frameworks are software tools that support the design of phone applications. These frameworks are generally specific for one platform or operative system. The main advantage of Apache Cordova (*ibid*.) is that it enables programmers to build applications using generic CSS3 (see entry), HTML5 (see entry) and JavaScript (see entry) code which can be later compiled and exported for different platforms (iOS, Android or Windows Phone). In practical terms, this means that if programmers want to design a mobile application for different devices, through Apache Cordova (*ibid*.) they will only need to use one software development tool instead of one specific tool for each device, which saves a lot of time.

Benchmarking: In assessment, the process by which standards of proficiency are established for particular groups or levels. Agreement on these standards can be reached through different methods. Once the baselines of language skills have been established, these can be compared against other samples of proficiency to provide analysis of levels of language ability, to set targets for language learning, to raise standards of teaching and learning or to set requirements for matriculation in educational programs and requirements for recruitment in the world of business. The most frequent form of benchmarking is that in which real samples of candidates to exams are selected as specimens for the different standards of proficiency. Once these specimens are selected, they are kept to be compared with others.

Broca's area: This is a part of the human prefrontal cortex, made up of Brodmann areas (see entry) 44 and 45, which is connected with speech production. Its discovery was not only important because it linked brain and language but also because it was the first time in which a cortical region was specifically connected with a piece of behavior. Nowadays it has become a staple of popular culture and is "featured in virtually every introductory psychology course and textbook and in many, if not all, introductory anatomy and linguistic courses, and for over a century, it has persisted as the focus of intense research and much debate" (Grodzinsky and Amunts, 2006:xiii).

Brodmann areas: These are considered to be an approximate map of human brain based on physiological and cytological coordinates. Although the physiology, the cytology and the histology of the human brain make it very difficult to delimit exact cortical areas, scientists use approximations for the location of the different regions in the human brain. These regions were originally

> defined based on microscopic patterns of nerve cell bodies (cytoarchitectonics) or myelin (myeloarchitectonics). The most widely used of these is based on numbers published by Korbinian Broadman (1868-1918) in 1909. This system divides the mammalian cortex into 'Brodmann's areas', which correspond to known, separable brain functions and to patterns of connections.
>
> (Woolsey *et al.*, 2003:10)

CA (Conversation Analysis Theory): This approach aims "to provide analytic descriptions of the organization of (inter)action, abstracting from the 'contents' of those (inter)actions" (Have, 2007:39) and "to discover the systematic properties of the sequential organisation of talk, the ways in which utterances are designed to manage such sequences, and the social practices that are displayed and embodied in talk-in-interactions" (Lazaraton, 2001:54). This method to analyze social interactions in conversations among individuals (native and non-native)

was first developed by the American sociologist Harvey Sacks. Having its own notation code, it is extensively used in different fields akin to anthropology. In linguistics, it has proved to be useful in the study of linguistic competence (vs. performance).

CEALM (*Centro de Estudios Avanzados en Lenguas Modernas*, Center for Higher Studies in Modern Languages): This is the center of the University of Jaén (Spain), in which language proficiency tests are developed and marked. In most Spanish universities obtaining a B1 or a B2 certificate in a foreign language is a requisite to complete study degrees and hence the importance of the CEALM in its context. In most cases, centers as the CEALM in Spain, in charge of developing and marking language proficiency exams, do not belong to the philology departments of universities due to the special needs that their tasks require. The CEALM is part of a working group of universities that foster mutual recognition in their exams and that share test specifications. These universities, situated in Andalusia are the universities of Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, Pablo de Olavide and Seville, all of which are in charge of developing and marking proficiency tests in their corresponding institutions.

CEFR (*Common European Framework of Reference*) (Council of Europe, 2001): The idea of a common operative educational framework was envisaged by the Council of Europe (see entry) to foster recognition and international mobility among state members even before the text that this entry refers to was published. After some years of debate and previous work, in 2001 the Council of Europe published this famous document that has become the most important reference for the study and teaching of languages in the history of Europe. The document, developed by expert Brian North, is a compendium of guidelines of good practice and of descriptions of different levels of proficiency in languages. Although it was explicitly not designed as a prescriptive tool, in the absence of a more

authoritative document, it is followed as the main reference as regards language learning and teaching all around Europe and in many other parts of the world. The descriptions of language proficiency levels that it contains are among the most important contributions of the document since, by aligning different procedures and standards to such descriptors, mutual recognition of international certificates and levels can be more easily achieved.

Central Executive (Baddeley, 1988, 2000a, 2000b, 2003): This refers to a view of human cognition system which departs from the traditional view of short-term memory and describes a memory model which is stable, productive and understandable. In his model, Baddeley describes a tripartite system, comprising a supervisory controlling system, the Central Executive aided by 2 slave systems, 1 which is specialized for processing language material, the Articulatory Loop, and the other concerned with visuo-spatial memory, the Visuo-Spatial Sketch Pad. This construct of language processing is interesting in the sense that it helps to conceptualize how different important cognitive process may work to produce languages. The system resembles some aspects of Chomskyan minimalist program (Chomsky, 1997).

Cognitive validity (also *theory-based validity*) (see also *predictive validity*, *concurrent validity*, *content validity*, *construct validity*, *face validity*, *context validity*, *scoring validity* and *consequential validity*): This is defined by Field (2011:69) as the validation that "entails constructing an empirically attested model of the target skill as employed by expert users under non-test conditions; then relating the processes which feature in the model to the specifications of the test under examination". A cognitive validity framework allows to determine "in a systematic way how the various processes which make up performance in a skill are represented, explicitly or implicitly, in the test criteria" (*ibid*.). Cognitive validity seems to be very close to construct validity but the former focuses, as its name suggests, on the cognitive layer of different aspects of linguistic

performance, as for example the different phases of verbal output that Levelt (1999:87) proposed, which included conceptualization, grammatical encoding, morpho-phonological encoding, phonetic encoding, articulation and self monitoring.

Communiqué: This is a French word used to refer to the proceedings of the meetings held by the countries that participate in the biennial conferences that have followed up the *Bologna Declaration* (1999) from its inception.

Competence (see also *performance*): The term refers to speakers' knowledge of the formal system of a language as opposed to the actual use that is made of it (performance). Hymes (1972) made interesting contributions to the notion of competence as well as Canale and Swain (1980), who broke down the knowledge of language into grammatical, sociolinguistic and strategic competence. Bachman (1995) took Canale and Swain's framework and established a linguistic model composed by language competence, strategic competence and psychophysiological mechanisms.

Concurrent validity: (see also *predictive validity*, *content validity*, *construct validity*, *face validity*, *cognitive validity*, *context validity*, *scoring validity* and *consequential validity*) This is the term used when the scores of a test are used to predict a criterion at the same time the test is given (Fulcher and Davidson, 2007:5). Crombach and Meehl (1955:282) defined it as follows:

> The investigator is primarily interested in some criterion which he wishes to predict […]. If the test score and criterion are determined at essentially the same time, he is studying concurrent validity. Concurrent validity is studied when one test is proposed as a substitute for another (for example, when a multiple-choice form of spelling test is substituted for taking dictation), or a test is shown to correlate with some contemporary criterion (*e.g.*, psychiatric diagnosis).

Consequential validity: (see also *predictive validity*, *concurrent validity*, *content validity*, *construct validity*, *face validity*, *cognitive validity*, *context validity* and *scoring validity*): Consequential validity studies the consequences and impact that tests and their scores may have, and comprises aspects such as washback (see entry), student's and teacher's perception, etc. For a wider discussion see Hawkey (2011).

Construct: This important term refers to the conceptualization of an abstract idea by means of which such idea is provided with 2 fundamental properties. On the one hand, the abstract idea enters into theoretical schemes that can be related in different ways to other theoretical schemes, to create a web of such constructs. On the other hand, when an abstract idea becomes a construct it can be observed and measured objectively. As an example, the partially-abstract idea of language will become a construct only if it can be related to other scientific schemes (linguistics, psychometrics, sociology) and if it can be observed and measured (*i.e.* through a test, through a set of rubrics, etc.). The most widely accepted definition of construct arose in the field of psychology and can be found in Kerlinger and Lee (2000:40). There are other interesting references in Fulcher (2007:7) and Davies *et al*. (2006:31).

Construct irrelevant variance: When designing a test, some things that should not appear on it may creep in. If, for example, it is necessary for the test taker to make a mathematical operation (whether small or big) in order to come up with the right answer, the mathematical operation may come in between the observed score and the real proficiency of the test taker. These are the cases in which we speak of construct irrelevant variance.

Construct under-representation: This has to do with whether the test actually does test as a broad range of aspects as it is appropriate. If the test falls short of

resources, items or tasks cannot elicit representative sample of language of language performance, necessary to make judgments on test takers' capacities. In such cases we can speak of construct under-representation.

Construct validity: (see also *predictive validity*, *concurrent validity*, *content validity*, *face validity*, *cognitive validity*, *context validity*, *scoring validity* and *consequential validity*): This refers to the theory on which a test is based. "The construct validity of a language test is an indication of how representative it is of an underlying theory of language learning" (Davies *et al*. 1999:33). Crombach and Meehl (1955:282) first defined it as follows:

> Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined." The problem faced by the investigator is, "What constructs account for variance in test performance?" Construct validity calls for no new scientific approach […]. Construct validity is not to be identified solely by particular investigative procedures but by the orientation of the investigator.

Content validity: (see also *predictive validity*, *concurrent validity*, *construct validity*, *face validity*, *cognitive validity*, *context validity*, *scoring validity* and *consequential validity*): This is "defined as any attempt to show that the content of the test is a representative sample from the domain to be tested" (Fulcher and Davidson, 2007:6). Crombach and Meehl (1955:282) defined it as follows:

> Content validity is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling systematically within this universe to establish the test.

Context validity: (see also *predictive validity*, *concurrent validity*, *content validity*, *construct validity*, *face validity*, *cognitive validity*, *scoring validity* and

consequential validity): In the model proposed by Galaczi and Ffrench (2011:113), context validity analyses the parameters that determine how a task is designed (response format, purpose, weighting, known criteria, order of items or time constraints), how a test is administered (physical conditions, uniformity of administration, security) and the demands of the test in terms of linguistic input and output (channel, discourse mode, length, nature of information, topic familiarity, lexical resources, structural resources and functional resources) and as regards interlocutors in case they are necessary, as for example in oral interviews (speech rate, variety of accent, acquaintanceship, number, gender).

Council of Europe: This is the most important institution that has been in charge of the European construction process since 1949, the year of its foundation. After 2 world wars, the Council of Europe aimed at protecting democracy and human rights, at promoting European unity and at enhancing the rule of law in Europe by fostering cooperation on different matters. Nowadays it has 48 member states, covers approximately 820 million people and it is one of the most important sources of policies in the continent, including linguistic policies.

CTT (Classical Test Theory, also *Classical True Score Measurement Theory*) (see also *MTT*): This is a branch of psychometrics used to analyze various types of statistics in language tests (such as item difficulty, levels of discrimination, the contribution that each item or part of a test makes to the internal reliability of the test, the relationship between various parts of a test or tests, the relationship between test taker characteristics and their performance on a test, etc.). It is still currently used but is frequently criticized for being sample dependent in the sense that the results obtained for one test through CTT may be different from the results obtained for the same test if it is sampled in a different population. Davies *et al*. (1999:22) define it as a theory according to which:

an observed score (on a test) is made up of a true score and an error score. The standard error of measurement of a test is an index of the extent to which the observed score is influenced by the error score. Since the purpose of a test is to achieve reliable observed scores, i.e. as close as possible to true scores, much of the effort put into test construction concerns ways of promoting and estimating test reliability. Although classical theory is still much in vogue, its inability to handle different types of error and its total reliance on the sample under test have been criticized.

Crombach's alpha: In CTT (see entry), this statistic is used as an estimate of the internal reliability and the internal consistency of a psychometric test, that is, it indicates how closely related a set of items are as a group. If a particular language test yields high values of this statistic it means that its items work consistently as a group to measure a particular ability.

CRUE (*Conferencia de Rectores de las Universidades Españolas*, Board of Rectors of Spanish Universities): This Spanish association is composed by 50 public universities and 26 private ones. CRUE is the main interlocutor between Spanish universities and the Spanish government and plays a key role in the developing of laws that affect higher education in its country. It also fosters relationships between society and higher education institutions both at a national and at an international level.

CSS3: This is the third revision of CSS (Cascading Style Sheets) programming, used for describing the visual aspect of digital documents, particularly in web programming. It enables the separation of the document content from the way it is presented in such a way that the layout, colors or fonts of a digital document can be specified. Basically, it allows web programmers to provide the static content of a web site with a particular format or style.

Descriptor: A descriptor is a short statement or sentence that describes a particular ability or characteristic of a language speaker. These descriptors are normally grouped in scales which describe levels of proficiency. If the performance of a speaker matches one or several of these descriptors, the speaker is said to be at the level of proficiency that the descriptors illustrate. Descriptors are frequently "can do" statements or the like. Descriptors are oriented to illustrate what speakers of a language are able to do, not what they are *not* able to do. In analytic scales descriptors are central because their clarity will determine the consistency of the interpretations made of them.

DEVA (*Dirección de Evaluación y Acreditación*, Board of Evaluation and Accreditation): This regional agency was in charge of assessing and certifying the activity of Andalusian universities. The agency fosters research, research and development, and relationships between society and Andalusian higher education institutions at a regional, national and international levels. This agency was responsible for the development of the quality specifications and the assessment of Andalusian higher-education language proficiency tests. Its competences belong now to AAC (see entry).

DGU (*Dirección General de Universidades*, General Board of Universities): At a higher level than DEVA (see entry), among other things, this Andalusian agency coordinates and follows up different joint activities in the field of higher education. It coordinates the most important policies that affect Andalusian universities and controls their budgets. It is also in charge of evaluating the quality of degrees in the Andalusian regional system of higher education.

Discrimination index: This is described as the capacity of test items to differentiate among candidates who have more or less of the trait that the test is designed to measure (Davies *et al*. 1999:96). It is a CTT (see entry) essential

feature of test measurement. A test with high item discrimination indexes is considered to be reliable in the sense that this test is suited to discriminate which candidates possess the abilities that the test aims to measure and which ones do not.

ECTS (European Credit Transfer System): Among European institutions, it is considered to be the standard for comparing the study attainment and performance in higher education. When students of European universities successfully complete a course or subject, they are awarded such credits. The credits are recognized by all higher-education institutions and, as a consequence, the system facilitates transfer and progression of students throughout Europe. One academic year corresponds to 60 ECTS credits that are equivalent to 1,500-1,800 hours of total workload, including tuition and independent work.

EHEA (European Higher Education Area): As described in its own website (EHEA, 2014a), the EHEA "was launched along with the Bologna Process' decade anniversary, in March 2010, during the Budapest-Vienna Ministerial Conference. As the main objective of the Bologna Process since its inception in 1999, the EHEA was meant to ensure more comparable, compatible and coherent systems of higher education in Europe. Between 1999 - 2010, all the efforts of the Bologna Process members were targeted to creating the European Higher Education Area, which became reality with the Budapest-Vienna Declaration of March, 2010. The next decade will be aimed at consolidating the EHEA and thus the current EHEA permanent website will play a key role in this process of intense internal and external communication."

European Research Area: The European Research Area groups different scientific research programs aiming at integrating the scientific resources of the European Union. It mainly focuses on transnational cooperation in the fields of medical,

environmental, industrial and socioeconomic research. It is likened to a research and innovation equivalent of the European Common Market for goods and services. Its main purpose is to increase the competitiveness of European research institutions by bringing them together and encouraging a more inclusive way of work. Increased mobility of knowledge workers and deepened multilateral cooperation among research institutions within European member states are central goals of the European Research Area.

Evo-devo: This term is a clipping of *evolution* and *development*, from the discipline of evolutionary development biology. It is a conceptual framework that tries to shed light on the unsolved question of how new, unprecedented biological traits that constitute evolutionary novelties (as for example human language) originate and survive in the course of species evolution. The evo-devo paradigm "tries to unveil, under an all-embracing conceptual umbrella, the rules and mechanisms which evolution has brought into play over time to generate the past and present biodiversity of life forms" (Baguñà and García-Fernández, 2003:465). The evo-devo discussion is one of the core focuses of attention of biolinguistics.

Facets: Facets (Linacre, 2014) a software package employed to analyze data through different Rasch mathematical models, which belongs to the so-called MTT (see entry) family. In this respect, Facets (*ibid*.) is very similar to Winsteps (Linacre, 2016), with the exception that the former also allows analyses through additional mathematical models, chiefly the multi-faceted Rasch measurement model for persons, items and raters. The benefits of Facets (Linacre, 2014) over Winsteps (Linacre, 2016) is that the former allows us to zero in more than one aspect (or facet) in a given analysis without necessarily using 2 or 3 parametric logistic models.

Face validity: (see also *predictive validity*, *concurrent validity*, *content validity*, *construct validity*, *cognitive validity*, *context validity*, *scoring validity* and *consequential validity*): Face validity is described by Davies *et al*. (1999:59) as the "degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer (such as the candidate taking the test or the institution which plans to administer it". If face validity is accompanied by practical validity, *i.e*., the extent to which institutions and candidates perceive the results of tests as both representative and practical to achieve specific goals (entering a country as immigrant, certifying a level of language to obtain a job or to enter an academic program, etc.), face validity becomes a powerful force driving the choices of candidates.

Facility value (also *difficulty value*): This term refers to the degree of facility or difficulty of a test item, and is calculated on the basis of a group's test overall performance. This statistic is used both in CTT (see entry) and MTT (see entry) as an individual indicator and as a factor of other calculations. In CTT, if in a sample of 100 candidates 74 responded correctly one particular item, this item is said to have a facility value index of 74%.

Foreign language: The term is used in contrast to second language (see entry). A foreign language is the language that is learnt through structured, conscious processes and not intuitively in parallel to the acquisition of a mother tongue. Foreign languages are learnt while second languages are acquired. We normally refer to foreign language learning when we speak about language learning educational programs. On the other hand, second languages are not necessarily learnt through formal instruction but in parallel with the mother tongue or the first language. The distinction between foreign and second language is established to differentiate the cognitive processes that underlie in both.

Gaussian function: In mathematics and statistics, this type of function represents normal distribution (see entry) and is characterized by a symmetric bell curve. Gaussian functions are named after the German mathematician that first described them, Johann Carl Friedrich Gauss.

Gold standard: In testing and benchmarking, this term refers to the standard of performance against which samples can be compared. During a benchmarking session, the gold standard is generally used to exemplify a specific level of proficiency in a particular skill or linguistic feature of such skill.

Holistic scoring (see *Analytic scoring*): is a "type of marking procedure which is common in communicative language testing whereby raters judge a stretch of discourse (spoken or written) impressionistically according to its overall properties" (Davies *et al*., 1999:75). A major advantage of holistic scoring over analytic scoring is that, if properly implemented, it allows for a faster marking pace. On the downside, raters consider holistic scoring less reliable than analytic scoring as the former frequently lacks specificity of details, which may lead 2 different raters to mark the same piece of performance differently. Generally speaking, with holistic scoring raters tend to be happy with the global consideration that they make of a piece of performance, but remain unhappily uncertain about the accuracy or replicability of their assessment.

HTML (Hypertext Markup Language) is the long established standard programming language for describing the architecture, contents and appearance on the World Wide Web. There have been different iterations of this language.

HTML5: is the fifth revision of the well-known HTML (see entry), the long established standard programming language for describing the architecture, contents and appearance on the World Wide Web. Roughly speaking, HTML5 is

the language in which most current websites are coded. Since it allows the use of application programming interfaces, it is particularly well suited to implement WebSQL (see entry) local databases.

ICC (Item Characteristic Curves): In MTT (see entry), this refers to a curve showing the relationship between the probability of a correct response and a person's ability (Davies *et al*., 1999:94).

Inter-rater reliability (see intra-rater reliability): In language testing, the extent to which 2 or more raters consistently rate the same sample of language performance using the same criteria. Given a sample of language performance (a piece of writing or a recorded conversation), if rater A and B use the same standards for marking (*i.e*. the same rubrics), the outcome of their appraisal should be the same regardless of which rater marks the test. Checking levels of inter-rater reliability is paramount for high-stakes tests since all candidates expect to be rated according to their performance and independently from the leniency or severity of raters.

Intra-rater reliability (see inter-rater reliability): In language testing, this refers to the extent to which a particular rater is consistent using particular marking criteria over time or over candidates. If a rater is given one particular sample of performance in January 1st, he is expected to mark it exactly the same if he is given the same sample in June 31st.

iOS (see Android): This is a mobile operating system created by Apple. It is distributed exclusively for Apple devices such as iPhone, iPad and IPod touch. It is among the most popular operating systems in the world, only second to Android (see entry).

IQ tests: These are several standardized tests aimed to assess human intelligence quotient (hence their name). These tests are rooted in the work of Francis Galton and Alfred Binet. The items of these tests may be visual or verbal and they are based on abstract-reasoning problems, arithmetic, vocabulary or general knowledge questions.

JavaScript: This is a type of programming language used to make web pages interactive. Along with HTML (see entry) and CSS (see entry), JavaScript makes up for most of the content that we frequently see on the Internet. While HTML and CSS languages are generally considered as static, the main advantage of JavaScript is that it is dynamic. In this way, by programming through JavaScript we can include dynamic content such as clocks which display the exact time, animations, videos, etc. in the design of a web or an application.

Kappa coefficient (also *κ statistics*): This CTT method is frequently used in clinic medicine to check agreement among medical doctors when they have to diagnose patients. The κ coefficient measures the overall percentage of agreement by means of the expression:

$$K = \frac{P_o - P_c}{1 - P_c}$$

Where P_o is the proportion of observed agreements and P_c is the proportion of agreement expected by chance (Sim and Wright, 2005:258). The resulting value tells us how reliable the measurements of our raters are once chance is removed from the equation. For data to be eligible for this type of analysis, these must follow certain specifications, namely 1) measurements must consist of different repeated trials, 2) each trial must have a discrete number of possible outcomes (ordinal categories) and 3) the trials must be independent in the sense that the outcome of one trial does not affect the outcome of others.

Linguistic universals: These are language patterns or phenomena which occur in all known languages. For example, it has been suggested that if a language has dual number for referring just to 2 of something, it also has plural number for referring to more than 2 (Richards and Schmidt, 2002:294). Recursion (see entry) has also been suggested as a linguistic universal.

Linux: Linux is an operating system like Windows or Mac OS. It is open source and allows certain possibilities that other operating systems do not. Programmers tend to choose this system because it does not lock functionalities from software packages and gives the opportunity of modifying the code easily.

Logistic model (also *logistic curve*): This is a mathematical model whose graphical representation in X and Y axes has "S" shape (sigmoid curve). It is defined by the equation

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

where

e = the natural logarithm base,

x_0 = the x-value of the sigmoid's midpoint,

L = the curve's maximum value, and

k = the steepness of the curve

Notice that other apparently sigmoid curves, for example the cumulative form of the normal distribution, are not logistic as they do not meet the formula above. Facets (Linacre, 2014) provides these logistic curves in its analysis, the ICC curves (see entry).

Logit: The easiest way to understand logits is to consider them as a mathematical ruler along which we can place the measurements taken from raters, items, candidates, etc. to see how these interact. The mathematical principles that underlie Rasch models create one particular such ruler for every measurement made and, thus, the ruler of research A is not necessarily the same as the ruler of research B. In the middle of this ruler we are always going to find an average or 0 measurement. Logits reflect the probability or odds of a response being awarded a particular mark. McNamara (1996:165) writes about the probabilities or odds of a particular response:

> The odds are expressed as a logarithm ('log' for short) of the naturally occurring constant e. We thus speak of the 'log odds' of a response, rather than the odds of a response, and the units of measurement scale constructed in this way are called 'log odds units' or logits (pronounced 'LOH-jits'; stress on the first syllable). The logit scale has the advantage that it is an interval scale – that is, it can tell us not only that one item is more difficult than another, but also how much more difficult it is. The interval nature of the ability measurements means that growth in ability over time can be plotted on the scale; this has attractive implications for the evaluation of the effectiveness of teaching […]. By convention, the average difficulty of items in a test is set at zero logits. Items of above-average difficulty will thus be positive in sign, those of below-average difficulty negative in sign. Ability estimates in turn are related to item difficulty estimates, so that a person of an ability expressed as 0 logits would have a 50 per cent chance of getting right an item of average difficulty.

MTT (Modern Test Theory, also *Item Response Theory* and *Latent Trait Theory*) (see also *CTT*): Most of the definitions of MTT agree on the fact that it relates the probability of an examinee's response to a test item to an underlying ability (Linden and Hambleton, 1997:v; Green, 2013:xii) and on the fact that "it encompasses any mathematical model which attempts to predict observations on a latent variable" (Linacre, 2003:926) (hence its alternative name of "Latent Trait Theory"). MTT can help us to predict how candidates, items, etc. will behave departing from a reduced data set. MTT tries to identify patterns in data which

researchers or test designers can use to draw conclusions, even if such data sets are reduced in size, which is another advantage. Set against CTT, MTT is not sample dependent.

Multinomial distribution: In statistics, a multinomial distribution expresses the probability of the possible results of an experiment after repeated trials in which each trial can result in a specified number of outcomes greater than 2 (this last one would be a binomial distribution). A multinomial distribution can show the odds of obtaining a value when rolling a dice because the dice can land on 1 of 6 possible values every time it is tossed. On top of this, for data to qualify for multinomial distribution analysis, trials must be independent (*i.e.* in our rubrics, the fact that candidate A obtains a particular band does not influence on the band that candidate B obtains), and the probability of each possible result must be constant (*i.e.* potentially, at the beginning of the test, there is equal chance for any candidate to obtain any band).

Naturalistic: In Chomskyan terminology, a naturalistic approach seeks to unify the study of language with the core of other natural sciences.

Non-word (also *pseudoword*)**:** A non-word is a unit of speech that appears to be an actual word in a given language but that has no real meaning. In a way, all the words that we know are non-words to us at some point, at the moment in which we first hear them. These non-words are usually made up to develop cognitive experiments as, for example, the one carried out by Ebbinghaus (1913) to investigate memory and forgetfulness. A very-well known example of non-words are those included in Lewis Carroll's poem *Jabberwocky*, in which the author creates a whole poem through words with no meaning but in which grammatical categories are recognizable.

Normal distribution: This is a type of distribution in which data tend to be around a central value with no bias left or right. In practical terms, it is the same as a Gaussian function (see entry).

Ontogeny (see also *phylogeny*): This is the word used to describe the origin and development of an organism, for example from the fertilized egg to mature form. When we refer to the ontogenesis of speech, we refer to the form in which such ability appears in children and the way in which it evolves until its maturity. In this respect, this term is close to the concept of language acquisition.

Performance (see also *competence*): The actual application of a speaker's competence (or knowledge of rules of language) to language communication.

Phylogeny (see also *ontogeny*): Phylogeny is the evolutionary development and history of a species. A rough example would be the way in which humans evolved from early primates to become *Homo Sapiens*. When we refer to the phylogeny of speech, we refer to the way in which this ability may have evolved from the early vocalic sounds of our ancestors to the complex systems that we utilize nowadays. Notice that we do not refer to phylogeny to define relationships between languages, as historical linguistics would do.

Predictive validity: (see also *concurrent validity*, *content validity*, *construct validity*, *face validity*, *cognitive validity*, *context validity*, *scoring validity* and *consequential validity*) Predictive validity is the term used when the test scores are used to predict some sort of future criterion (Fulcher and Davidson, 2007:5). Crombach and Meehl (1955:282) first defined it as follows:

> The investigator is primarily interested in some criterion which he wishes to predict. He administers the test, obtains an independent criterion measure on the

same subjects, and computes a correlation. If the criterion is obtained some time after the test is given, he is studying predictive validity.

Primary trait scales: Similar to holistic scales, these are developed to obtain more information than regular holistic scales can provide. They predetermine criteria for writing or speaking on a particular topic and, consequently, they are task-specific. When the scale is designed to analyze more than 1 primary trait, some experts speak about multi-trait scales.

Psychometrics: The measurement of psychological traits such as intelligence or language ability through mathematical and statistical procedures. In language testing, psychometrics provides relevant data to analyze the properties of raters, candidates and items.

R: is a free language and software environment for statistical computing and graphics. It is based on S language and provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques. One of its strengths is the ease with which quality plots can be produced, including mathematical symbols and formulae (RDCT, 2016b).

Rater: Raters are qualified professionals who rate the performance of test takers. They frequently rate specific dimensions or overall performance of candidates on the basis of previous marks (as when assigning a final grade to one candidate) or by judging specific aspects of their performance (as when they use rubrics to analyze the oral or written performance of test takers). In this respect, "a distinction is often made between the *marker*, indicating a less skilled role, and the *rater*, which is a role requiring professional training" (Council of Europe,

2001:38). A rater may be a clerical marker or may be not. A rater may be the same person that determines whether an answer is correct or not on the basis of unambiguous keys or may take on the job of a clerical marker to, for example, rank-order a group of test takers and to establish a final grade for them.

Recursion: This term refers to the inherent property of languages that allows them to create infinite sequences through a limited number of elements. One well-known example of recursion is that of the possibility of creating clauses of infinite length: "I know the girl", "I know the girl who lives next door", "I know the girl who lives next door that you are dating", etc.

Reliability: Reliability is the extent to which we can be sure of the fact that a particular measurement device (our tests or rubrics) will yield the same results if they are applied in a consistent fashion to different samples. If we apply test A to person X, the results obtained must be the same that we obtain when we use test A with person Y of the same characteristics of X. Every time we use a thermometer we obtain an accurate value of the temperature of a person at one particular time and that is why we consider this measurement tool accurate.

Second language (see *foreign language*): This term is used to referred to refer to either of the 2 or more languages that a person can learn during natural language development thanks to exposure with 2 or more different languages. Second languages are normally acquired (not learnt) in multilingual contexts without the necessity of formal training. A bilingual person would have a first and a second language while a person that learns one language beyond adolescence would have a first and a foreign language. The distinction between second and foreign language is established to differentiate the cognitive processes that underlie in both.

Scoring validity (see also *predictive validity*, *concurrent validity*, *content validity*, *construct validity*, *face validity*, *cognitive validity*, *context validity* and *consequential validity*): Scoring validity can be understood as a superordinate term for all aspects of test reliability, that is to say, all aspects of the testing process that impinge on the consistency and dependability of test scores, all of which influence on the trustworthiness of the information provided by test results (Taylor and Galaczi, 2011:171).

Short-Term Memory: "A proposed intermediate memory system in which information had to reside on its journey from sensory memory to long-term memory" (Anderson, 1995:461). The use of this term has been replaced in modern literature by working memory (see entry).

Specification file: In Facets (Linacre, 2014), this is the txt file which contains all the necessary instructions for the software to carry out its mathematical calculations. A correct design of the specification file is critical for the accuracy of results.

SPSS: This is a software package (IBM, 2016) originally released in 1968 by SPSS Inc., and later acquired in 2009 by IBM. SPSS is not specifically designed for the study of languages but it is extensively used in the field since it allows different types of psychometric analyses in language tests like facility value, discrimination index or internal reliability. SPSS also analyses bivariate statistics (t-tests, ANOVA, correlations, nonparametric tests, etc.), predictions for numerical outcomes (linear regression) and prediction for identifying groups (factor analysis, cluster analysis, etc.). These types of analyses belong to the so-called CTT family.

Standard setting: is defined by Cizek and Bunch (2007:5) as "the process of establishing one or more cut scores on examinations". For example, if one test creates 2 performance categories, *pass* or *fail*, 1 single cut score will have to be defined, the one that separates the scores that will be considered as *pass* from the scores that will be considered as *fail*. If within the category of *pass* we wanted to create another one for the best candidates, let us say the category *distinction*, then we would have to define 2 cut scores, one to differentiate *fail* from *pass* scores and another one to differentiate between *pass* and *distinction*, and so on and so forth. Standard setting procedures normally require iterations of judgments of expert raters on particular items and their results, which makes it a costly procedure in terms of time and money. To avoid the cost of standard setting, correction boards frequently pre-set a cut score on a particular test and apply it after correction.

Stakeholder: In testing, this term has gained special importance in recent times. The implications of high-stakes testing have led to considering its impact on society from the very beginning of test design. Davies *et al*. (1999:184) define stakeholders as

> all those who have a legitimate interest in the use or effect of a particular test, such as the candidates, their teachers and parents/families, the test constructors and their clients who have commissioned the test, the receiving institutions (including government bodies, eg, (*sic*) Ministries of Education and of Immigration) in the case of a selection test

Textmapping: This is a technique to design listening and reading tests based on the consensus of experts. When only 1 rater designs items to elicit specific information from a text or a recording, the resulting items derive from the opinion of what is prominent according to this particular rater. Through textmapping, the same reading or listening passage is subject to the opinion of at least 3 different

raters. Only when at least 2 of these 3 raters reach consensus on what ideas from the passage are relevant are the items designed. Generally speaking, texts can be mapped to find general and supporting ideas or specific information and details.

UI (User Interface): This is the space (physical or digital) in which interactions between humans and machines occur. A screen is the interface between us and computers or mobile phones. In industrial design and digital development this term also refers to the appearance of the interface and the mechanics that it employs to communicate humans and machines.

Unification: In Chomskyan terms, this refers to the merge of all the disciplines that study language into a single discipline. This discipline would be likely to be studied with the core of natural sciences. The idea is related to the naturalistic approach (see entry) to the study of language that Chomsky proposes. There are different examples in the history of sciences in which different disciplines have unified, being the case of chemistry and physics a recent one (Chomsky, 2000:106).

UX (User Experience): This term refers to the process of designing a satisfactory relationship between users and consumption goods. In digital technologies this is closely linked to aspects of usability, accessibility and pleasure provided by the good (a mobile phone digital application, for example).

Validity (see also *predictive validity*, *concurrent validity*, *content validity*, *construct validity*, *face validity*, *cognitive validity*, *context validity*, *scoring validity* and *consequential validity*): This term refers to the extent to which one assessment, test, etc. measures what it is supposed to measure. In test design (along with other disciplines) this concept is linked to *reliability*.

VPN (Virtual Private Network): This is a type of virtual connection between machines used to share data through public networks as if they were directly connected to each other. If we imagine digital connections as multiple roads that communicate with each other, a VPN is like a tunnel between 2 points in these roads that no other road can cross and that no one can access except those who control the machines at the end of the tunnel. This type of virtual connection is normally used for security reasons.

Washback: (also *backwash*) is defined as the effect of testing on instruction and is said to be either positive or negative (Davies *et al*., 1999:225; McNamara, 1996:23). An example of negative washback would be that of language instruction oriented to pass a test based on outdated constructs or tasks. In the Spanish environment, the instruction received by last-year high school students to pass their English University Entrance exam is mostly an example of negative washback. Since this entrance exam contains no speaking or listening sections, these skills are neglected during the last years of secondary education. On the other hand, washback can be positive when a testing procedure encourages good teaching practices. An example of positive washback can be found in most courses oriented to modern proficiency exams (IELTS, Cambridge suite of exams, TOEFL, etc.). Although with different degrees of validity, all these exams encourage test takers to acquire a balanced mastery of the 4 traditional skills.

WebSQL: is a web page application programming interface for storing data in databases. It is supported by most web browsers and allows website developers to build fully fledged web applications which can store information locally in the device of the user and allow applications to work off-line. It is fully compatible with HTML5 programming protocols.

Winsteps: is a software package (Linacre, 2016) employed to analyze data through different Rasch mathematical models, namely Multiple-Choice, Rating Scales and Partial Credit models. The mathematical models that it uses belong to the so-called MTT family.

Working memory (a modern form accepted for short-term memory): This term is opposed to that of Short-Term Memory in the sense that it perceives this range of memory as active instead of passive, as is the case of the former:

> Working memory is thought of as an active system for both storing and manipulating information during the execution of cognitive tasks such as comprehension and learning. In the influential model of Baddeley, working memory consists of two storage components and a central executive function. The two storage components are the articulatory loop, which holds traces of acoustic or speech-based material for a few seconds (longer if the material is rehearsed) and the visuospatial sketchpad for the storage of verbal and visual information. The central executive is a limited capacity, supervisory attentional system used for such purposes as planning and trouble shooting.
>
> Richards and Schmidt (2002:591)

REFERENCES

AAC. 2013. *Resolución del 18 de marzo de 2013, de la Dirección Gerencia de la Agencia Andaluza del Conocimiento, por la que se hace público el trámite a seguir para la evaluación del procedimiento de acreditación del dominio de lenguas extranjeras en las Universidades Andaluzas*. <https://goo.gl/RW5rI1> (28/09/2016). (Resolution of the Managing Board of the Andalusian Agency of Knowledge which Establishes the Procedure to Certify Foreign Language Proficiency at Andalusian Universities).

Abdullaev, Y. 2006. "Brain activity related to semantic and phonological aspects of language", in F. J. Chen (ed.), *Brain Mapping and Language*. New York (NY): Nova Biomedical, 15–71.

ACLES. 2016. "Guía explicativa al formulario de solicitud de acreditación". Online publication. ACLES. <https://goo.gl/lzo8nC> (10/10/2016). (Guide to the Application Form for Accreditation).

Alderson, C. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, C. 2001. "The shape of things to come: will it be the normal distribution?", in C. Weir, I. Vidaković and E. Galaczi (eds.). *European Language Testing in a Global Context*. Cambridge: Cambridge University Press, 1–26.

Ali, A. (2016, March 29). "Theresa May 'wrongly deported 48,000 students' after BBC Panorama exposes TOEIC scam". *Independent*. <http://goo.gl/YHrBhK> (12/7/2016).

ALTE. 2011. *Manual for Language Test Development and Examining for Use with the CEFR*. Strasbourg: Language Policy Division of the Council of Europe.

ALTE Code of Practice, The. 1994. ALTE. <http://goo.gl/WyEDmr> (12/07/2016)

Álvarez, P. 2016. "El gran examen escolar de la LOMCE, pendiente de los pactos políticos". *El País* 25/01/2016. <http://goo.gl/ITZMjz> (26/09/2016). (The Great School Exam of LOMCE, Pending Political Agreements).

Anderson, J. 1995. *Cognitive Psychology and its Implications*. New York (NY): W. H. Freeman & Co.

AngularJS: Google. 2016 [2010]. *AngularJS*. Computer application (Single-page application framework). <https://goo.gl/QRwvka> (28/09/2016).

Anthoine, E., L. Moret, A. Regnault, V. Sbille and J. Hardouin. 2014. "Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures". *Health and Quality of Life Outcomes* 12: 176. <http://goo.gl/PjuXKg> (26/09/2016).

Apache Cordova: Apache Cordova (developer) and J. Bowser, M. Brooks, R. Ellis, D. Johnson, A. Kadri, B. Leroux, J. MacFadyen, F. Maj, E. Oesterle, B. Whitten, H. Wong, and S. Abdullah (authors). 2016. *Apache Cordova*. Computer application (Mobile development framework). <https://goo.gl/Oizgqp> (28/09/2016).

Arabski, J. and A. Wojtaszek (eds.). 2010. *Neurolinguistic and Psycholinguistic Perspectives on SLA*. Bristol: Multilingual Matters.

Asthana, H. 2015. "Wilhelm Wundt". *Psychological Studies* 60.2: 244–248.

Asimov, I. (1997). *Caves of Steel*. London: Harper Collins Publishers.

Bachman, L. 1995. *Fundamental Considerations in Language Testing*. Hong Kong: Oxford University Press.

Bachman, L. and A. Cohen (eds.). 1998. *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Bachman, L. and A. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. and A. Palmer, 2010. *Language Assessment in Practice*. Oxford: Oxford University Press.

Baddeley, A. 1988. *Working Memory*. Oxford: Oxford University Press.

Baddeley, A. 2000a. "The episodic buffer in working memory". *Trends in Cognitive Science* 4.11: 417–423.

Baddeley, A. 2000b. "Working memory". *Current Biology* 20.4: 136–140. <http://goo.gl/QdPBrH> (26/09/2016).

Baddeley, A. 2003. "Working memory and language: an overview". *Journal of Communication Disorders* 36: 189–208.

Baguñà, J. and J. García-Fernández. 2003. "Preface". *International Journal of Developmental Biology* 47.7/8: 465–705. <http://goo.gl/OYaJxf> (26/09/2016).

Bekoff, M. and P. W. Sherman. 2004. "Reflections on animal selves". *Trends in Ecology and Evolution* 19.4: 176–180.

Berezow, A. 2012. "Why psychology isn't a science". *Los Angeles Times* 13/07/2012. <http://goo.gl/x7BHAB> (26/09/2016).

Bergen Communiqué. 2005. *The European Higher Education Area. Achieving the Goals. Communiqué of the Conference of European Ministers Responsible for Higher Education*. <http://goo.gl/E1ERmw> (20/07/2016).

Berlin Communiqué. 2003. *Realising the European Higher Education Area. Communiqué of the Conference of Ministers responsible for Higher Education*. <http://goo.gl/V79zKv> (20/07/2016).

Bernroider, G. and S. Roy. 2005. "Quantum entanglement of K+ ions, multiple channel states and the role of noise in the brain", in N. Stocks, D. Abbot and R. Morse (eds.), *Proceedings of SPIE Reprint* 5841: 205–214. Austin (TX): Society of Photographic Instrumentation Engineers.

Berwick, R. and N. Chomsky. 2011. "The biolinguistics program: the current state of its development", in Di Scuillo and Boeckx (2011:19-41).

BOE 1970: "Ley 14/1970, de 4 de agosto, General de Educación y Financiamiento de la Reforma Educativa". *Boletín Oficial del Estado*, 4th

August, 1970. 187: 12525–12546. (General Law on Education and on Financial Resources for the Reform of Education).

BOE 1980: "Ley Orgánica 5/1980, de 19 de junio, por la que se regula el Estatuto de Centros Escolares". *Boletín Oficial del Estado*, 27th June, 1980, 154: 14633–14636. (Organic Law Regulating the Statute of Schools).

BOE 1985: "Ley Orgánica 8/1935, de 3 de Julio de 1985, reguladora del Derecho a la Educación". *Boletín Oficial del Estado*, 4th July, 1985, 159: 21015–21022. (Organic Law Regulating the Right to Education).

BOE 1990. "Ley Orgánica 1/1990, de 3 de octubre de 1990, de Ordenación General del Sistema Educativo". *Boletín Oficial del Estado*, 4th October, 1990, 238: 28927-28942. (Organic Law on the General Regulation of the Educational System).

BOE 2002. "Ley Orgánica 10/2002, de 23 de diciembre, de Calidad de la Educación". *Boletín Oficial del Estado*, 24th December, 2002, 307: 45188–45220. (Organic Law on the Quality of Education).

BOE 2006. "Ley Orgánica 2/2006, de 3 de mayo, de Educación". *Boletín Oficial del Estado*, 4th May, 2006, 106: 1758–17207. (Organic Law on Education).

BOE 2007. "Real Decreto 1393/2007, 29 October, por el que se establece la ordenación de las enseñanzas universitarias oficiales". *Boletín Oficial del Estado*, 30th October, 2007, 30: 1–25. (Royal Decree Regulating Official University Teaching).

BOE 2013. "Ley Orgánica 8/2013, de 9 de diciembre, para la mejora de la calidad educativa". *Boletín Oficial del Estado*, 10th December, 2013, 295: 97858–97921. (Organic Law on the Improvement of the Quality of Education).

BOJA 2005. "Acuerdo de 22 de marzo de 2005, del Consejo de Gobierno, por el que se aprueba el Plan de Fomento del Plurilingüismo en Andalucía". *Boletín Oficial de la Junta de Andalucía*, 5th April, 2005, 65: 8–39.

(Agreement of the Council of Government, Sanctioning the Plan for the Promotion of Plurilingualism in Andalusia).

BOJA 2011a. "Orden de 28 de junio de 2011, por la que se regula la enseñanza bilingüe en los centros docentes de la Comunidad Autónoma de Andalucía". *Boletín Oficial de la Junta de Andalucía*, 12th July, 2011, 135: 6–19. (Order Regulating Bilingual Teaching in Schools of the Autonomous Community of Andalusia).

BOJA 2011b. "Orden de 29 de junio de 2011, por la que se establece el procedimiento para la autorización de la enseñanza bilingüe en los centros docentes de titularidad privada". *Boletín Oficial de la Junta de Andalucía*, 12th July, 2011, 135: 20–23. (Order Establishing the Procedure to Authorize Bilingual Teaching in Private Schools).

BOJA 2013. "Orden de 18 de febrero de 2013, por la que se modifican la de 28 de junio de 2011, por la que se regula la enseñanza bilingüe en los centros docents de la Comunidad Autónoma de Andalucía, y la de 29 de junio de 2011, por la que se establece el procedimiento para la autorización de la enseñanza bilingüe en los centros docentes de titularidad privada". *Boletín Oficial de la Junta de Andalucía*, 5th March, 2013, 44: 11–12. (Order Modifying the Order Regulating Bilingual Teaching in Schools of the Autonomous Community of Andalusia, and the Order Establishing the Procedure to Authorize Bilingual Teaching in Private Schools).

BOJA 2015. "Orden de 19 de mayo de 2015, por la que se regula el procedimiento para el reconocimiento de acreditación de los niveles de competencia linguistica en lenguas extranjeras, de acuerdo con el Marco Común Europeo de Referencia para las lenguas, para el profesorado de enseñanza bilingüe en el ámbito de la Comunidad Autónoma de Andalucía". *Boletín Oficial de la Junta de Andalucía*. 9th June, 2015, 109: 8–14. (Order Regulating the Procedure to Recognize Certifications of Levels of Linguistic Competence in Foreign Languages, According to the Common

European Framework of Reference, for Bilingual Teachers in the Autonomous Community of Andalusia.)

Boeckx, C. 2013. "Biolinguistics: fact, fiction and forecast". *Biolinguistics* 7: 316–328.

Bolhuis, J., I. Tattersall, N. Chomsky and R. Berwick. 2014. "How could language have evolved?". *PLoS Biology* 12.8: 316–328.

Bologna Beyond 2010. 2009. *Report on the Development of the European Higher Education Area. Background Paper for the Bologna Follow-up Group Prepared by the Benelux Bologna Secretariat*. <http://goo.gl/7BY7Le> (21/07/2016).

Bologna Declaration. 1999. *Joint declaration of the European Ministers of Education convened in Bologna on 19 June 1999*. <http://goo.gl/MyNJsg> (25/05/2015).

Bologna Policy Forum Statement. 2010. <http://goo.gl/SgfAE2> (21/07/2016).

Brindley, G. 1998. "Describing language development? Rating scales and SLA", in in Bachman and Cohen (1998:112–140).

Brown, G. and G. Yule. 1983. *Teaching the Spoken Language*. Cambridge: Cambridge University Press.

Broca, P. 1861 [2006]. "Comments regarding the seat of the faculty of spoken language followed by an observation of aphemia (loss of speech)", reprinted in Y. Grodzinksy and K. Amunts (eds.), *Broca's Region*. New York (NY): Oxford University Press, 291–304.

Brown, A. 2003. "An examination of the rating process in the revised IELTS Speaking Test". *IELTS Research Reports* 6: 1–30.

Bruce, E. and L. Hamp-Lyons. 2015. "Opposing tensions of local and international standards for EAP writing programmes: who are we assessing for?". *Journal of English for Academic Purposes* 18: 64–77.

Bucharest Communiqué. 2012. *Bucharest Communiqué. Making the Most of Our*

Potential: Consolidating the European Higher Education Area. <http://goo.gl/OpmNUM> (21/07/2016).

Buck, G. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.

Budapest and Vienna Declaration on the European Higher Education Area. 2010. <http://goo.gl/AkOnBH> (21/07/2016).

Canale, M. and M. Swain. 1980. "Theoretical bases of communicative approaches to second language teaching and testing". *Applied Linguistics* 1: 1–47.

Carroll, J. 1968. "The psychology of language testing", in A. Davies (ed.), *Language Testing Symposium. A Psycholinguistic Perspective*. London: Oxford University Press, 46–49.

Catell, J. 1905. "Examinations, grades and credits". *The Popular Science Monthly* 66: 367–378. <https://goo.gl/oO1RBL> (18/07/2016).

CC. 2011. *Convenio de colaboración entre la Universidad de Almería, la Universidad de Cádiz, la Universidad de Córdoba, la Universidad de Granada, la Universidad de Huelva, la Universidad de Jaén, la Universidad de Málaga, la Universidad Pablo de Olavide y la Universidad de Sevilla, para la acreditación de lenguas extranjeras*. <http://goo.gl/Vkl3Xc> (22/07/2016). (Agreement of Collaboration between the University of Almería, the University of Cadiz, the University of Cordoba, the University of Granada, the University of Huelva, the University of Jaen, the University of Malaga, the University Pablo de Olavide and the University of Seville on Accreditation of Foreign Languages).

Chadwick, E. 1864. "Statistics of educational results". *The Museum: a Quarterly Magazine of Education, Literature, and Science* 3: 479–484. <https://goo.gl/BMdbVD> (17/08/2016).

Chee, M. , E. Tan and T. Thiel. 1999. "Mandarin and English single word processing studied with functional magnetic resonance imaging". *The Journal of Neuroscience* 19.8: 3050–3056.

Chen, F. (ed.). 2006. *Brain Mapping and Language*. New York (NY): Nova Biomedical Books.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge (MA): The MIT Press.

Chomsky, N. 1997. *The Minimalist Program*. Cambridge (MA): The MIT Press.

Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*. Cambridge (MA): Cambridge University Press.

CIEP (Centre international d'études pédagogiques). 2008. *Spoken performances illustrating the 6 levels of the Common European Framework of Reference for Languages*. DVD. <http://goo.gl/Ipi0A8> (26/09/2016).

Cizek, G. and M. Bunch. 2007. *Standard Setting*. Thousand Oaks (CA): SAGE Publications.

Clark, J. 1979. "Direct vs. semi-direct tests of speaking ability", in E. Briere and F. Hinofotis (eds.), *Concepts in Language Testing: Some Recent Studies*. Washington (DC): TESOL, 35–49.

Code of Fair Testing Practices in Education. 2004. Washington (DC): Joint Committee on Testing Practices in Education.

Colman, A. 2016 [2015]. *A Dictionary of Psychology*. Online version. Oxford: Oxford University Press. <http://goo.gl/KNwaC1> (11/07/2016).

Consejería de Educación (Junta de Andalucía). 2013. *Guía Informativa para Centros de Enseñanza Bilingüe*. (Guide for Bilingual Schools).

"Cordova Overview". 2016. *Cordova*. Webpage. <https://goo.gl/uwIyTv> (26/09/2016).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2007. *Guide for the Development of Language Education*

Policies in Europe. Main Version. Strasbourg: Council of Europe. <http://goo.gl/Gu9jYS> (26/09/2016).

Council of Europe. 2009. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Strasbourg: Council of Europe. <https://goo.gl/iurZ42> (27/07/16).

Council of Europe. 2011. *Manual for Language Test Development and Examining for Use with the CEFR*. Strasbourg: Council of Europe.

Council of Europe. 2016. Website. <http://goo.gl/fwdWOz> (18/07/2016).

Crombach, L. and P. Meehl. 1995. "Construct validity in psychological tests". *Psychological Bulletin* 52: 281–302.

Cyr, P., K. Smith, I. Broyles and C. Holt. 2014. "Developing, evaluating and validating a scoring rubric for written case reports". *International Journal of Medical Education* 5: 18–23.

Damasio, A. 1996. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York (NY): Gosset/Putnam.

Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley and T. McNamara. 1999. *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Dean, J. 2012. *Rubrics in Language Assessment*. Honolulu (HI): National Foreign Language Resource Center.

Descartes, R. 1840. *Discours de la Méthode*. Edited by A. Lorquet. Paris: Jules Delalain et Cie. (Discourse on the Method).

Descartes, R. 1991. *The Philosophical Writings of Descartes*. Vol. 3. Translated by J. Cottingham, R. Stoothoff, D. Murdoch and A. Kenny. Cambridge: Cambridge University Press.

Deygers, B. and K. Van Gorp. 2015. "Determining the scoring validity of a co-constructed CEFR-based rating scale". *Language Testing* 32.4: 521–541.

Dicolo, J. 2009. "IBM to acquire SPSS, adding to acquisitions". *The Wall Street Journal* 30/07/2009. Online. <http://goo.gl/vXQ0hZ> (12/07/2016).

Dick, P. 1996. *Do Androids Dream of Electric Sheep?* New York (NY): Del Rey.

Di Scuillo, A. and C. Boeckx. 2011. *The Biolinguistics Enterprise. New Perspectives on the Evolution and Nature of the Human Language Faculty*. Oxford: Oxford University Press.

EALTA. 2006. *EALTA Guidelines for Good Practice in Language Testing and Assessment*. <http://goo.gl/ltC8lV> (12/07/2016).

East, M. 2009. "Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing". *Assessing Writing* 14: 88–115.

Ebbinghaus, H. 1913. *Memory. A Contribution to Experimental Psychology*. New York (NY): Teachers College, Columbia University.

Edgeworth, F. 1888. "The statistics of examinations". *Journal of the Royal Statistical Society* 51.3: 599–635.

EHEA 2014a. "How does the Bologna Process Work?". *European Higher Education Area*. Website. <http://goo.gl/5rJATQ> (26/09/2016).

EHEA 2014b. "History". *European Higher Education Area*. Website. <http://goo.gl/X54gCH> (26/09/2016).

Ek, J. van and J. Trim. 1991. *Waystage 1990*. Cambridge: Cambridge University Press.

Extra, G. and K. Yağmur. 2012. *Language Rich Europe. Trends in Policies and Practices for Multilingualism in Europe*. Milan: Cambridge University Press.

Fedorenko, E. and N. Kanwisher. 2009. "Neuroimaging of language: why hasn't a clearer picture emerged?". *Language and Linguistics Compass* 3.4: 839–865.

Ferguson, C. 2015. "'Everybody knows psychology is not a real science': public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public". *American*

Psychologist 70.6: 527–542.

Ffrench, A. 2003. "The development of a set of assessment criteria for Speaking Tests". *Cambridge ESOL Research Notes* 13: 8–16. <https://goo.gl/Y59fLB> (26/09/2016).

Field, A. 2014. *Discovering Statistics Using IBM SPSS Statistics and Sex and Drugs and Rock'n'Roll*. London: SAGE Publications Ltd.

Field, J. 2011. "Cognitive validity", in Taylor (2011: 65–111).

Fleiss, J. 1986. *The Design and Analysis of Clinical Experiments*. New York (NY): Wiley.

Fulcher, G. 2003. *Testing Second Language Speaking*. London: Longman Pearson Education.

Fulcher, G. 2009. "Test use and political philosophy". *Annual Review of Applied Linguistics* 29: 3–20.

Fulcher, G. 2015. "Research timeline. Assessing second language speaking". *Language Teaching* 48.2: 198–216.

Fulcher, G. and F. Davidson. 2007. *Language Testing and Assessment*. New York (NY): Routledge.

Galaczi, E. and A. Ffrench. 2011. "Context validity", in Taylor (2011: 112–170).

Graz Declaration. 2003. *Graz Declaration 2003. Forward from Berlin: the Role of the Universities to 2010 and Beyond*. European University Association. <http://goo.gl/XqsAe1> (20/07/2016).

Green, R. 2013. *Statistical Analyses for Language Testers*. Eastbourne: Palgrave Macmillan.

Green, R. and C. Spoetl. 2011. *Building up a pool of standard setting judges: problems solutions and insights*. PowerPoint presentation of a paper delivered at The Eighth Annual Conference of EALTA, Siena, Italy 5th-8th of May, 2011. <http://goo.gl/Ht8KiY> (27/09/2016).

Grodzinsky, Y. and K. Amunts (eds.). 2006. *Broca's Region*. Oxford: Oxford University Press.

Halback, A. and A. Lázaro. 2015. *La acreditación del nivel de lengua inglesa en las universidades españolas: actualización 2015*. Madrid: British Council. <http://goo.gl/jMotox> (21/07/2016). (The Certification of English Language Levels in Spanish Universities).

Hambleton, R. and R. Jones. 1993. "Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development". *Educational Measurement: Issues and Practice* 12.3: 38–47.

Harris, D. 1989. "Comparison of 1-, 2-, and 3-parameter IRT models". *Educational Measurement: Issues and Practice* 8.1: 35–41.

Hasegawa, M., P. Carpenter and M. Just. 2002. "An fMRI study of bilingual sentence comprehension and workload". *Neuroimage* 15: 647–660. <http://goo.gl/3HRYWu> (26/09/2016).

Have, P. ten. 2007. *Doing Conversation Analysis: A Practical Guide*. London: Sage.

Hawkey, R. 2011. "Consequential validity", in Taylor (2011: 234–258).

Hawkey, R. and M. Milanovic, 2013. *Cambridge English Exams. The First Hundred Years*. Cambridge: Cambridge University Press.

Harris, J. 1973. "Linguistics and language teaching: applications versus implications", in J. Jankowsky (ed.), *Georgetown University Round Table of Languages and Linguistics*. Washington (DC): Georgetown University Press, 11–18.

Herodotus. 1996. *Histories*. Translated by G. Rawlinson. Ware: Wordsworth Classics.

Howatt, A. 2004. *A History of English Language Teaching*. Oxford: Oxford University Press.

Hymes, D. 1972. "On Communicative Competence", in J. Pride, and J. Holmes

(eds.), *Sociolinguistics*. Baltimore (MD): Penguin Books, 269–293.

IBM Inc. 2016. *SPSS Statistics for Windows*. Version 24.0. Computer software programme. New York (NY): IBM Corporation.

Jakobson, R. 1962. *Selected Writings*. The Hague: Mouton & Co.

Janssen, G., V. Meier and J. Trace. 2015. "Building better rubric: Mixed methods rubric revision". *Assessing Writing* 26: 51–66.

Jenkins, L. 2011. "Biolinguistic Investigations: Genetics and Dynamics", in A. Di Scuillo and C. Boeckx (2011:126–134).

Johnson, A. and G. Svingby. 2007. "The use of scoring rubrics: reliability, validity and educational consequences". *Educational Research Review* 2: 130–144.

Kerlinger, F. and H. Lee. 2000. *Foundations of Behavioral Research*. Fort Worth (TX): Harcourt College Publishers.

Khalifa, H. and A. Ffrench. 2009. "Aligning Cambridge ESOL examinations to the CEFR: issues and practice". *Cambridge ESOL Research Notes* 37: 10–14. <https://goo.gl/nJEEuH> (26/09/2016).

Khalifa, H. and A. Salamoura. 2011. "Criterion-related validity", in Taylor (2011: 259–292).

Kim, K., N. Relkin, K. Lee and J. Hirsch. 1997. "Distinct cortical areas associated with native and second languages". *Nature* 388: 171–174. <https://goo.gl/TPtFkI> (26/09/2016).

Klein, D., R. Zatorre, B. Milner, E. Meyer and A. Evans. 1994. "Left putaminal activation when speaking a second language". *Neuroreport* 5: 2295–2297.

Knoch, U. 2009. *Diagnostic Writing Assessment*. New York (NY): Peter Lang.

Krashen, S. 2002 [1981]. *Second Language Acquisition and Second Language Learning*. Internet edition. First printed edition published by Pergamon Press Inc. <https://goo.gl/vSdH9Z> (27/09/2016).

Kunnan, A. 2004. "Test fairness", in M. Milanovic and C. Weir (eds.), *European Language Testing in a Global Context*. Cambridge: Cambridge University Press, 27–48.

Lane, C. 2012. "Congress should cut funding for political science research". *The Washington Post* 04/06/2012. <https://goo.gl/bHCXWm> (27/09/2016).

Lawler, J. and L. Selinker. 1971. "On paradoxes, rules, and research in second language learning". *Language Learning* 21: 27–43.

Lazaraton, A. 2001. "Qualitative research methods in language test development and validation", in M. Milanovic and C. Weir (eds.), *European Language Testing in a Global Context*. Cambridge: Cambridge University Press, 51–71.

Lenneberg, E. 1967. *Biological Foundations of Language*. New York (NY): Wiley.

Leuven and Louvain-la-Neuve Communiqué. 2009. *The Bologna Process 2020. The European Higher Education Area in the New Decade*. <http://goo.gl/P0wfTk> (21/07/2016).

Levelt, W. 1999. "A blueprint of the speaker", in C. Brown and P. Hagoort (eds.), *The Neurocognition of Language*. Oxford: Oxford University Press, 83–122.

Lilienfeld, S. 2012. "Is psychology a science?". *The Conversation*. Blog. <http://goo.gl/I32PI7> (11/07/2016).

Linacre, J. M. 1999. "Investigating rating scale category utility". *Journal of Outcome Measurement* 3.2: 103–122.

Linacre, J. M. 2003. "What is IRT theory? A tentative taxonomy". *Rasch Measurement Transactions* 17.2: 926–927. <http://goo.gl/oXnNzS> (27/06/2016).

Linacre, J. M. 2014. *Facets Rasch Measurement Computer Program*. Chicago: Winsteps.com. Version 3.71.4.

Linacre, J. M. 2016. *Winsteps Measurement Computer Program*. Chicago: Winsteps.com. Version 3.92.1.

Lisbon Convention. 1999. Council of Europe and UNESCO. *Convention on the Recognition of Qualifications concerning Higher Education in the European Region*. Lisbon, 11/04/1997. CETS 165. <http://goo.gl/2i9fzW> (25/05/2015).

Lisbon Declaration. 2007. *Europe's Universities beyond 2010: Diversity with a Common Purpose*. Brussels: European University Association <http://goo.gl/Qp6NQ9> (27/09/2016).

Linden, W. van der and R. Hambleton. 1997. *Handbook of Modern Item Response Theory*. New York (NY): Springer.

London Communiqué. 2007. *Towards the European Higher Education Area: Responding to Challenges in a Globalized World*. <http://goo.gl/51s0Oe> (20/07/2016).

Luoma, S. 2004. *Assessing Speaking*. Cambridge: Cambridge University Press.

Macneilage, P. 2008. *The Origin of Speech*. Oxford: Oxford University Press.

Maddieson, I. 2009. *Patterns of Sounds*. Cambridge: Cambridge University Press.

Marcelo, C., I. Sanz, P. Romero and E. Megía. 2013. *Procedimiento de acreditación de dominio de lenguas extranjeras en las universidades andaluzas*. Sevilla: Agencia Andaluza del Conocimiento, Consejería de Economía, Innovación, Ciencia y Empleo, Junta de Andalucía. <https://goo.gl/JLwbLP> (27/09/2016). (Procedure to Certify Foreign Language Proficiency in Andalusian Universities).

Marian, V., M. Spievey and J. Hirsch. 2003. "Shared and separate systems in bilingual language processing: converging evidence from eyetracking and brain imaging". *Brain and Language* 86: 70–82.

Martínez, A. 2011. *La evaluación de las lenguas*. Granada: Octaedro. (Foreign Language Evaluation).

Matthews, P. 2014. *The Concise Oxford Dictionary of Linguistics*. Online version. Oxford University Press. <http://goo.gl/ZFRKJO> (12/07/2016).

McNamara, T. 1996. *Measuring Second Language Performance*. New York (NY): Longman.

McNamara, T. and R. Adams 1994. "Exploring rater behavior with Rasch techniques". Paper presented at the Annual Language Testing Research Colloquium (Princeton, NJ, March 21-23, 1991). (Also available as ERIC Document Reproduction Service No. ED 345498, <http://goo.gl/qpGYXb> (20/08/2016)).

McNamara, T. and U. Knoch. 2012. "The Rasch wars: the emergence of Rasch measurement in language testing". *Language Testing* 29.4: 555–576.

MCU. 1988. *Magna Charta Universitatum*. <http://goo.gl/ksVU0T> (25/05/2015).

Menon, P. 2016. "48,000 students wrongly deported from the U.K.". *The Hindu* 29/03/2016. <http://goo.gl/jmzs5i> (20/07/2015).

Messick, S. 1992. *The Interplay of Evidence and Consequences in the Validation of Performance Assessments*. Princeton (NJ): Educational Testing Service.

Nebrija, A. de 1984. *Gramática de la lengua castellana*. Edited by A. Quilis. Madrid: Editorial Nacional. (Grammar of the Castilian Language).

New Oriental. 2015. Website. <https://goo.gl/boN1I7> (29/09/2016).

NHGRI (National Human Genome Research Institute). 2016. "An overview of the Human Genome Project". Last modified May 2016. <https://goo.gl/Eytu5t> (27/09/2016).

North, B. 2000. *The Development of a Common Framework Scale of Language Proficiency*. New York (NY): Peter Lang.

North, B. and G. Schneider (1998). "Scaling descriptors for language proficiency scales". *Language Testing* 15.2: 217–263.

Osterhout, L., A. Poliakov, K. Inoue, J. McLaughlin, G. Valentine, I. Pitkanen, C. Frenck-Mestre and J. Hirschensohn. 2008. "Second-language learning and changes in the brain". *Journal of Neurolinguistics* 21: 509–521.

PADLE. 2015. *Procedimiento de acreditación de dominio de lenguas extranjeras (gestión, elaboración y evaluación de exámenes de dominio y reconocimiento de acreditaciones externas)*. Jaén: CEALM. <https://goo.gl/Ps8Nn2> (23/07/2016). (Procedure to Certify Foreign Language Proficiency (Management, Design and Evaluation of Proficiency Exams and Recognition of External Certifications)).

PAI. 2011. *Propuestas sobre la acreditación de idiomas. Informe elaborado por la "Comisión para el análisis y estudio de la acreditación y formación en idiomas" de la CRUE (Conferencia de Rectores de las Universidades Españolas) y aprobado en la Asamblea General de la CRUE (Santander, 8 de septiembre de 2011)*. <http://goo.gl/N0fqfn> (21/07/2016). (Proposal about the Certification of Languages. Report Drafted by the "Board for the Analysis and Study of Language Certification and Teaching" of the CRUE (Board of Rectors of Spanish Universities) and Sanctioned in the General Meeting).

Panadero, E. and A. Jonsson. 2013. "The use of scoring rubrics for formative assessment purposed revisited: a review". *Educational Research Review* 9: 129–144.

Papageorgiou, S., R. Tannenbaum, B. Bridgeman and Y. Cho. 2015. *The Association between* TOEFL IBT® *Test Scores and the Common European Framework of Reference (CEFR) Levels*. (ETS Research Memorandum 15-06). Princeton (NJ): Educational Testing Service. <https://goo.gl/l40RBh> (27/09/2016).

Piantadosi, S. and E. Gibson. 2013. "Quantitative standards for absolute linguistic universals". *Cognitive Science* 38.4: 736–756.

Pinker, S. 1995. *The Language Instinct*. London: Penguin.

Plato. 2001. *Cratylus*. Translated by B. Jowett. South Bend (IN): Informotions Inc. <http://goo.gl/G4m77J> (16/07/2016).

Popper, K. 2002. *The Logic of Scientific Discovery*. London: Routledge Classics.

Prague Communiqué. 2001. *Towards the European Higher Education Area. Communiqué of the meeting of European Ministers in charge of Higher Education in Prague on May 19th 2001*. <http://goo.gl/jOfs4H> (20/07/2016).

QuarkXPress: *Quark, Inc*. 2016 [1987]. *QuarkXPress*. Computer application (Desktop Publishing). <https://goo.gl/wSj4hk> (27/09/2016).

Quirk, R., S. Greenbaum, G. Leech, J. Startvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Reavis, G. 1999. *The Animal School*. Peterborough: Crystal Springs Books.

Reise, S., A. Ainsworth, and M. Haviland. 2005. "Item Response Theory. Fundamentals, applications, and promise in psychological research". *Current Directions in Psychological Science* 14.2: 95–101.

RDCT (R Development Core Team). 2016a. *R Computer Program*. Auckland: GPL <https://goo.gl/PceqSX> (27/09/2016).

RDCT (R Development Core Team). 2016b. *The R Project for Statistical Computing*. Auckland: RDCT <https://goo.gl/ipqBN7> (27/09/2016).

Richards, J. and R. Schmidt. 2002. *Longman Dictionary of Language Teaching and Applied Linguistics*. London: Longman.

Robins, R. 1997. *A Short History of Linguistics*. Thetford: Routledge.

Robinson, A. 2003. *The Story of Writing*. London: Thames & Hudson.

Roux, F. and V. Lubrano. 2006. "Language studied by electrostimulation in bilingual patients", in F. Chen (ed.), *Brain Mapping and Language*. New York (NY): Nova Biomedical Books, 111–133.

Saussure, F. de. 1995 [1916]. *Cours de linguistique générale*. Edited by C. Bailly and A. Séchehaye, with the collaboration of A. Riedlinger. Critical edition by Tullio De Mauro. Paris: Éditions Payot.

Savard, J. and L. Laforge (eds.). 1981. *Proceedings of the 5th Congress of*

L'association Internationale de Linguistique Appliquée. Montréal: Les presses de l'Université Laval.

Schrödinger, E. 1935. "Die Gegenwärtige Situation in der Quantenmechanik". Naturwissenschaften 23: 807–812. (The Present Situation in Quantum Mechanics).

Schrödinger, E. 2013. *What is life?* Croydon: Cambridge University Press. <http://goo.gl/nXouWg> (15/08/2016).

Second Bologna Policy Forum Statement. 2010. <http://goo.gl/FAjqMR> (21/07/2016).

Sim, J. and C. Wright. 2005. "The Kappa Statistic in seliability studies: use, interpretation and sample size requirements". *Physical Therapy* 85.3: 257–268.

Simos, P., E. Castillo, J. Fletcher, D. Francis, F. Maestu, J. Breier, W. Maggio and C. Papanicolaou. 2001. "Mapping of receptive language cortex in bilingual volunteers by using magnetic source imaging". *Journal of Neurosurgery* 95.1: 76–81.

Skinner, B. F. 1957. *Verbal Behavior*. New York (NY): Appleton-Century-Crofts.

Smith, A., R. Rush, L. Fallowfield, G. Velikova and M. Sharpe. 2008. "Rasch fit statistics and sample size considerations for polytomous data". *BMC Medical Research Methodology* 8. Online publication. <https://goo.gl/i2mTBU> (28/09/2016).

Sorbonne Declaration. 1998. *Sorbonne Joint Declaration. Joint Declaration on Harmonisation of the Architecture of the European Higher Education System by the four Ministers in charge for France, Germany, Italy and the United Kingdom*. Paris, the Sorbonne, May 25 1998. <http://goo.gl/vO3JV8> (25/05/2015).

Spolsky, B. 1995. *Measured Words*. Oxford: Oxford University Press.

Statement by the Bologna Policy Forum. 2009. <http://goo.gl/RUYg6X> (27/07/2016).

Statement of the Fourth Bologna Policy Forum. 2015. <http://goo.gl/Es3uN8> (21/07/2016).

Statement of the Third Bologna Policy Forum. 2012. *Beyond the Bologna Process: Creating and Connecting National, Regional and Global Higher Education Areas*. <http://goo.gl/BwE6r1> (21/07/2016).

Stemler, S. 2004. "A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability". *Practical Assessment, Research and Evaluation* 9.4. Online publication. <https://goo.gl/RgqQOX> (29/07/2016).

Taylor, L. 2003. "The Cambridge approach to speaking assessment". *Research Notes* 13: 2–4. <https://goo.gl/TTcT93> (27/09/2016).

Taylor, L. (ed.) 2011. *Studies in Language Testing. Examining Speaking*. Cambridge: Cambridge University Press.

Taylor, L. and E. Galaczi. 2011. "Scoring Validity", in Taylor (2011: 171–233).

Thorndike, E. 1910. "Educational measurements of fifty years ago". *Journal of Educational Psychology* 4: 551–552.

Thurstone, L. 1928. "Attitudes can be measured". *American Journal of Sociology* 33: 529–554.

Turner, C. 2000. "Listening to the voices of rating scale developers: identifying salient features for second language performance assessment". *The Canadian Modern Language Review* 56.4: 555–584.

Turing, A. 1950. "Computing machinery and intelligence". *Mind* 49: 433-460.

Tyldum, M. (dir.) 2014. *The Imitation Game*. Film. Nora Grossman.

UCLES. 2012. *IELTS Guide for Teachers. Test Format, Scoring and Preparing Students for the Test*. <https://goo.gl/OzQdTO> (24/07/2016).

Vidaković, I. and E. Galaczi. 2013. "The measurement of speaking ability 1913-2012", in Weir *et al*. (2013:257-346).

Weigle, S. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C., I. Vidaković and E. Galaczi. 2013. *Measured Constructs: A History of Cambridge English Language Examinations 1913-2012*. Cambridge: Cambridge University Press.

Wilkins, D. 1976. *Notional Syllabuses*. Oxford: Oxford University Press.

Williams, M. and R. Burden. 2010. *Psychology for Language Teachers*. Cambridge: Cambridge University Press.

Winsteps. 2016. "Estimation Considerations". Website. <http://goo.gl/c9uVb2> (12/8/2016).

Woolsey, T., J. Hanaway and M. Gado. 2003. *The Brain Atlas. A Visual Guide to the Human Central Nervous System*. Hoboken (NJ): John Wiley & Sons.

Yerevan Communiqué. 2015. <http://goo.gl/5ZB2wU> (21/07/2016).

Young, J., Y. So and G. Ockey. 2013. *Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments*. Online publication. Educational Testing Service (ETS). <https://goo.gl/TZ3LK4> (29/09/2016).

Appendix: Rubric

	Language Vocabulary (range and control) Grammar (range and control) Errors	**Pronunciation** General pronunciation Articulation of sounds Prosody (stress, rhythm, intonation)	**Interaction** Information exchange Initiate, mantain and end conversation Cooperation	**Discourse** Cohesion Thematic development Fluency
5	• **Vocabulary** range and control go from good to high to express most general topics and are sufficient to provide clear descriptions, viewpoints and more abstract or cultural topics such as music and films. • **Grammar** control is relatively high and displays complex sentence forms. • **Errors** are minor, not recurring and do not lead to misunderstanding when dealing with simple or complex ideas.	• In **general pronunciation**, he/she can most frequently use appropriate intonation, place stress correctly and articulate individual sounds clearly. • The **articulation of sounds** is reasonably clear in the majority of sounds of the target language in extended speech. • In **prosodic features**, he/she can produce smooth, intelligible spoken discourse with only occasional lapses in control of stress, rhythm and/or intonation, which do not affect intelligibility.	• **Information exchange** takes place to communicate detailed information reliably and to synthesise and report ideas from different speakers. • **Initiates, maintains and ends** discourse appropriately with effective turntaking in a range of topics that go beyond familiar topics or personal interests. • **Cooperates** in the discussion by confirming comprehension, inviting others, giving feedback, follow up statements and inferences and summarizing to help focus the talk.	• The **cohesion** devices used are limited in number but help to link utterances into clear, coherent discourse. • The **thematic development** shows clear descriptions or narratives, expanding and supporting his/her main points with relevant supporting detail and examples. • Is **fluent** to produce stretches of language with a fairly even rhythm but can be hesitant as he/she searches for patterns and expressions and there are few noticeably long pauses.
4	• **Vocabulary** range and control are sufficient to express him/herself about topics like family, hobbies and interests, work, travel, current events and abstract cultural topics such as music and films. • **Grammar** range and control are efficient and accurate in a variety of sentence structures that still display mother tongue influence. • **Errors** occur, but are very scarce, minor and it is clear what he/she is trying to express, even when expressing more complex and unfamiliar topics.	• In **general pronunciation**, he/she can most frequently use appropriate intonation, place stress correctly and articulate individual sounds with effort. • The **articulation of sounds** is reasonably clear in the majority of sounds in extended speech; mispronunciations occasionally occur. • In **prosodic features**, he/she can produce intelligible spoken discourse with frequent lapses in control of stress, rhythm and/or intonation.	• **Information exchange** takes place to give details, to summarize and to give opinion through short, sustained interventions. • **Initiates, maintains and ends** a discussion on a familiar topic, using a suitable phrase to get the floor. • **Cooperates** by exploiting a basic repertoire of strategies to help keep a conversation going and can summarize the point reached in a conversation to help focus the talk.	• **Cohesion** is achieved through frequent connectors, which are used to link simple sentences and a choherent discourse with occasional imprecisions. • **Thematic development** is clear in narrative discourse but lacks subsidiary supporting detail or examples. • Is **fluent** to express him/herself with relative ease. Despite some problems with formulation resulting in pauses and 'cul-de-sacs', he/she is able to keep going effectively without help.
3	• **Vocabulary** range and control are sufficient to express him/herself about topics like family, hobbies and interests, work, travel and current events. Lexical limitations may cause repetition and hesitation. • **Grammar** range and control are used efficiently throughout simple, correct sentence structures that still display mother tongue influence. • **Errors** are scarce, minor and do not intrude or impede communication when expressing simple thoughts, although major errors still occur when expressing more complex and unfamiliar topics.	• **General pronunciation** is most frequently intelligible; can approximate intonation and stress at both utterance and word levels. • The **articulation of sounds** is clear in a high proportion of the sounds, in extended stretches of production; is intelligible throughout, despite a few systematic mispronunciations. • **Prosodic features** (e.g. stress, intonation, rhythm) are generally employed to support the message he/she intends to convey, though with some noticeable influence from other languages he/she speaks.	• **Information exchange** takes place frequently but only to pass on straightforward factual information. • **Initiates, maintains and ends** simple, face-to-face conversation on topics that are familiar or of personal interest. • **Cooperates** by inviting others into the discussion and repeating back words to confirm understanding and keep the development of ideas.	• **Cohesion** is achieved through the most frequently ocurring connectors, which are used to link simple sentences, to tell a story or to describe something as a simple list of points. • In **thematic development** he/she can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. • Is **fluent** enough to keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.
2	• **Vocabulary** is sufficient to conduct routine, everyday activities involving familiar situations and topics beyond simple survival needs. • **Grammar** is limited to correct sentence structures that display heavy influence of the mother tongue. • **Errors** are frequent and intrude with simple and complex ideas, although it is usually clear what he/she is trying to say.	• **General pronunciation** is most frequently intelligible; tries to approximate intonation and stress at both utterance and word levels but does not always succeed. • The **articulation of sounds** is intelligible throughout, despite occasional mispronunciation of individual sounds and words he/she is less familiar with. • In **prosodic features**, he/she can convey the main point of his/her message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks.	• **Information exchange** is scarce and limited to pass on straightforward information. • **Initiates, maintains and ends** only short conversations through simple techniques. • **Cooperates** in simple exchanges without undue effort and responds to suggestions and agrees or disagrees with the interlocutor.	• **Cohesion** may go beyond simple connectors but is not able to narrate a short sequence of events. May be able to describe something as a simple list of points. • In **thematic development** he/she has to struggle to relate simple narratives or descriptions as a linear sequence of points, and he/she occasionally succeeds. • Is **fluent** to make him/herself understood in short contributions, even though pauses, false starts and reformulation are very evident.
1	• **Vocabulary** is sufficient for the expression of basic communicative needs and for coping with simple survival needs. • **Grammar** is limited to simple, correct structures, basic sentence patterns and memorised phrases and brief everyday expressions to satisfy simple needs. • **Errors** are systematic (tends to mix up tenses and forget to mark agreement). Nevertheless, it is usually clear what he/she is trying to say.	• **General pronunciation** is clear enough to be understood, but conversational partners will need to ask for repetition from time to time. • The **articulation of sounds** is intelligible throughout, provided the interlocutor makes an effort to understand specific sounds. • In **prosodic features**, the interlocutor must struggle to understand the main point of his/her message due to a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks.	• **Information exchange** only takes place to provide personal information or limited information on familiar and routine operational matters. • Does not **initiate, maintain or end** conversations and is only able to handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord**.** • **Cooperates** only in very simple situations, chiefly to indicate when he/she is following or not.	• **Cohesion** is achieved by linking groups of words with simple connectors like 'and', 'but' and 'because'. • **Thematic development** is reduced to describing something as a simple list of points. • Is **fluent** to construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.