DIPARTIMENTO DI PSICOLOGIA
DEI PROCESSI DI SVILUPPO
E SOCIALIZZAZIONE

FACOLTÀ DI MEDICINA
E PSICOLOGIA

SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato in Psicologia Sociale, dello Sviluppo e della Ricerca Educativa

TESI DI DOTTORATO

English Language Knowledge of First-Year University Students on Performance-Based Tests

Dottoranda
Snežana Mitrović

Tutors
Prof. Pietro Lucisano
Prof. Guido Benvenuto

Ciclo XXX

Anno Accademico 2017 – 2018

II

DIPARTIMENTO DI PSICOLOGIA
DEI PROCESSI DI SVILUPPO
E SOCIALIZZAZIONE

FACOLTÀ DI MEDICINA
E PSICOLOGIA

SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato in Psicologia Sociale, dello Sviluppo e della Ricerca Educativa

Tesi di Dottorato

Dottoranda
Snežana Mitrović

Tutors
Prof. Pietro Lucisano
Prof. Guido Benvenuto

XXX Ciclo

English Language Knowledge of First-Year University Students on Performance-Based Tests

Anno Accademico
2017 – 2018

Composizione grafica a cura dell'Autore

Contents

List of Figures

List of Tables

Introduction

Motivation for the research

It is difficult if not impossible to tell with certainty how many people in the world speak English as a mother tongue, second or foreign language. In 2006, in his English Worldwide, David Crystal, a linguist, writer, and lecturer, reported that around 400 million people spoke English as a first language or mother tongue, 400 million people as their second language and 600 to 700 million as a foreign language, totaling 1,400 to 1,500 speakers worldwide. He believes that English has become so independent from any form of social control that nothing would be able to stop its ever more frequent use as a global lingua franca and that "there are no precedents for languages achieving this level of use" (Crystal, 2006, p. 422). In his opinion, this has been caused by different political and economic changes as well as thanks to the press, advertising, popular music, traveling, etc.

The European Commission policy "united in diversity" promotes language learning and linguistic diversity in Europe[1]. An ambitious goal of the Barcelona objective, agreed in 2002 by the EU's governments, is to enable citizens of the European Union to communicate in two languages other than their mother tongue. The Italian Ministry of Education Decree 509 of 3 November 1999[2] introduced as mandatory the knowledge of another European language in addition to Italian. The recent Gelmini reform made English mandatory for all types of upper-secondary schools in Italy. According to the 2010 Guidelines of the Italian Ministry of Education, "Indicazioni Nazionali," the aims and objectives of the fifth-year upper-secondary school curriculum for English correspond to the CEFR level B2. However, according to Education First[3], the only comparative study on the level of English in different European countries published in 2016, Italy ranks 28th among 72 countries, with a moderate level of English, aligned to the CEFR level B1.

When enrolling at an Italian university, students are required to demonstrate the knowledge of English at a certain CEFR level and pass the so-called university qualifying exam, or "idoneità" in Italian. Since

[1] https://ec.europa.eu/education/policy/multilingualism_en

[2] http://www.miur.it/0006Menu_C/0012Docume/0098Normat/2088Regola.htm

[3] http://www.ef.com/__/~/media/centralefcom/epi/downloads/full-reports/v6/ef-epi-2016-english.pdf

many universities give exemption from this exam to the holders of an internationally recognized certificate in English, many students decide to take an internationally recognized exam in English during the last year of the upper-secondary school or when they start the university. Are the students, independent of whether or not they possess a certificate in English, able to communicate in English? Are they able to write an inquiry email to get the information they need or to ask for directions in English in a foreign-speaking country? Simply put, are they able to perform real-life tasks? To the researcher's knowledge, similar studies have not been done in Italy. Answering these questions is the main motivation for this research.

Purpose of the research and research questions

The aim of the present research is to investigate the level of English language knowledge of the first-year university students through performance-based assessment of real-life tasks. The questions that this research seeks to answer are:

1. How do Italian students, after they have finished high school, perform on written and spoken extended production tasks that reflect everyday real-life activities and situations?

2. Are their speaking and writing skills at the CEFR B2 level of English language knowledge (as per the Ministry of Education Guidelines)?

3. What is their level of acquisition in different areas of language knowledge such as organizational and pragmatic knowledge and their individual components?

Both written and spoken tasks were designed for the purpose of the research, as well as accompanying assessment scales based on Bachman and Palmer's (1996, 2010) framework of language knowledge. The framework consists of organizational and pragmatic knowledge, where organizational knowledge encompasses grammatical knowledge and textual knowledge, while pragmatic knowledge comprises functional and sociolinguistic knowledge. This framework, along with the Common European Framework of Reference descriptors, was used as the basis for the test and scales development.

Delimitations of the research

The concept of communicative competence is discussed in detail in Chapter Three of Part Two. The modern definitions of communicative competence (for example, Canale, 1983; Bachman & Palmer, 1996, 2010) see it as comprising not only language knowledge but also the strategic competence, defined as "higher-order metacognitive strategies that provide a management function in language use" (Bachman & Palmer, 2010, p. 48). This research recognizes this distinction but focuses on the language knowledge component.

Another significant delimitation concerns the sample. Namely, the participants of the present research, all 189 of them, come from the Department of Educational Sciences of the Faculty of Psychology and Medicine, University of Sapienza, Rome. In terms of English language knowledge, according to the research findings, the group is quite heterogeneous. These findings, however, cannot be generalized to all Italian first-year university students but can only be indicative of university-level students.

Organization of the thesis

The research report consists of the Introduction, nine Chapters, Conclusion, and Appendices, one of which is the Glossary.

The Chapters are divided into three parts: Part One: Background to the Research, Part Two: Theoretical Background, Part Three: The Research. The Introduction is followed by Part One, where Chapter One provides an overview of English as a Foreign Language in Italy, including the Ministry of Education Degrees and national and international research projects, as well as a short review of English as a foreign language exams present in Italy. The chapter is followed by Chapter Two, which summarizes the history of the Common European Framework of Reference for Languages, as well as its advantages and disadvantages.

In Part Two: Theoretical Background, Chapter One provides a critique of the theories and models of communicative competence and language, starting from the first ones, until present-day models. Chapter Two describes the history of second language assessment divided into three historical periods. Chapter Three starts with the definition and history of performance-based assessment, then discusses its development and ends with the implications of performance-based assessment. Chapter Four

addresses the issue of validity in second language assessment and presents different approaches to the concept of validity.

Chapter One of Part Three provides a detailed description of the research methodology employed in the present research, including its basis in the theoretical background provided in Part Two of the research report. Chapter Two deals with the study validation and addresses different types of test validity. Finally, in Chapter Three, the findings of the research and their interpretation are presented, followed by a detailed analysis of student performance.

The Conclusion provides a summary of the findings, then discusses the limitations of the research, to end with a summary of advantages to the approach adopted in the research.

Finally, he Appendices provide both analytic and holistic scales used in the research, as well as the writing test, the speaking test, the student questionnaire and speaking test student responses. A Glossary of basic terms in foreign language teaching and assessment is also provided as Appendix I.

Part One

Chapter One
Italy and the English Language

1.1. English as a foreign language in the Italian education system

Starting from 1999, Europe has seen major changes in the higher education system, initiated by the Bologna Declaration, signed and adopted by 29 European countries at the University of Bologna on 19 June 1999, which instigated the Bologna process whose aim was to restructure and harmonize the higher education systems of European countries (Higher Education and Research, Council of Europe)[4]. The main objectives of the Bologna Declaration were to enable comparability of higher education degrees coming from different European countries, establish a credit system, enable mobility of students, teachers, and researchers across European countries and to adopt a two-cycle system of higher education consisting of undergraduate and graduate studies, which later became a three-cycle system consisting of bachelor's degree, master's degree and PhD degree. Each of the cycles is defined in terms of the number of European Credit Transfer and Accumulation System (ECTS), with each university course carrying a certain number of points. By now, 46 European countries have joined the Bologna Process.

The Bologna Declaration immediately triggered changes in Italy as well, starting with the Ministry of Education Decree 509 of 3 November 1999, published in the Official Gazette on 4 January 2000[5]. In addition to adopting the objectives of the Bologna Declaration, it introduced as mandatory the knowledge of another European language in addition to Italian. The level of the knowledge of a foreign language would be determined by individual universities depending on the course of studies.

The same year, the Ministry of Education initiated a project called Progetto Lingue 2000, which found its grounds in the Council of Europe's premise that every citizen of the European Union should be able to "communicate in two languages other than their mother tongue" (Multilingualism – Education and Training, European Commission)[6].

[4] http://www.coe.int/t/dg4/highereducation/EHEA2010/BolognaPedestrians_en.asp

[5] http://www.miur.it/0006Menu_C/0012Docume/0098Normat/2088Regola.htm

[6] https://ec.europa.eu/education/policy/multilingualism_en

Progetto Lingue 2000 started from the 1999/2000 school year and covered foreign language teaching from nursery to upper secondary school and introduced changes to the number of languages studied at school: one foreign language starting from nursery throughout the mandatory education, and another one from the lower secondary school. A significant change here was the introduction of a first foreign language in the last three years of classical lyceums. The number of hours per language and per school was also redefined: 100 hours in nursery, 300 hours in the primary school, 300 hours in the lower secondary school and 200 hours in the upper secondary school. The second foreign language, however, would be taught starting from the lower secondary school, 240 hours a year, while in the upper secondary school 200 hours per year. Furthermore, the project defined the foreign language objectives in terms of the Common European Framework of Reference for Languages levels: for the first foreign language CEFR level B1/B2 at the end of the upper secondary school and CEFR level A2/B1 for the second foreign language. Importance is also given to the exposure to foreign languages, in the form of study holidays, TV programs and movies in the original language, especially for the students aged 16 to 18. Freedom was given to the schools to decide on the curricula and the exact number of hours[7].

Another significant change was the introduction of the possibility for students to take internationally recognized exams, which was to prepare students for their studies, whether in Italy or abroad. With the Ministerial Decree 49 of 24 February 2000, ETCS points were given for the first time for certified foreign language knowledge, without any reference to a specific awarding body. The decree, however, did state that certificates released in Italy, by an awarding body recognized in the country of origin, did not need to be legalized or authenticated[8].

At the time, there was only one English language awarding body present, Cambridge. As a result, private language schools that offered language courses started opening and new awarding bodies started arriving. With English being the most accessible and popular language, the majority of courses were English language courses and the first new awarding bodies to arrive were British.

In January 2002, the Italian Ministry of Education signed an agreement with several foreign examination boards, including five English language

[7] http://www.edscuola.it/archivio/norme/programmi/progettolingue.pdf

[8] https://archivio.pubblica.istruzione.it/argomenti/esamedistato/secondo_ciclo/quadro/dm49_00.htm

ones. This agreement gave schools in Italy access to internationally recognized certificates as well as funding for extracurricular language activities through the European Union projects and European structural and investment funds. For the first time in 2007, the Ministry of Education at the request of the British Embassy in Rome, with Protocol no. 8075, referred schools to the British Council for the list of recognized English language awarding bodies. Consequently, although they were fully autonomous, universities started referring to the British Council for internationally recognized certificates in English.

On 7 March 2012, the Ministry of Education published another decree, no. 3889, where it listed the awarding bodies that they recognized. This decree has been updated regularly and new awarding bodies added to the list. The latest decree with the list of recognized awarding bodies was published on 28 February 2017.

After Progetto Lingue 2000, there have been a few reforms of the education system in Italy, some of which included changes to the number of languages taught in lower and upper secondary schools and to the number of hours of the teaching of these languages.

With the Moratti reform, and Law no. 53 of 28 March 2003[9], the teaching of another language of the European Union was introduced, and English became the first foreign language taught in Italian schools. Appendix D of the Decree of 17 October 2005 provides the number of hours of English to be taught in different types of upper secondary schools: a total of 528 hours for the linguistic lyceum, that is 4 hours a week for a year and 3 hours a week for the rest of the lyceum. For the economic lyceum, 3 hours a week with a total of 495 hours in the five years of school, and for all other ones 2 hours a week, for a total of 330 hours in the five years. The decree refers to the CEFR for the objectives of the curriculum: CEFR level C1 for the linguistic lyceum, lower C1 level for the economic lyceum and CEFR level B2 for all other lyceums. This was amended by the Ministry of Education Guidelines of 4 February 2010[10], which in Appendix A state that the objective of the English language curriculum of all lyceums is CEFR level B2. The number of hours becomes the same for all lyceums and other types of upper secondary schools except for the linguistic lyceum: 3 hours a week, with a total of 99 hours a year or 495 hours over the five years. For the linguistic

[9]https://archivio.pubblica.istruzione.it/mpi/progettoscuola/allegati/legge53_03.pdf

[10] http://www.edscuola.it/archivio/norme/programmi/licei_2010.pdf

one, on the other hand, the number of hours per week is 4 for the first two years, with a total of 132 hours a year, while in the last three years, it is the same as for other lyceums: 99 hours a year, that is 3 hours a week. The 132 hours in the first and the second year include 33 hours of lessons with a mother-tongue teacher. These changes were part of the Gelmini reform, which also made English mandatory for all types of upper secondary schools. At the same time, English became the only language taught in most upper secondary schools. In the linguistic lyceum, for example, at least another foreign language is taught. In addition, all students in the last year of upper secondary schools learn one non-language subject in a foreign language through CLIL.

According to Cifre chiave sull'insegnamento delle lingue a scuola in Europa 2012, a report published by Eurydice[11], the Education Information Network in Europe, 97.8% of students in upper secondary schools in Italy study English as the first foreign language.

1.2. Proficiency in English of Italians

Although student academic performance in different subject areas is evaluated by different research projects such as PISA (Programme for International Student Assessment), there are not many research projects or comparable studies of English language proficiency. The only ones that the researcher is aware of are the Education First English Proficiency Index and the self-reported data about the knowledge of English published by Eurobarometer on behalf of the European Commission.

1.2.1. Education First English Proficiency Index

Education First English Proficiency Index (EF EPI) is published by EF Learning Labs, part of EF Education First, an international education company founded in 1965 and specializing in language courses and training, academic study programs, educational travel and cultural exchange. They are present in 116 countries with 539 offices and schools.

11 http://eacea.ec.europa.eu/education/eurydice/documents/key_data_series/143IT_HI.pdf

According to the sixth edition of EF EPI report[12], Italy ranks 28th among 72 countries, with moderate proficiency in English. What does it tell us about the proficiency in English of Italians?

The ranking is based on 950,000 test takers who completed three different Education First English test in 2015. All the three tests are multiple-choice tests, two out of which are open online tests, while the third one is an online placement test that Education First uses for student placement in their English courses. The open tests comprise 30 questions and are adaptive, meaning that the questions that the students need to respond are based on the correctness of their previous response. All tests assess reading and listening comprehension only.

For a country to be included in the index, there needs to be a minimum of 400 test takers from the country. 46.3% of the sixth edition index test takers are female, while the median age of female test takers is 28 years. The median age of the male respondents is two years higher.

The scores of the test takers are translated into percentages and aligned to the CEFR levels, while each country is assigned to a proficiency band to enable comparability across countries. The five proficiency bands used to report the results are very low proficiency, low proficiency, moderate proficiency, high proficiency and very high proficiency. The very high proficiency band has been aligned to the CEFR level B2 high, while high, moderate and low proficiency bands are aligned to the CEFR B1 level. Finally, the lowest, band, very low proficiency, corresponds to the CEFR A2 level.

Education First, however, recognize the limitations of the test, the first one being that the test is completed by those who have access to the Internet and those who decide to complete it, which most often means that they are studying English and for that reason what to have their knowledge or progress assessed.

In addition, starting from 2015, EF publishes the EF English Proficiency Index for Schools[13] of secondary and tertiary education, as a study of the acquisition of English skills by secondary and tertiary students. The first edition is based on 130,000 students from 16 countries, one of which is Italy, and it provides information on the rate of improvement in English over the period of a year in three different age

[12] http://www.ef.com/__/~/media/centralefcom/epi/downloads/full-reports/v6/ef-epi-2016-english.pdf

[13] http://www.ef.edu/epi/reports/epi-s/

groups: 13 – 15, 16 – 18 and 19 – 21, where data are based on a comparison of different groups of students of different ages.

The test used in the research is the EF Standard Test of English, which is available online and free of charge. Among the 16 countries, Italy is the only one where students of different ages progress at the same pace, with the average score of CEFR A2 at the age of 15 for both listening and reading. The average score at the age of 20, however, is CEFR B1 in reading, and a lower B2 level in listening. The report does not state the number of Italians that participated in the research. The general approach of EF is that any school can participate and they state that some schools tested all their students while some participated in the research with only one of their classes. For that reason, the report findings cannot be considered to represent a whole nation's proficiency in English.

One drawback of the index is the conclusions made based on the research. Firstly, it is not explained how the EF levels have been aligned to the CEFR. More importantly, the predictions of performance based on the EF test scores, divided into bands, are not only for reading and listening, but also for the productive language skills, that is speaking and writing. Similarly, the CEFR descriptors used in the reports to enable score reading and comparability of scores are the general ones and include Can Do statements related to speaking and writing skills as well.

Despite its limitations, the EF English proficiency index remains one of the few international research projects concerning English language proficiency. The moderate proficiency in English of the Italians who have completed the test and ranked the country 28th position, according to EF, corresponds to the CEFR level B1.

1.2.2. Eurobarometer Report

The Standard Barometer[14] is a series of public opinion surveys conducted by Kantar TNS, a market research and information group, on behalf of the European Commission. 1,036 Italians participated in their survey on Europeans and their languages, published in 2012[15]. The data were gathered from 25 February to 11 March 2012. According to the report, 70% of interviewed Italians find English to be the most important

[14] http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/General/index

[15] http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/yearFrom/1974/yearTo/2012/surveyKy/1049

language, other than their mother tongue, for their personal development. 88% believe that everyone in the European Union should be able to speak at least one language in addition to their mother tongue but only 36% prefer to watch foreign movies and programs with subtitles, rather than dubbed, and only 34% of the Italian interviewees think they can hold a conversation in English.

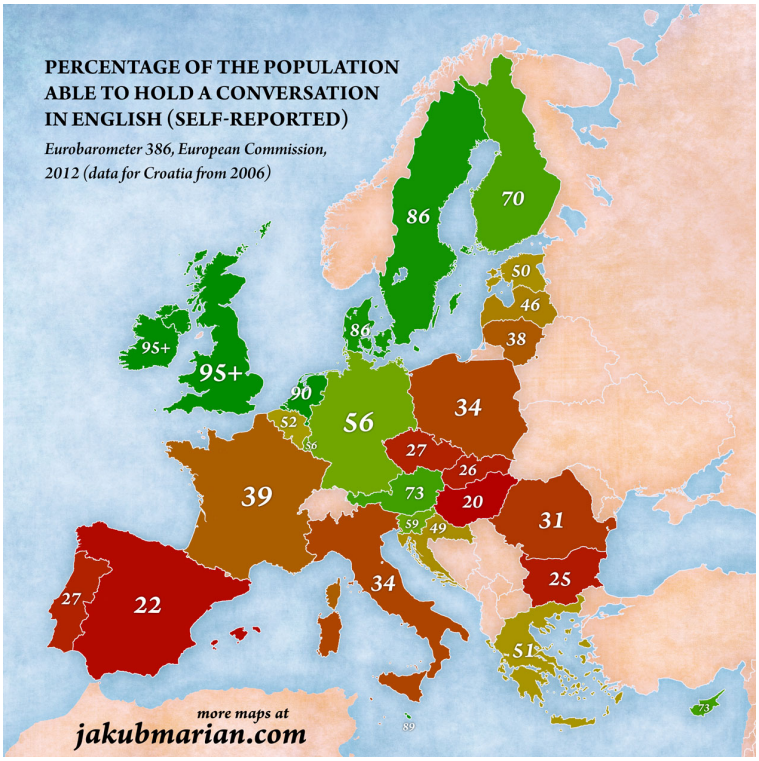

Figure 1. Percentage of the population able to hold a conversation in English (self-reported). Based on Eurobarometer 365, European Commission.[16]

1.3. English as a Second Language (ESOL) exams in Italy

There are a number of different exams of English as a Second language, also known as ESOL exams. These certificates are offered and awarded by different examinations boards mostly based in the UK. They can be different in format and the skills that they cover: some of them test only

[16] Taken from: https://jakubmarian.com/map-of-the-percentage-of-people-speaking-english-in-the-eu-by-country/

one of the four skills (reading, listening, writing and speaking), some all four in separate tests and some use a single test to test more than one skill. Most often, they are available at different levels and their comparability is possible thanks to their alignment to the Common European Framework of Reference for Languages.

For an English language exam to be internationally recognized, it needs to be awarded by an examinations body recognized by Ofqual (The Office of Qualifications and Examinations Regulation), a non-ministerial government department which regulates qualifications, examinations and assessments in England.[17]

According to the Italian Ministry of Education Decree of 28 February 2017 the following English language awarding bodies or English language exams are recognized in Italy:

- Cambridge ESOL,
- City and Guilds (Pitman),
- Edexel / Pearson,
- Educational Testing Service (ETS),
- English Speaking Board (ESB),
- International English Language Testing System (IELTS),
- Pearson - LCCI,
- Pearson - EDI,
- Trinity College London,
- Department of English, Faculty of Arts - University of Malta,
- National Qualifications Authority of Ireland - Accreditation and Coordination of English Language Services,
- Ascentis,
- AIM Awards;
- Learning Resource Network (LRN),
- British Institutes,
- Gatehouse Awards.

There are a number of reasons why Italian students of English take internationally recognized or recognized by the Ministry of Education exams of English: to obtain a study visa or to settle in an English-speaking country, to enroll at a foreign University or to get exemption from their English language exam at the University. The universities have the freedom to decided which certificate or certificates and at which of the CEFR levels they will accept to give a certain number of ETCS points and

[17] https://www.gov.uk/government/organisations/ofqual

consequently exemption from a part of or the entire English language exam.

The most popular English language exams in Italy are Cambridge ESOL exams, IELTS, TOEFL (ESB), JETSET LCCI Pearson and Trinity College exams. They owe their popularity to the recognitions by different universities across Italy. IELTS and TOEFL are assessment tests, that is they evaluate the level of the test taker's knowledge across the four skills (reading, listening, writing and speaking), while the rest are all level-based and all include all four skills, except for the Trinity's Graded Examinations in Spoken English.

None of the awarding bodies has made public the number of test takers in Italy per year. Some of them do publish validation research data (Cambridge, Pearson, IELTS) and the percentage of students according to their score (IELTS) and grade statistics (Cambridge). Without the number of students on which the statistics are based however the published data have little meaning.

Chapter Two
CEFR: Common European Framework of Reference for Languages

2.1. Introduction

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment was created by the Language Policy Division of the Council of Europe between 1989 and 1996 after twenty years of research in the field. The principal goal was to provide an easily understandable and comprehensive framework for learning, teaching and assessing foreign languages as well as to provide a basis for all those involved in teaching a foreign language, the design of foreign language syllabi and exam construction.

2.2. Origins and history

The CEFR was not the first attempt at defining levels of foreign language proficiency. The first publication in the field dates to 1970s when the Council of Europe released the first document that describes a level of foreign language proficiency: *Threshold Level,* published in 1975 (van Ek, 1975) and republished in 1990 (van Ek & Trim, 1990a). The same year, *Waystage* (van Ek & Trim, 1990b) was published and finally in 2001 (Van Ek & Trim, 2001) *Vantage Level*. In the process of the CEFR development, Threshold, Waystage, and Vantage were attached to CEFR levels A2, B1 and B2 respectively. These three publications "are purely descriptive, and the distance between Waystage and Threshold is not based upon any empirical evidence, but the intuition of the authors" (Fulcher, 2004b, p. 256).

It was Wilkins who, as early as 1978, first proposed seven levels of language proficiency. At the 1991 Rüschlikon Symposium, Carroll presented a nine-level framework proposal, seven out of which were the Wilkins's ones. In 1993, as a follow-up to the symposium, a project by the Swiss National Science Research Council was undertaken and lasted until 1996, when it resulted in the first version of the Common European Framework. The project was run in four phases (Council of Europe, 2001, pp. 217 - 218): intuitive, qualitative, quantitative and interpretation phase.

In the intuitive phase, the existing scales of language proficiency were analyzed and deconstructed into descriptive categories in relation to the CEFR's action-oriented approach. This was followed by the qualitative phase, in which recordings of teachers' discussions and student performance were analyzed to verify that the metalanguage used by practitioners was adequately presented, which was followed by 32 workshops with teachers whose tasks were to sort descriptors into categories, make qualitative decisions about the clarity, accuracy, and relevance of the descriptions as well as sort descriptors into three pre-defined bands of proficiency (low, middle or high). In the qualitative phase, over a period of two years, 12 questionnaires, each with 50 descriptors, that in the previous phase classified most consistently, were administered. The teachers' task was to assess representative learners for each descriptor on a rating scale from 0 to 4. The teachers' interpretations of the descriptors were then analyzed using the Rasch rating scale model. This phase of the project involved almost 300 teachers from lower secondary, upper secondary, vocational schools and schools for adults and around 2,800 students originating from 500 classes. The first seven questionnaires, administered in the first year of the phase, focused on Interaction and Production and were limited to English as a Foreign language. In the second year of the qualitative phase, the descriptors for the spoken interaction were reused, this time surveying French and German proficiency as well. At the same time, a Reception survey was added, and self-assessment and some examination information were added to the teacher assessment. Finally, in the interpretation phase, cut-points were established and translated into language proficiency levels (Common Reference Levels), the global scale, a self-assessment grid and a performance grid.

2.3. Contents

In the process, Threshold, Waystage and Vantage were incorporated in the Common European Framework levels, which in its final version describes foreign language proficiency using six levels (A1, A2, B1, B2, C1 and C2), reflecting its original idea to make it possible to compare language courses, tests and examinations across languages and countries. Apart from a general description of each of the levels, it also provides an analysis of communicative contexts, themes, tasks and purposes as well as descriptions of competences (knowledge, skills, and attitudes) at different levels (Council of Europe, 2001, p. 23).

The document can be divided into two parts: the descriptive scheme and illustrative scales. In its Chapter 2, (p. 9) Approach adopted, there is a detailed scheme overview (further detailed in the following chapters of the document) where language use and learning is described as follows:

Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of competences, both general and in particular communicative language competences. They draw on the competences at their disposal in various contexts under various conditions and under different constraints to engage in language activities involving language processes to produce and/or receive texts in relation to themes in specific domains, activating those strategies which seem most appropriate for carrying out the tasks to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences.

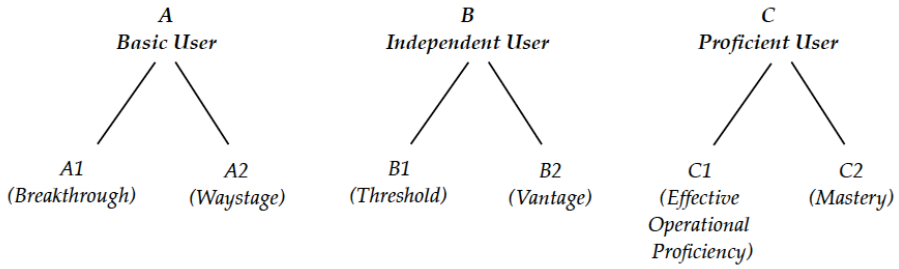

Figure 2. The common reference levels. Reprinted from *Common European Framework of Reference for Languages: Learning, teaching, assessment* (p. 23) by Council of Europe, 2001

It distinguishes between general competences, the ones which are not language specific (knowledge, skills and know-how, existential competence and ability to learn) and communicative language competences, comprising linguistic, sociolinguistic and pragmatic competences. It further defines four categories of language activities: reception, production, interaction and mediation, in four different domains: public domain, personal domain, occupational domain and educational domain; as well as the strategies activated when communicating and learning conditioned by the context, text type and conditions or constraints in different situations.

The other and, according to Alderson (2007, p. 661), more useful part of the CEFR are the illustrative scales, which "have provided an apparently concrete operationalization of the Descriptive Scheme." The illustrative scales, divided into Communicative Activities (reception, interaction, and production), Communication Strategies and Communicative Language Competence (linguistic, sociolinguistic and pragmatic). The scales all comprise "Can Do" statements, which provide descriptions of the aims and objectives at each of the CEFR levels. In addition, the Global Scale provides a general proficiency description at the six levels. The CEFR comprises a total of 57 illustrative scales, 34 of them for communicative language activities (listening comprehension, reading comprehension, spoken interaction, written interaction, spoken production, written production and working with text), seven on communication strategies (divided into reception, interaction and production) and 13 for different aspects of communicative language competence (broadly divided into linguistics, sociolinguistics and pragmatic). It also provides three summary tables (global scale, self-assessment grid and qualitative aspects of spoken language use).

As the chapter title says, the approach adopted is an action-oriented one, according to North (2007, p. 656), "the heuristic behind the CEFR's Descriptive Scheme": the language learner or user is seen as a "social agent", with different tasks to complete in different situations and environments to achieve a certain goal.

2.4. Criticism and Defense

Ever since it was designed, the CEFR has been criticized by multiple authors for its theoretical basis (or the lack of it) and origin as well as for practical issues such as the vagueness of the terminology used and consequently validity issues.

One of the loudest opponents of the CEFR, Fulcher, looks at the origins of the CEFR to claim that "it relies upon scaling descriptors (and) has no basis in second language acquisition theory" (2012, p. 384). To make his point, he also cites North (2000, p. 573 in Fulcher, 2012, p. 384): "what is being scaled is not necessarily learner proficiency, but teacher/raters' perception of that proficiency - their common framework." Fulcher (2004a, 2004b) also looks back on the phases of development of the CEFR and describes the process in detail to make his point. Also, according to Fulcher, because of the widespread use of the CEFR, teachers have actually started believing that the CEFR scales "represent an acquisitional

hierarchy, "that is the order in which students learn a language" (Fulcher, 2004b, p. 260). Finally, according to him, the two levels that were published before the CEFR and later on included in its final version, Waystage and Threshold "are purely descriptive and the distance between Waystage and Threshold is not based upon any empirical evidence, but the intuition of the authors" (Fulcher, 2004b, p. 256).

Another reason why the CEFR has been criticized is its terminology. As Alderson (2007, p. 661) points out, the language used in the illustrative scales is "not easy to understand, often vague, undefined and imprecise." and that "it became apparent that language terms lacked definitions, there were overlaps, ambiguities, and inconsistencies in the use of terminology." Quite often different words or expressions with similar meaning are used without indicating whether or not they are used as synonyms. This particularly refers to the reference-level or illustrative descriptors or scales, where quite often, quite similar descriptions appear at different levels. Morrow (2004, p. 7) describes readers' reaction to the CEFR as a "completely baffling plethora of terminology."

A limitation that some other authors have pointed to is that the CEFR is language-independent (Alderson, 2007; Little, 2007) and does not make any reference to specific languages. According to Little, it does describe certain language functions but not how the functions can be realized in different languages (p. 645).

The question that emerges from the limitations that different authors have addressed, as well from its abstract nature, is whether the CEFR can actually be called a "framework," or it is, as Fulcher says (2004b, p. 258), simply a model. When addressing the CEFR issues, several authors have referred to different models of language competence: *The Manual for Language Test Development and Examining for use with the CEFR* lists the ones of Bachman (1990), Canale and Swain (1980), and Weir (2005a) and maintains that the action-oriented approach of the CEFR includes the major elements of a model of language competence: general and communicative language competences, language activities, language processes, etc. Similarly, Alderson (2007, p. 660) notes that the influence of Wilkins (1976), Canale and Swain (1980) as well as Bachman (1990) is evident in the CEFR. Alderson, as well as Morrow (2004) also mention the influence of communicative language teaching on the development of the CEFR. Originally, the Council mostly focused on teaching language to adults, which led to the development of notional-functional syllabi and the communicative teaching approach, which was a way to create conditions for the development of the CEFR, where the focus switched

from teaching to creating specific learning objectives for different levels. However, in his article "Are Europe's tests based on an unsafe framework?", Fulcher claims that the CEFR has no underlying theory and no content specifications" (para. 10). In his response to Fulcher and to CEFR's defense, North (2004a, para. 4) stresses that "the CEFR draws on theories of communicative competence and language use in order to describe what a language user has to know and do in order to communicate effectively and what learners are typically expected to do at different levels of proficiency. It doesn't try to define what should be taught (content specifications), let alone state how it should be taught (methodology)". Similarly, Morrow (2004, p. 7) refers to the full title of the CEFR and stresses the importance of the phrase "of reference," reminding that its main aim was to act as a frame of reference, that it is a descriptive framework and not a set of suggestions, recommendations or guidelines, simply put, that it is for "description, not prescription."

And the intended use of the CEFR was precisely this one - to help specify learning objectives at different levels, to help teachers and learners teach and learn, to help test designers and decision makers.

Alderson (2007) and Fulcher (2004b) both focus on the unintended and unfortunate uses of the CEFR, both arguing that the CEFR is being used in a wrong way and for unintended purposes. While Fulcher explicitly warns of the danger of linking tests to the CEFR levels and using its levels to compare scores across different tests, Alderson provides more context and discusses the use of the CEFR by non-linguists or non-specialists for defining standards of language proficiency or its use in inappropriate contexts. Similarly, Fulcher (2004b, p. 261) warns of the dangers of institutionalization and the use of the CEFR by awarding bodies and public institutions claiming that "it is not possible to use a description at the model level to meaningfully link tests that have been designed for different purposes, and hence a variety of different construct definitions."

Weir (2005b) addresses this particular issue: the problem of developing comparable examinations and tests based on the CEFR. In this respect, he discusses four problematic areas: context validity, theory based validity, scoring validity and transparency problems. Context validity "is concerned with the social dimensions of the task, including the setting of the task, and in particular the linguistic and social demands" (Weir, 2005b, p. 284). In other words, the setting (purpose, response, format, time constraints), as well as the demands (linguistic channel, discourse mode, length, topic, lexical, structural and functional), need to be defined in the test specifications. Weir then moves on to discuss the

theory-based validity, that is cognitive and metacognitive processing that participants perform when completing the task; and scoring validity: the need for clearly specified criteria and evidence on test raters and rating reliability. However, this is where the CEFR, according to Weir, is lacking.

Nevertheless, many European institutions, including Ministries of Education (one of them being the Italian one), use the CEFR levels to state the entry requirements for different study courses and positions, the danger of which Fulcher has warned. In that respect, North (2004b, p. 78) suggests "studying relevant CEF scales, stating what is and what is not assessed and what level of proficiency is expected as a basis for relating examinations to the CEF."

Considering that North was one of the CEFR creators, he has published a number of articles to the defense of the CEFR, in most of them stressing its intended use and the advantages of having a context-free framework. He reminds that "the aim of the CEF is to empower and to facilitate, not to prescribe or control" (2004a, para. 3), and that the descriptors are not based on a second language acquisition theory because at the time the CEFR was designed, there was not enough research in second language acquisition to base the CEFR descriptors or scales on. He also looks back on the origins of the CEFR and stresses that there are only few illustrative descriptors have been validated empirically in other validation studies as well, not only in the process of the CEFR design.

Finally, Little (2007, p. 649) goes back to the original idea behind the CEFR to state that the CEFR can serve only as a starting point, and in that way confirms Morrow's point of view.

In response to the criticism, as well as to help linking examinations to the CEFR, in 2003, the Council of Europe published *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, and Assessment,* which focuses on the importance of theoretical and empirical evidence to support the validation of the claim. It further stresses the need for test content specifications, which would then be mapped to the content of the CEFR (Fulcher, 2004b, p. 262). The document guides test providers/designers to the different stages of test design, including the linking process, test specifications, standardization training and benchmarking, standard setting procedures as well as validation.

Despite the criticism, CEFR still remains the starting point in the area of language learning, teaching and assessment and is used by the majority of European awarding bodies that refer to the CEFR levels in order to

enable a comparison of the existing language proficiency tests and "help score users interpret the meaning of test scores" (Papageorgiou, 2010; Tannenbaum & Wylie, 2008; Taylor & Jones, 2006, as cited in Papageorgiou, S., Xiaoming, X., Morgan, R., & S. Youngsoon, 2015). The "common framework of teacher/raters" perception of proficiency that North (2000, p. 573) mentions has become common in many different ways: language teaching course books are aligned to it, exam providers align their tests to it and the levels have become "a common currency in language education" (Alderson, 2007, p. 660) Finally, according to Kane (2012, p. 8) meaning can be added to the scores by referencing them to achievement levels such as CEFR.

2.5. Conclusion

The CEFR levels have become commonly accepted and in a way defined, at least in terms of syntax and vocabulary. Publishers offering foreign language course books align their books the CEFR; in the same way, test providers attach their exams to one of the CEFR levels. What has spontaneously happened over the years, since the CEFR was first released is that publishers and awarding bodies have designed their own syllabi and specification including grammar, vocabulary and communicative functions on which they base their course books or tests. Those working in the area of language teaching or assessment, when talking about one of the CEFR levels, have a clear idea, or at least think they have a clear idea, what a person whose English is supposed to be at that level is expected to know. This "idea" is based on the syllabi and specifications of different publishers and awarding bodies.

Part Two

Chapter One
Theoretical Models of Communicative Language Competence and Second Language Performance

1.1. Introduction

Many linguists and researchers have attempted to define communicative competence as opposed to performance and provided their own theoretical models[18] of second language performance. These models have influenced the field of language testing since

> Language tests involve measuring a subject's knowledge of, and proficiency in, the use of a language. A theory of communicative competence is a theory of the nature of such knowledge and proficiency. One cannot develop sound language tests without a method of defining what it means to know a language, for until you have decided what you are measuring, you cannot claim to have measured it. (Spolsky, 1989, p. 140)

It is necessary to define what it means to know a language to be able to design a test that would adequately assess it. Considering that, originally, the main idea behind performance-based testing was to assess "actual performances of relevant tasks" (McNamara, 1996, p. 6), it has never been considered necessary to have and employ an explicit model of performance in performance-based testing. However, to be able to make inferences on candidates' abilities based on their performance on this type of tests, employment of such a model is necessary (McNamara, 1996; Messick, 1994, 1995, 1995; Bachman, 1990).

According to Bachman (1990, p. 82), the first models for describing language proficiency distinguished skills (reading, listening, writing and speaking) from components of language knowledge (grammar, vocabulary, phonology/graphology) but their limitation was that they did not indicate how the skills and knowledge are related. Also, they failed to recognize the context of discourse and situation. In time, different linguists such as Hymes (1972) and Halliday (1978) started recognizing the need for introducing into the model of language proficiency factors other than the language itself (Bachman, 1990, p. 82).

[18] Model, theoretical model and theory are used interchangeably in this Chapter.

The following authors have given their contribution to defining the knowledge of a language, constituents of the knowledge, underlying factors and their relation to actual performance in order to propose models of second language communicative ability. The linguists who proposed these models variously use terms such as "model of language proficiency," "communicative competence" or "communicative language ability" (CLA) (Fulcher & Davidson, 2007, p. 36). Each of these models has three dimensions or components (McNamara, 1996, p. 48):

1) factors of knowledge of a language
2) factors that underlie the knowledge and that enable an individual to perform communicative tasks involving language
3) how instances of language use are seen in relation to the two preceding dimensions.

1.2. Hymes on Competence and Performance

One of the earliest works on communicative competence, Hymes's paper "On Communicative Competence" (1972) was provoked by Noam Chomsky's distinction between competence and performance, published in his *Aspects of the Theory of Syntax* (1965) as well as his desire to contribute to the study of the "language problems of disadvantaged children" (Hymes, 1972, p. 269). Namely, Chomsky's distinction between competence and performance is based on the view of an ideal speaker-listener, where he sees language competence as the speaker-hearer's perfect language knowledge, which is subconscious, while performance is seen as an actual application of this knowledge, with false starts, deviations from rules, etc. (Chomsky, 1965, p. 3). Hymes finds this distinction limiting and considers it necessary to challenge this understanding in order to provide insight into a number of linguistic problems. Furthermore, Chomsky's view of competence and performance does not consider the relevance of social context and sociolinguistic norms of appropriateness, that is knowing the culturally specific rules (Hymes, 1972, p. 272) but implies that only in ideal conditions would performance reflect competence.

The main difficulty of Chomsky's theory, according to Hymes, is that he associates competence with grammaticality and performance with acceptability, whereas in his opinion, grammatical competence is just one of several types that constitute communicative competence.

In his attempt to redefine the notions of competence and performance, Hymes (1972, p. 280) proposes two contrasts:

1) (underlying) competence v. (actual) performance;
2) (underlying) grammatical competence v. (underlying) models/rules of performance.

According to McNamara (1996, p. 55), Hymes here distinguishes between performance models, that is ability as potential and actual use, which is the realization of the potential. Unlike Chomsky, who leaves no space for sociolinguistic norms or social context, Hymes also distinguishes between communicative competence, which he then divides into knowledge and ability for use, and performance. By introducing the notion of ability for use, Hymes creates space for elements of communicative competence other than grammatical competence.

To provide an explanation as to how these types of communicative competence may interact, he proposes four judgments or features of instances of language use:

1) Whether (and to what degree) something is formally *possible*;
2) Whether (and to what degree) something is *feasible* in virtue of the means of implementation available;
3) Whether (and to what degree) something is *appropriate* (adequate, happy, successful) in relation to a context in which it is used and evaluated;
4) Whether (and to what degree) something is in fact done, actually *performed*, and what its doing entails. (Hymes, 1972, p. 281, emphasis in original)

According to Hymes, "something possible within a formal system is grammatical, cultural, or, on occasion, communicative" (Hymes, 1967 as cited in Hymes, 1972, p. 285). Here Hymes expands on Chomsky's idea of grammatically correct sentences to be the only ones possible or acceptable and includes cultural as well as communicative in addition to grammatical.

In the second judgment, Hymes explains that what the linguistic theory considers performance and acceptability lacks a general term in relation to cultural behavior and he proposes the term "feasible." This is because some sentences, although can be said to be grammatically correct or possible are not likely to be feasible.

In the third judgment, Hymes considers to what extent something is appropriate in a particular context, maintaining that the grammatical and

cultural appropriateness are not separable one from another. A sentence can be grammatically possible and feasible but not appropriate to the context. On the other hand, linguistic theory places appropriateness under performance and at the same time acceptability.

Finally, "something may be possible, feasible and appropriate and not occur" (Hymes, 1972, p. 286), which can be explained by our idea about what is or is not commonly accepted.

Hymes believes that an understanding of these four judgments is necessary in order to be able to interpret behavior in relation to the culture and context, maintaining that "there are rules of use without which the rules of grammar would be useless" (Hymes, 1972, p. 278). In this way, he uses Chomsky's term competence but redefines it to include not only grammar but also sociocultural features (Munby, 1978).

Widdowson (2001) rightly recognizes that Hymes does not indicate how the features of language use are related. He equates Hymes's notion of "possible" to Chomsky's notion of linguistic competence, which concerns grammatical knowledge. Consequently, the ability to distinguish grammatical from ungrammatical sentences without context is then part of communicative competence. The problem here, however, is that these components of competence do not relate one to another and it is not known how they interact for "the whole is a function and not a sum of its parts" (Widdowson, 2001, p. 13). Similarly, Hymes's notion of possible encompasses only what is grammatically possible but not what is lexically possible. Finally, some expressions, such as elliptical phrases, that would be impossible in isolation would at the same time be possible or appropriate in a certain context. This is why Widdowson maintains that Hymes's theory only partly deals with communicative competence, as "communication involves not identifying separate features, but exploiting relationships between them" (Widdowson, 2001, p. 13).

Nevertheless, this was a pioneer work in the area of communicative competence as it was Hymes who coined the term and introduced the importance of context, that is the ability to use the grammatical competence in different communicative contexts or settings.

1.3. Halliday's View of Language Knowledge

Hymes was not the only one who reacted to Chomsky's idea of linguistic competence of the ideal native speaker.

Halliday (1978) adopts a similar standpoint and agrees with Hymes that Chomsky's linguistics is limiting. He believes that, while it is

reductionist due to its idealization of natural language, it still demonstrates that natural language can be studied as a formal system. In his definition of competence, Chomsky idealizes the natural language, and in that way attributes physiological side-effects, mental blocks, statistical properties of the system, subtle nuances of meaning, etc. to performance. Having said that, Halliday (1978, p. 38) maintains that that leaves two options: to accept the distinction and decide to study performance, which Hymes does in his works (and studies sociolinguistic competence as part of communicative competence), or to reject the distinction due to the high level of idealization and accept the mess in order to study a potential, that is what one could do, which in his opinion is objective, unlike competence, which he considers subjective.

Therefore, Halliday (1978, p. 39) describes language as a system consisting of semantics, grammar (including vocabulary) and phonology, where he defines grammar as the system of what the speaker can say, semantics as the system of what the speaker can mean, the "meaning potential", by means of which the grammar system is realized. Finally, considering factors other than the language itself, the semantic system is the realization of what he calls "behavior potential," what the speaker can do. In other words, an individual has at their disposal a number of potential behaviors ("can do") which are not necessarily linguistic. If the individual chooses to express themselves linguistically, they have a number of semantic options, that is a number of meanings to choose from ("can mean"). Having decided on the intended meaning, the individual chooses from linguistic options at their disposal ("can say").

We can see that while Chomsky, when talking about competence, refers to knowledge of grammar and of other aspects of language, and when talking about performance, refers to the actual use of a language, Halliday, one of the few researchers who does not recognize the distinction, believes that this distinction is unnecessary and can be misleading (Canale and Swain, 1980, p. 3). According to Canale and Swain, one of the most critical elements of his research is his notion of a "meaning potential," where a user's behavior is determined by a social system, depending on which he chooses from the semantic and consequently grammatical options. Similarly to Hymes' notion of "communicative competence" and his idea of appropriateness in relation to the context, Halliday's "meaning potential" is constrained by the society (Halliday, 1978).

1.4. Campbell and Wales

Campbell and Wales (1970 as cited in Canale and Swain, 1980) adopt a standpoint similar to the one of Hymes, asserting that producing utterances which are appropriate to the context is more important than producing grammatically correct ones. Furthermore, unlike Halliday, they recognize the distinction between communicative competence and performance (Canale and Swain, 1980, p. 4).

1.5. Munby

Finally, Munby, in his model of communicative competence, talks about three components: a sociocultural orientation, a socio-semantic view of linguistic knowledge and rules of discourse. Like Halliday, he believes that the choice of semantic options depends on the context, that is the social structure. (Munby, 1980 as cited in Canale and Swain, 1980, p. 21).

Munby (1978) discusses the previous theories of communicative competence and starts by accepting Chomsky's distinction between competence and performance. However, Munby believes that none of the two provides a place for competency as they do not include sociocultural significance. He then goes on to make a distinction between actual performance and underlying rules of performance, which Chomsky fails to do and which Hymes considers part of underlying competence. He agrees with Hymes that grammatical correctness or accuracy is not the most significant feature, again confirming his view on the significance of appropriateness to the context as well as that his view of communicative competence includes grammatical competence (Canale and Swain, 1980).

The theoretical model of communicative competence in a second language that Munby proposes comprises three major components: the sociocultural orientation, the socio-semantic basis of linguistic knowledge and the discourse level of operation, each of which is then divided into other components.

The first constituent of his sociocultural orientation, that according to Canale and Swain (1980) is based on Hymes's work, is "competence and the community" (Munby, 1978, p. 23), where he explains that there is no perfect competence nor a homogeneous speech community, but communities that consist of members with different levels of competence. For this reason, communicative competence needs to be seen in relation to the communicative needs of the community. The second constituent of

Munby's sociocultural orientation is contextual appropriacy, where he reiterates that the knowledge of a foreign language is not sufficient for effective communication, but there needs to be the knowledge of the rules of use and language appropriate to the social context. He then introduces the concept of "language variety," which in his opinion is "characterized by its selection and use of linguistic forms for its constitutive communicative acts or functions" (Munby, 1978, p. 24). The choice of variety will, of course, depend on the context. As the third constituent, Munby proposes communicative needs of the learner, on which the speech functions and communicative acts that will be taught depend.

The socio-semantic basis of linguistic knowledge as its first constituent has language as semantic options deriving from the social structure. Using Halliday's concept of meaning potential, and the importance of semantic options for converting behavior options into linguistic options. He stresses the need for teaching linguistic forms from the standpoint of meaning (Munby, 1978). A communicative approach is the second component of the socio-semantic basis of linguistic knowledge. According to Wilkins (1972 as cited in Munby, 1978, p. 25), a language curriculum needs to be based on notional categories, or what we use language for. Having mastered these, a learner would then choose from the set of linguistic forms for the notional categories that he wishes to express. This is the teaching approach that, according to Munby, helps develop communicative competence.

Finally, the third component of Munby's model is the discourse level of operation, where he quotes Sinclair, Coulthard, Forsyth, and Ashby (1972, as cited in Munby, 1978 p. 24) to define discourse as "the level between grammar and non-linguistic organization", which as such is both written and spoken and consisting of terms like speech act, speech event and speech situation. He sees speech act as the central discourse unit, which has both formal features as well as rules of occurrence.

1.6. Widdowson on Communicative Competence

Widdowson (1978) not only supports Hymes's distinction between knowledge and ability for use but also stresses the importance of the distinction for language teaching: "Someone knowing a language knows more than how to understand, speak, read and write sentences. He also knows how sentences are used to communicative effect" (Widdowson, 1978, p. 1). He also distinguishes between language usage, utterances such as "This is my hand.", which are examplificatory expressions but

meaningless as they have no communicative value, and language use, which are utterances that have actual communicative value. He advocates for teaching not only rules of grammar but also rules of use, that is rhetorical rules, as the knowledge of language includes both grammar and communicative competence (Widdowson, 1979).

Even though Widdowson's model is structurally quite similar to the one of Hymes (Fulcher, 1998, p. 283), he introduces new terms for the model components (McNamara, 1996: 59): "rules", that is the knowledge of linguistic and sociolinguistic conventions, which Hymes calls "knowledge", and procedures of interpretation and creation of discourse coherence, or what he calls "communicative competence", which is Hymes's "ability for use". Widdowson (1979) restricts the notion of communicative competence to the knowledge of the rules and introduces the term "communicative capacity" to talk about the procedures of interpretation and creation of discourse coherence.

The previously discussed theories of first language performance have significantly influenced the developments of second language testing. According to McNamara (1996, p. 66), the two most influential adaptations of Hymes' model of performance in the field of second language testing are the ones of Canale and Swain (1980), later adapted by Canale (1983), and Bachman and Palmer (1982), later adapted by Bahman (1990).

1.7. Canale and Swain's Model of Communicative Competence

The first complete model of communicative competence and performance in the field of second language assessment was the one of Canale and Swain (1980), presented in their paper "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing," later adapted by Canale (1983).

Relying on Hymes's model, they adopt the term "communicative competence" and distinguish it from "communicative performance," where communicative performance is seen as instances of language use in real communicative situations (Fulcher & Davidson, 2007, p. 38). They believe that "there are rules of language use that would be useless without rules of grammar" (Canale and Swain, 1980, p. 5) and include grammatical competence under communicative competence, together with sociolinguistic competence and strategic competence.

With regard to grammatical competence, it includes "knowledge of lexical items and of rules of morphology, syntax, sentence-grammar

semantics, and phonology" (Canale and Swain, 1980, p. 29) and is essential for expressing accurately the literal meaning of utterances.

Sociolinguistic competence, on the other hand, has two constituents or sets of rules: sociocultural rules of use and rules of discourse, which are necessary for understanding and interpreting utterances for social meaning. The knowledge of sociocultural rules is essential for determining the appropriateness of utterances in specific contexts, considering "factors such as topics, roles of participants, settings, and norms of interaction" and is also concerned with genre and register. The rules of discourse, on the other hand, are concerned with rules in terms of the cohesion or grammatical links and coherence (appropriate combination of communicative functions) of groups of utterances (Canale and Swain, 1980, p. 30).

Finally, strategic competence consists of verbal and non-verbal communication strategies used to "compensate for breakdowns in communication due to performance variables or to insufficient competence" (Canale and Swain, 1980, p. 30). These communication strategies can be used to compensate for either lack of grammatical competence or sociolinguistic competence.

McNamara (1996, p. 62) focuses on the inconsistencies and contradictions of Canale and Swain's model. Firstly, Canale and Swain intentionally exclude Hymes's "ability for use" from their model and use the term "communicative competence" to encompass language knowledge only, not leaving any room for underlying knowledge or what Hymes calls "ability for use". They state clearly that the decision has been made for two reasons: the notion has not been examined thoroughly in any research and because they "doubt there is any theory of human action that can adequately explicate 'ability for use'" (1980, p. 7). However, they maintain that the ability for use is included in their model under the notion of communicative performance, which in their opinion is the realization of the constituents of communicative competence and "their interaction in the actual production and comprehension of utterances" and consequently define it as "the actual demonstration of this knowledge (constituents of communicative competence) in *real* second language situations and for *authentic* communication purposes" (Canale and Swain, 1980, p. 6, emphasis in original). They also maintain that competence cannot be directly measured but only through performance. McNamara, first of all, argues that they do not address the issue of ability for use at all, and in addition, what they describe as strategic competence

as part of communicative competence is actually an aspect of performance and not language knowledge.

Fulcher and Davidson (2007) believe that the model of communicative competence is nevertheless relevant to language testing for three reasons. Firstly, it implies that tests need to contain both tasks that include actual performance and item types that measure knowledge. Secondly, Canale and Swain maintain that the criticism of discrete point tests in the communicative revolution in the 1970s was not grounded and finally, it is a model that could be used "if it were more 'fine grained'" in second language testing as a basis on which criteria for the evaluation of language performance would be developed in order to interpret the scores in terms of what the language users would be able to do in a non-test situation (Fulcher & Davidson, 2007, p. 39).

1.8. Canale's Adaptation of the Model (1983)

Three years after the publication of their work, Canale (1983) acknowledged its main shortcoming: the failure to include in the model the way in which the model components interact with each other, that is, the ability for use, and expanded the notion of communicative competence to include the underlying skills and not only the knowledge, thus adopting Hymes's standpoint (McNamara, 1996, p. 64). He also chooses to use the term "actual communication" and not "communicative performance," and sees it as a manifestation of knowledge and skills in concrete situations (Fulcher & Davidson, 2007, p. 41).

In his subsequent works (1983), he introduced another component to the model: discourse competence, which, in the Canale and Swain model is a constituent of the sociolinguistic competence. To support his notion of discourse competence, Canale uses Widdowson's notion of coherence. According to McNamara (1996) however, Widdowson considers his notion of coherence as part of communicative capacity whereas Canale's notion of discourse is a matter of knowledge. Furthermore, Widdowson uses the term "ability" referring to what is needed to activate the knowledge and distinguishes it from both knowledge and procedures of interpretation. Widdowson himself recognizes the faultiness of Canale's notion of discourse in the fact that he separates cohesion from coherence, "for cohesion without coherence makes no sense" (Widdowson, 2001, p. 14).

Furthermore, Widdowson (2001, p. 14), maintains that "grammatical" in Canale and Swain's (1980) and consequently in Canale's (1983) models

can be equated with Hymes's "possible," whereas "appropriate" as encompassing the sociolinguistic and the discourse component. "Feasible" and "performed," the third and the fourth of Hymes's judgments, however, do not correspond to any of the components of Canale's model.

Sociolinguistic competence, on the other hand, in Canal's adaptation of the model, refers to sociocultural rules only and is seen as the appropriateness of meaning and form but also as the appropriateness of non-verbal behavior (Fulcher & Davidson, 2007, p. 41). As we can see, grammatical competence in Canale's model includes lexical knowledge, but Canale fails to explain how sociolinguistic competence is involved in deciding on grammatically or lexically appropriate forms (Widdowson, 2001, p. 14).

Finally, Canale elaborates on the notion of strategic competence as well, which in the 1980 model had a compensatory role, in order to include the ability "to enhance the rhetorical effect of utterances" Canale, 1983b, as cited in McNamara 1996, p. 65). Even though his strategic competence is considered a constituent of the communicative competence, in his explanation of the notion, it is more similar to Widdowson's notion of the "exercise of communicative capacity" (McNamara, 1996, p. 65) and as such not a type of knowledge but a "capacity for strategic behaviour in performance" (McNamara, 2015, p. 18). In addition, Bachman (1990, p. 99) argues that both Canale and Swain's and Canale's definition of strategic competence are limited as they do not describe how the strategic competence operates. Finally, Widdowson (2001, p. 14) confirms that sociolinguistic competence in Canale's model is not a competence at all "but the process of relating the others" and Canale fails to define how this is done. If these competences are not brought together in any way, there is actually no communication (Widdowson, 2001, p. 15).

The drawbacks of Canale and Swain's and Canale's model are best summed up by Widdowson (2001, p. 15): "if there is no way, direct or otherwise, of relating this underlying competence to actual performance, it cannot represent the reality of what people do with their knowledge when they communicate".

1.9. Bachman and Palmer's Construct Validation (1982)

The previous models of communicative competence have been further elaborated on by Bachman and Palmer (1982), Bachman (1990) and

Bachman and Palmer (1996, 2010) in order to develop an explicit model of language ability for the purpose of foreign language testing.

Bachman and Palmer started their investigations into the notion of communicative competence as early as in 1982, when they examined the construct validity of some tests of communicative competence and a hypothesized model (Bachman & Palmer, 1982). Following the Canale and Swain (1980) model, where they "developed a framework which not only defines several hypothesized components of communicative competence but also makes the implicit claim that components of communicative competence comprise distinct underlying abilities" (Bachman and Palmer, 1982, p. 449), and maintaining that this one or previous models were not empirically validated, Bachman and Palmer employed a multitrait-multimethod design, with three hypothesized traits and four methods.

The three traits in question are grammatical competence, pragmatic competence, and sociolinguistic competence. Grammatical competence is seen as comprising range and accuracy of morphology and syntax. In this initial model, Bachman and Palmer do not consider phonology and graphology part of grammatical competence because they view these two as channels and not components. The second trait, pragmatic competence that is "the ability to express and comprehend messages" (Bachman & Palmer, 1982, p. 450) comprises vocabulary, cohesion and organization or coherence. Finally, sociolinguistic competence is seen as including distinguishing between different registers, nativeness, and control of non-literal, figurative language and relevant cultural allusions.

The four methods employed in the design are an oral interview, writing sample (a variety of writing tasks ranging from short answers to extensive composition), multiple-choice method and self-rating. Each of the test parts tests each of the traits.

The results of the study indicate that "the components of what they called grammatical and pragmatic are closely associated with each other, while the components they described as sociolinguistic competence are distinct" (Bachman, 1990, p. 86).

1.10. Bachman (1990)

The advantage of having an explicit model is that it can be studied, criticised, its implications understood, its assumptions questioned, and it can then be improved. However, a model needs to be empirically tested: not just criticised

speculatively. Bachman's model has been much referred to, but little operationalised, to date. (Alderson, 1997, p. 5)

The Bachman model (1990), the most influential one, derives from Hymes and Canale & Swain (Alderson, 1997, p5).

Bachman (1990, p. 81) himself confirms that his definition of communicative language ability is consistent with earlier definitions of Hymes (1972, 1973), Munby (1978), Canale and Swain (1980) and Canale (1983). He describes it as "the ability to use language communicatively" involving both the knowledge of language and the capacity for using the knowledge, that is how language is used to achieve certain communicative goals. What makes his framework more refined is his attempt to investigate and describe how the different components of the framework interact with each other and with the context of language use (Bachman 1990, p. 81). Furthermore, it clearly distinguishes "knowledge" from a "skill" and their individual constituents (Fulcher & Davidson, 2007, p. 42).

The model that Bachman (1990) proposes includes three components: language competence, strategic competence and psychophysiological mechanisms.

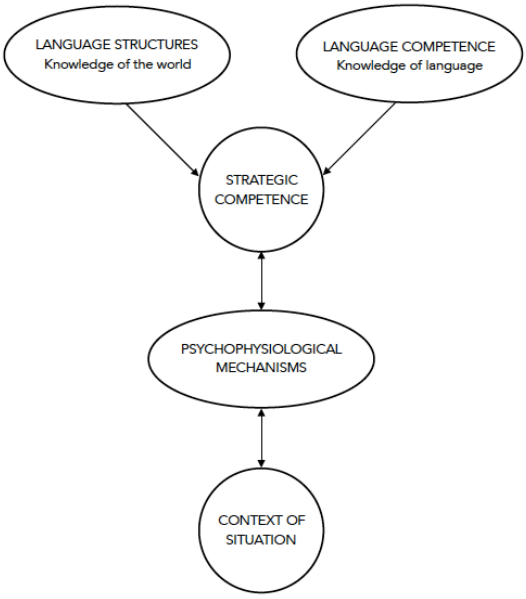

Figure 3. Components of communicative language ability in communicative language use. Reprinted from *Fundamental Considerations in Language Testing*, (p. 85), by L.F. Bachman, 1990, Oxford: Oxford University Press.

Language competence. Bachman's language competence or knowledge of a language is based on the empirical findings of Bachman and Palmer's (1982) study and comprises two types of competence: organizational and pragmatic.

Organizational competence, defined as comprising "those abilities involved in controlling the formal structure of language for producing or recognizing grammatically correct sentences, comprehending their prepositional content, and ordering them to form texts" (Bachman 1990, p. 87) is then seen as comprising grammatical and textual competence.

Grammatical competence includes four independent competences involved in understanding and producing grammatically correct utterances: knowledge of vocabulary, morphology, syntax and phonology/graphology, whereas textual competence is concerned with the knowledge involved in organizing these utterances to form texts (cohesion and rhetorical organization).

The second component of language competence, pragmatic competence is concerned with "the relationships between utterances and the acts or functions that speakers (or writers) intend to perform through these utterances" (Bachman, 1990, p. 89) and is seen as consisting of illocutionary competence, that is competence to use language to express a wide range of functions (ideational, manipulative, heuristic and imaginative) and sociolinguistic competence, the knowledge of sociolinguistic rules of appropriateness of these functions in different contexts: sensitivity to dialect or variety, sensitivity to differences in register, sensitivity to naturalness and ability to interpret cultural references and figures of speech (p. 97).

A difference that Bachman introduces here is the separation of two elements of discourse competence, cohesion, and coherence: while coherence is still seen as part of textual competence, coherence is transformed into illocutionary competence (Fulcher & Davidson, 2007, p. 44).

The language functions are classified into four groups (Bachman, 1990, pp. 92 - 93):

1) ideational functions, as defined by Halliday (1973, p. 20 as cited in Bachman 1990, p. 92): "expressing meaning in terms of our experience of the real world"; "expressing propositions, exchanging information about knowledge and feelings";
2) manipulative functions: affecting the world around us, further divided into instrumental (using language to get things done),

regulatory (controlling or manipulating others) and interactional (forming, maintaining or changing interpersonal relationships);
3) heuristic functions: extending our knowledge of the world around us;
4) imaginative functions: creating or extending our own environment for humorous or esthetic purposes.

Strategic competence. Bachman again stresses the importance of language use as a dynamic process, that is the ability to assess the relevance of information in a specific context and negotiate the meaning to achieve a communicative goal. He defines strategic competence as a "general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task" (Bachman 1990, p. 106). Although it is not to be dismissed as a competence that cannot be measured, he considers it beyond the scope of his book. He does, however, see it as consisting of three components: assessment component, planning component, and execution component.

The assessment component enables us to:
1) identify the information (the language variety, dialect) that is needed to realize a particular communicative goal in a particular context;
2) decide what language competences (native language, second or foreign language) we have at our disposal for achieving the communicative goal;
3) determine which abilities and knowledge we share with the interlocutor;
4) evaluate the extent to which the communicative goal has been realized.

The planning component is what enables us to retrieve grammatical, textual, illocutionary and sociolinguistic items from language competence and formulate a plan on how to use them to get the information needed.

Execution component concerns the relevant psychophysiological mechanisms to implement the plan that is to realize an utterance (Bachman, 1990, pp. 101-103).

Psychophysiological mechanisms. Bachman considers the third component of his model, psychophysiological mechanisms, as necessary to be able to fully describe language use and defines these mechanisms as neurological and physiological processes in the execution phase of language use (Bachman, 1990, p. 107).

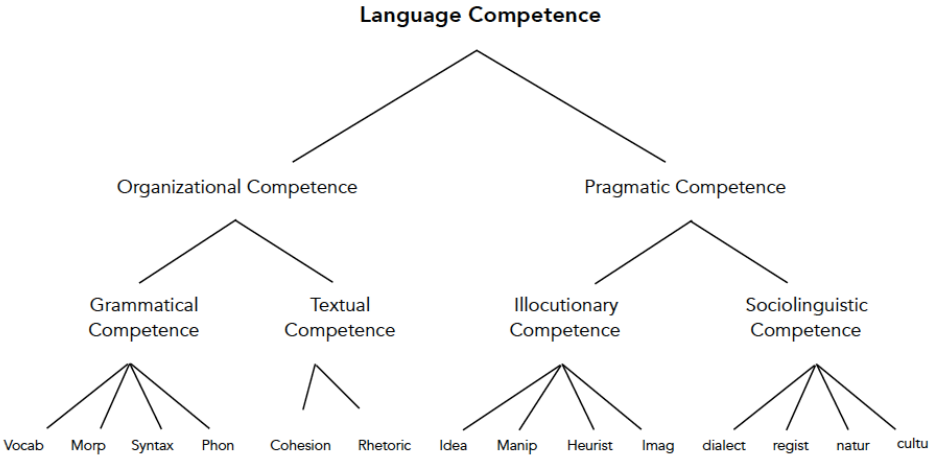

Figure 4. Components of Language Competence. Reprinted from *Fundamental Considerations in Language Testing*, (p. 87), by L.F. Bachman, 1990, Oxford: Oxford University Press.

Bachman's model relies on Canale and Swain's model of communicative competence but reorganizes the constituents and categories of the model. In addition, what Canale and Swain do not consider in their model, and what Hymes calls ability for use, is here expanded and categorized as strategic competence (McNamara, 1996, p. 69). In his overview of Bachman's model, McNamara (pp. 68 - 69) recognizes that the Canale and Swain's discourse competence is here redistributed: cohesion as part of textual competence and coherence as divided between illocutionary competence and strategic competence. Furthermore, the notion of strategic competence is redefined by Bachman and seen as a separate category not characterized as knowledge and no longer part of language competence but the ability for use, "ability, capability or capacity" (p. 69). What McNamara considers problematic is that although Bachman recognizes that strategic competence may be a source of difficulty in interpretation of test scores, the extent to which different tasks involve strategic competence remains unknown. Furthermore, although strategic competence does include cognitive factors, it is not specified which ones or of what type, whether they are the ones that Hymes includes in his model or not (McNamara, 1996, p. 70). Widdowson (2001, p. 16) agrees with McNamara and suggests that

what is missing is "some superordinate node" that would relate components of the model.

Another issue that McNamara addresses is the "overlap between *illocutionary competence* and *strategic competence*" (p. 71, emphasis in original) due to the fact that illocutionary competence, as defined by Bachman (1990, p. 91): "a sentence type whose form is not generally associated with the given illocutionary act, and whose interpretation depends very heavily on the circumstances under which the act is performed," is not merely language competence as he classifies it but may be better defined as part of strategic competence. Although Bachman does include strategic competence in his framework, unlike Canale and Swain (1980), due to the overlap between illocutionary competence and strategic competence, his model has the same difficulties as the one of Canale and Swain.

Despite the difficulties mentioned above, McNamara (1996, p. 71) maintains that Bachman's reorganization of categories of communicative competence and the introduction of strategic competence is a crucial step as it provides a theoretical framework that can be empirically validated in second language testing.

1.11. Bachman and Palmer's Model of Language Ability (1996, 2010)

In their *Language Testing in Practice* (1996) and consequently *Language Assessment in Practice* (2010), Bachman and Palmer revised the original model of language competence proposed by Bachman in 1990. After the first publication of *Fundamental Considerations in language testing* in 1990, Bachman was the most cited single text in the following decade and together with Bachman and Palmer's *Language Testing in Practice* (1996), these two books "declare a distinctive position about language testing, both theoretically and practically" (McNamara, 2003, p. 466).

The most obvious difference that Bachman and Palmer introduced to the Bachman (1990) model are the terms: whereas Bachman (1990) talks about a model of language competence, organizational and pragmatic competence, etc., in Bachman and Palmer (1996, 2010) these are called language knowledge, organizational and pragmatic knowledge, etc.

In the 1996 and 2010 model, language knowledge is part of a bigger structure: Bachman and Palmer start their description of the model with the term "language ability", which is defined as "a capacity that enables language users to create and interpret discourse" (Bachman and Palmer, 2010, p. 33). Language ability then comprises two components: language

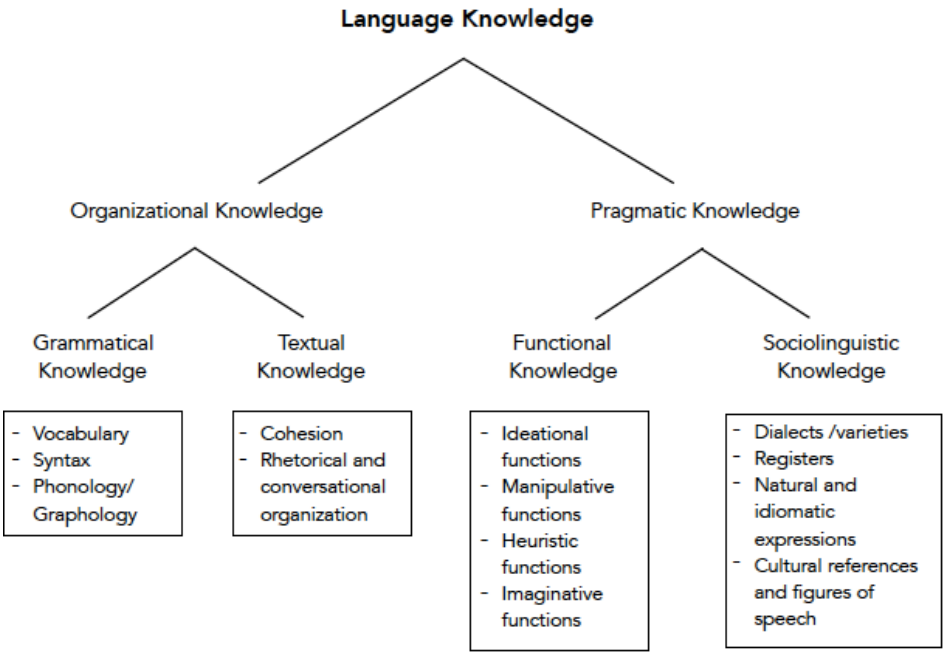

Figure 5. Areas of Language Knowledge. Reprinted from *Language Testing in Practice*, (p. 68), by L.F. Bachman and A. S. Palmer, 1996, Oxford: Oxford University Press.

Language knowledge. "Language knowledge can be thought of as a domain of information in memory that is available to the language user for creating and interpreting discourse in language use" (Bachman & Palmer, 2010, p. 44). Language knowledge is further divided into organizational and pragmatic knowledge.

Organizational knowledge. Similar to the 1990 definition of organizational competence, organizational knowledge concerns formal elements of language necessary for understanding and producing grammatically acceptable utterances, sentences, and texts. It encompasses two components: grammatical knowledge (how individual utterances or sentences are organized), further divided into knowledge of vocabulary, knowledge of syntax and knowledge of phonology/graphology; and textual knowledge (how utterances or sentences are organized to form texts, written or spoken), which is then divided into knowledge of

cohesion and knowledge of rhetorical or conversational organization (Bachman & Palmer, 2010, p. 45).

Knowledge of cohesion is defined as "producing or comprehending the explicitly marked relationships among sentences in written texts or among utterances in conversation" (Bachman & Palmer, 2010, p. 45), whereas knowledge of rhetorical or conversational organization concerns "conventions for sequencing units of information in written texts" (p. 46) and how participants in a conversation manage the exchange of information.

Pragmatic knowledge. Bachman and Palmer (1996, 2010) kept the notion of pragmatic knowledge as defined in Bachman (1990) with some changes to the terminology. What is called "illocutionary competence" in the 1990 model, has been renamed to functional knowledge to include different types of language functions: ideational, manipulative, heuristic and imaginative functions. Pragmatic knowledge is what relates utterances or sentences and texts to the communicative goals (functional knowledge) and the language use setting (sociolinguistic knowledge). The second component of pragmatic knowledge, sociolinguistic knowledge, is what makes it possible for a speaker to create and interpret language appropriate to the context. It further includes the appropriate use of genres, dialects or varieties, registers, natural or idiomatic expressions, cultural references and figures of speech (Bachman & Palmer, 1996, 2010) and is as we can see almost identical to the 1990 definition of sociolinguistic competence, with the addition of the knowledge of genres.

Strategic competence. The 1990 strategic competence and psychophysiological mechanisms have been merged and defined as strategic competence in the 1996 and 2010 model. It is considered to be consisting of metacognitive strategies necessary for language use. These strategies are divided into goal setting appraising and planning. Areas of metacognitive strategy use are listed in Figure 6.

Goal setting (deciding what one is going to do)
- Identifying the language use or assessment task to be attempted
- Choosing one or more tasks from a set of possible tasks (sometimes by default, if only one task is understandable)
- Deciding whether or not to attempt to complete the tasks(s) selected.

Appraising (taking stock of what is needed, what one has to work with, and how well one has done)
- Appraising the characteristics of the language use or assessment task to determine the desirability and feasibility of successfully completing it and what resources are needed to complete it
- Appraising our own knowledge (topical, language) components to see if relevant areas of knowledge are available for successfully completing the language use or assessment task
- Appraising the degree to which the language use or assessment task has been successfully completed.

Planning (deciding how to use what one has)
- Selecting elements from the areas of topical knowledge and language knowledge for successfully completing the assessment task
- Formulating one or more plans for implementing these elements in a response to the assessment task
- Selecting one plan for initial implementation as a response to the assessment task.

Figure 6. Areas of Metacognitive Strategy Use. Reprinted from *Language Assessment in Practice*, (p. 49), by L.F. Bachman and A. S. Palmer, 2010, Oxford: Oxford University Press.

The most significant changes to the 1990 model (McNamara, 1996, p. 72; Fulcher and Davidson, 2007, p. 45) are the introduction of the so-called affective schemata, as one of the attributes that are not part of language ability, strategic competence conceptualized as a set of metacognitive strategies and some changes to the terminology, for example, the term "topical knowledge", previously referred to as "knowledge structures". Affective schemata, defined as (Bachman & Palmer, 2010, p. 42) "feelings we associate with specific kinds of topical knowledge", is the first attempt "to deal explicitly in a model of language communicative ability with the aspect of [Hymes'] *ability for use* which relates to affective or volitional factors" (McNamara, 1996, p. 74, emphasis in original). Bachman and

Palmer (2010, p. 42) justify their decision to add affective schemata to their model:

> The affective schemata provide the basis on which language users appraise, consciously or unconsciously, the characteristics of the language use task and its settings in terms of past emotional experiences in similar contexts. The language user's affective schemata, in combination with the characteristics of the particular language use task, determine, to a large extent, his affective response to the context. The affective responses of language users may thus influence not only whether they even attempt to use language in a given situation, but also how flexible they are in adapting their language use to variations in the setting.

McNamara (1996, p. 74), however, believes that Bachman and Palmer's view of affective schemata is contradictory as it is seen of different relevance in different language performance situations. Namely, they use an example of an emotionally charged topic as a type of topic that could potentially disable a student from performing their best while, on the other hand, maintain that "emotional responses can also facilitate language use" (Bachman & Palmer, 2010, p. 42). The introduction of affective schemata is still recognized as "potentially far-reaching development" (McNamara, 1996, p. 75) and "a major step in making the model much more complex" (Fulcher & Davidson, 2007, p. 45).

Other attributes which are not related to language knowledge that Bachman and Palmer include in the model include "personal attributes" (age, sex, nationality, resident status, length of residence, native language, level and type of general education, and type and amount of preparation or prior experience with a given assessment), "topical knowledge" (referred to as "knowledge structures" or "knowledge of the world" in the 1990 model) and "cognitive strategies". As previously explained, strategic competence is now seen as a set of metacognitive strategies or "higher-order processes that explain the interaction of the knowledge and affective components of language use" (Fulcher and Davidson, 2007, p. 45). "Affective schemata" and "cognitive strategies" are the two cognitive components of the ability for use, unlike "personal attributes" and "topical knowledge."

Finally, each of the metacognitive strategies listed under "sociolinguistic competence" is seen as interacting with the model components (McNamara, 1996, p. 75), making the model much more complex and refined than any previous one.

1.12. Conclusion

Each of the described models has contributed to defining communicative competence to a certain extent. A summary of the models is illustrated by Figure 7:

Writer	Model of knowledge	Model of performance	Actual use
Chomsky	competence	performance	
	grammatical competence	pragmatic competence	actual performance
Hymes	communicative competence		performance
	knowledge	ability for use	
Halliday	[rejects the distinction between 'competence' and 'performance']		
Campbell & Wales	communicative competence		performance
Munby	communicative competence		performance
Widdowson	communicative competence	communicative capacity	
	rules	procedures	
Canale and Swain	communicative competence	[unable to be modelled]	communicative
Canale	communicative competence		actual communication
	knowledge	skill	
Bachman (1990)	communicative language ability		language use
	language competence/ knowledge	strategic competence	

Figure 7. Models of knowledge and performance. Adapted from *Measuring Second Language Performance*, (pp. 54,56,58), by McNamara, T. F. (1996). Measuring Second Language Performance. London: Longman.

Most often, these models focus on language competence or language knowledge, leaving the notion of communicative competence either unexplored or unclear. As Widdowson (2001, p. 17) points out, the complex process of communicative competence is analyzed in terms of different components, but the dynamic interrelationships between the components remain undefined. What we are left with then are individual

features and using such a model in teaching or testing would mean returning to Chomsky's structuralist approach. Widdowson then goes on to claim that communicative competence cannot be measured and as the most salient feature that is both teachable and testable is linguistic competence. He finds an appropriate definition of linguistic competence in the works of Halliday and his notion of "meaning potential" or realization of the potential, where linguistic knowledge is not seen as separate from other components and with no defined relationship, but its combination with other components is a realization of its own potential. To make language testing "art of the possible" again, he proposes focusing on the "meaning potential," that is, how "a general linguistic capacity for communication … gets realized in the particular circumstances of real-life communication" (p. 19).

Chapter Two
History of Foreign Language Assessment

2.1. Introduction

Foreign language testing has come a long way since its first instances occurred as early as in the 1900s. Over the period of more than 100 years, the field has been strongly influenced by the development of other disciplines or fields such as psychology and foreign language teaching and has changed radically over the years.

The history of language testing development can be roughly divided into three periods (Spolsky, 1978, p. v):

- the pre-scientific era (1913-1945),
- the psychometric-structuralist era (the 1960s) and
- the integrative-sociolinguistic era (1975 on).

2.2. The pre-scientific era (1913-1945)

The most important characteristic of the first tests was that there was no scientific background or an underlying linguistic theory: "the emphasis was on language use, though some attention was paid to form in the grammar and phonetics sections" (Weir, 2005a, p. 5). To illustrate the development of language testing over the years, Weir (2005a) uses the Cambridge ESOL Certificate of Proficiency in English (CPE), which, according to him, is the first serious test of English as a foreign language still in existence. The CPE was initially divided into two parts: Written and Oral and included tasks such as translation, essay writing, literature, dictation and reading, and conversation. At the time, phonetics was the main concern of linguistics and language studies and that reflected in the field of assessment. In addition, the prevalent method of language teaching was grammar translation, where students are required to translate texts from one language to another as a means of mastering the language and growing intellectually.

Weir also points to the different approaches to validity and reliability in the USA and the UK in that period. Whereas in the UK and consequently Europe, the stress was on what was tested, the American

approach was more focused on how it was tested. Although the general approach to validity and reliability was that they were taken for granted (Davies, 2003, p. 356), in the European approach the main concern was construct validity even if the psychometric qualities of the test were not perfect, that is, it was only important that the test "felt fair" (Spolsky, 1995 as cited in Miyata-Boddy and Langham, 2000, p. 77). Spolsky (1978) criticizes this approach to language testing because of its lack of concern for objectivity or reliability. The greatest authority of the time were the examiners or teachers, who, using their experience and knowledge, would judge a student's performance essay. On the other hand, the American approach sacrificed some aspects of validity in the pursuit of reliability (Weir, 2005a).

2.3. The psychometric-structuralist era (the 1960s)

What Spolsky (1978, p. v) calls the "psychometric-structuralist era" started in the 1960s and was marked by a growing interest in improving test reliability. The name itself reflects the significance of the influence of the structural linguistics that, in that period, identified elements of language to be tested. The loudest of the structuralists to support this approach was Lado (1961), who, according to Choi (1989, p. 97), saw language as "a finite system of an exhaustive list of items" as other structuralists. In other words, by means of a structural contrastive analysis, each language skill can be broken into small parts and items that could be tested in order to provide information about the candidate's ability to handle that particular item (Miyata-Boddy and Langham, 2000, p. 76) and, in that way, provide information about the candidate's knowledge of the language assuming "that knowledge of the elements of a language is equivalent to knowledge of the language" (Morrow, 1981, p. 11). This approach is known as "discrete point testing," meaning that it tested individual language components separately, for example, there were separate tests of grammar and separate tests of vocabulary.

Lado (1961) was also the first one to propose a scientific approach to langue testing. He saw a language as a system of habits of communication, such as form and meaning, and distribution at several levels of structure (sentence, clause, phrase, etc.). Considering that an individual is not aware that their use of a language is based on a complex system of habits, they use the same habits as they would in their own native language. Therefore, the native language habits are transferred to the foreign language use. Where and when the two languages (the native

language and the foreign language) are similar or the same, the individual communicates successfully. However, where they are different, there is a gap and the individual needs to learn the new units and patterns. For that reason, the learning problems can be predicted by comparing the two languages and testing the problems meant testing the language. This approach was described by Morrow (1981, p. 11) as atomistic, one that breaks down the concept of knowing a language into isolated segments.

At the same time, importance was given to the psychometric characteristics of a test. Psychometrics, as the science of the measurement of cognitive abilities, started emerging in that period and it focused on improving the reliability of tests, or "consistency of estimation of candidates' abilities" (McNamara, 2015, p. 14). As McNamara explains, tests consisted of a number of small items aimed at assessing the same language component, for example, a test of vocabulary would consist of a number of multiple-choice items, each aimed at testing vocabulary. The most important test characteristics were that items be considered objective as it was considered the basis of reliability, and the test reliable with concurrent validity (Shohamy & Reves, 1985, p. 48).

Consequently, testing literary and cultural knowledge became less important. In the 1960 revision of the CPE, there were major changes to the writing part: the phonetics was left out to introduce Use of English and English Language. The test takers could choose whether or not to take the English Literature part of the test. The Use of English was assessed using multiple-choice tests, which was a result of the growing interest in the test objectivity and internal consistency (Weir, 2005a). At the same time, in the United States, the aim was to deliver a large number of exams each year, and for that reason, multiple-choice tests were used. This required the development and production of automatic marking machines to mark multiple-choice tests rapidly.

As Morrow (1981, p. 11) points out, there are several shortcomings to this approach. It quickly became evident that it was impossible to design discrete items that would test one language segment only. More importantly, the main disadvantage of this approach was the wrong assumption about language knowledge being a sum of individual elements of language and the fact that the ability of a learner to combine these elements in order to communicate was not taken into account. Similarly, Choi (1989, p. 98) maintains that the tests did not reflect natural or communicative language-use contexts.

Spolsky (1978, p. vi) also points out two major drawbacks of this approach: firstly, since this type of items required a written response, they

were limited to reading and listening, and the types of items used did not reflect newer ideas about language teaching and learning. Secondly, "a new set of experts added notions from the science of language to those from the science of educational measurement."

There were few advantages, however, the most important one being that both course developers and testers needed to have a clear idea of the language elements that they wanted to teach and test (Choi, 1989, p. 98), which was not the case in the pre-scientific period. Another advantage was that this kind of tests provided easily quantifiable data and seemed to be more reliable (Miyata-Boddy and Langham, 2000, p. 76).

2.4. The integrative-sociolinguistic era (1975 on)

The integrative-sociolinguistic era started in the early 1970s with the introduction of integrative and pragmatic tests. According to McNamara (2015, p. 14), this was because the existing types of tests did not meet the need of "assessing the practical language skills of foreign students wishing to study at universities in Britain and the US" as they focused on the knowledge of the formal linguistic system, that is, the individual language features, and not how that knowledge was used to actually communicate. At the same time, there was a major shift in the field of language teaching: the communicative language teaching, which would, in the years to come, provoke other changes in the field of language testing.

2.4.1. Integrative and pragmatic tests

The first tests to emerge in this period were the integrative ones. The name itself stresses the need for integrating individual language components as well as the context in which the knowledge of the language is used.

A new approach to testing was offered by Oller (McNamara, 2015, p. 15), who tried to define language tests in terms of the "language processing operations required of learners" (p. 92). He proposes a pragmatic test, as opposed to the discrete point test, where the learner needs to "process sequences of elements in a language that conform to the normal contextual constraints of that language and which requires the learner to relate sequences of linguistic elements via pragmatic mapping to extralinguistic context" (Oller, 1978, p. 38) In other words, he advocates

for measuring the learner's ability to use and combine different language features (grammatical, lexical, contextual and pragmatic knowledge) in the way they are used in real-life situations: considering the constraints of the context as well. Oller's proposal later became known as the Unitary Competence Hypothesis. As the first tests in the psycholinguistic-sociolinguistic era, the integrative ones, were quite expensive and time-consuming as they implied speaking in oral interviews, composing written texts, etc., Oller proposed what he considered more efficient tests: cloze tests (a gap-filling test) and dictation, which according to him, tested the learner's ability to combine the individual language features unlike the discrete point tests.

According to Bachman (1990), Oller was one of the first linguists, after Carroll (1941) to do construct validation research: research into the relationship between performance on language tests and the abilities that underlie that performance. However, Bachman maintains that even though in his initial research Oller believed to have discovered a "g-factor," a unitary trait, or "general language proficiency," after additional studies, Oller himself admitted that his hypothesis was wrong (Bachman, 1990, p. 6).

Even before Oller himself admitted to having been wrong, drawbacks of this type of tests became evident. According to McNamara (2015, p. 16), cloze tests measured the same type of knowledge, e.g., vocabulary and grammar, as the discrete point tests, and not communicative language skills.

Morrow (1981, p. 15) maintains that cloze and dictation are tests of language competence and help determine the level of language proficiency of a learner. However, neither of the two types actually tested the learner's ability to use the language, that is how they would perform when they actually needed to transform their competence (the knowledge of how language works) into performance. This contrast between competence and performance became a burning issue with the birth of communicative language testing, several years later.

The test types that Oller proposed were also an object of study of Alderson's Ph.D. thesis (1978). According to Weir (1990, p. 4), Alderson demonstrated that results of cloze tests depend on the number of deleted items and where the deletions began. Furthermore, Weir points out that the fact that this type of test provides information on candidate's linguistic competence only but not on their performance, that is, what they can or cannot do (p. 6). This limitation of discrete point and integrative tests gave rise to communicative language testing.

Nevertheless, Oller's proposal was still pioneering work as it introduced construct validation in language testing research.

2.4.2. Communicative language testing

A language acquisition and teaching theory that heavily influenced the field of language assessment was the communicative language teaching. In the early 1970s, the existing approaches to language teaching, the audio-lingual method, in particular, did not satisfy the needs of the large number of immigrant workers in Europe. According to Savignon (1983, p. 1), it was this, together with the rich British linguistic tradition that included both social and linguistic context in the language behavior description, that gave birth to the development of the notional-functional concepts of language use and consequently communicative language teaching, where the main focus is on the communicative needs of the learner. It was the period when the Van Ek published his first notional-functional syllabus, *Threshold Level English* (1975), which was based on the analysis of communicative needs of learners. It was here that the term "communicative" was attached to such a syllabus for the first time, and through it, it became the underlying principle of the CEFR as well.

The most influential of the theories that emerged in this period and influenced the field of language testing are the theory of communicative competence and performance by Hymes (1972), Halliday's (1972) theory of competence and performance, Campbell and Wales's theory of competence and performance Mynby's (1978) theory of communicative competence, Widdowson's (1978) view of communicative competence, then Canale and Swain's (1980) model of communicative competence and its adaptation by Canale (1983) and finally Bachman's (1990) model of communicative language ability as well as its modification by Bachman and Palmer in 1996 and 2010. These have been discussed in Chapter One of Part Two.

The main trait of the tests of this period, which were influenced by different models of communicative competence, was the idea to test not only grammar but also the ability to use the language, that is performance (Weir, 1990, p. 7). Consequently, the importance of social functions and aspects of language were recognized.

Furthermore, communicative language testing was basically a reaction against the importance of the roles of validity and reliability, especially in the United States during the 1960s (Fulcher, 2000, p. 483).

According to McNamara (2015, pp. 16 - 17), the main features of communicative language tests are:

1. They were performance tests, requiring an assessment to be carried out when the learner or candidate was engaged in an extended act of communication, either receptive or productive or both.
2. They paid attention to the social roles candidates were likely to assume in real-world settings, and offered a means of specifying the demands of such roles in details.

A definition of communicative language testing is also provided by Harding (2014, p. 187): "a 'communicative' language test is often understood to be a test where language is assessed in context, often involving authentic tasks and interaction".

Morrow (1981, pp. 16 - 17) on the other hand lists the features that conventional tests do not measure: interaction, unpredictability, context (of situation and linguistic), purpose, performance, authenticity and behavioral outcome and goes on to propose what he calls "The Promised Land", that is a test of communicative ability.

Among the features that according to him a test measuring communicative ability needs to have are: a) to be criterion-referenced, that is linked to authentic tasks, b) to have content, construct and predictive validity, c) to rely on qualitative modes of assessment and d) not to focus on objectivity; reliability needs to be subordinate to face validity. In his opinion, it was the task of communicative language testing to redefine the notions of reliability and validity, where reliability would become subordinate to face validity (pp. 17 - 18).

A test that he proposes, as it has all the mentioned features, is a performance-based test, which measures what a learner or a candidate can actually do. This was one of the first descriptions of communicative language testing (Harding, 2014, p. 188).

Similarly, according to Fulcher (2000, pp. 489 - 493), communicative tests need to involve performance, authenticity and to be scored on real-life outcomes. As he rightly recognizes, the following became the "buzzwords of early communicative language testing": "real-life tasks," "face validity," "authenticity" and "performance." However, these buzzwords or terms remained undefined for a long period and provoked discussions by other linguists.

Bachman (1990, pp. 301 - 302), for example, differentiates between what Morrow considered "real-life" authenticity from the so-called

"interactional-ability" approach to defining authenticity. Namely, the idea behind "real-life" authenticity is to mirror the real world, and a test is considered authentic if it replicates non-test language performance. On the other hand, the "interactional-ability" approach to defining test authenticity sees authenticity as "a function of the interaction between the test taker and the test task." Harding (2014, p. 189) recognized this as "two separate forms of CLT [communicative language testing] that functioned under the same banner": the real-life approach, which is atheoretical but authentic, and the interactional authenticity approach, which considers the underlying traits of communicative ability and maintains that both approaches are still employed when designing tests.

Whichever of the approaches to authenticity is adopted, there remains the problem of sampling appropriate tasks for the test and generalization of results. To be able to make inferences about the candidate's ability outside the test situation, there would need to be quite a high number of tasks and contexts. However, according to Weir (1990, p. 15), the specificity of tasks makes it nearly impossible to generalize.

Bachman (1991, as cited in Miyata-Boddy, 2000, p. 81) addresses the same problem and maintains that in order to be able to generalize the results we need to be sure that a) the language test corresponds to the language abilities in non-test language use, and b) that the characteristics of the test tasks correspond to the features of a target language use context.

Other issues with the communicative approach to language testing have been raised by other authors. Weir (1990, p. 13), for example, maintains that the holistic and qualitative approach to the assessment of communicative language testing requires a different view of reliability. He also points out that the nature of the criterion-referenced approach to testing communicative language ability requires attention. Another issue he raises is inter-rater reliability (p. 32), where as a solution he proposes clearly established assessment criteria as well as rater training.

In time, different problems with and drawbacks of communicative language testing, which are also replicated in performance-based assessment, have been recognized and addressed. These will be further discussed in the following chapter in relation to performance-based assessment, while issues pertinent to the study will be fully addressed in the Chapter One of Part Three: Methodology.

Harding (2014, p. 190) concludes that "much mainstream language testing is now 'communicative' in the sense that it draws on existing theories of communicative language ability and utilises 'real-life' tasks,

paying heed to authenticity and often including interactive performance."

2.5. Conclusion

As we have seen, different types of tests have been proposed over the years. At a certain point in time, each one of them was a burning issue only to be replaced with a different, new one. However, most of them have stayed, and the existing tests most often combine the different types of items and testing approaches proposed over the years (Morrow, 1981). The fact remains that test authenticity and performance were proposed (Morrow, 1981; McNamara, 1996; Fulcher, 2000) as one of the main features of communicative language tests.

Chapter Three
Performance-Based Assessment: Past and Present

3.1. Defining Performance-based Assessment

The origin of performance-based assessment does not lie in applied linguistics but has a long history in other fields (McNamara, 1996, p. 6). When describing performance-based assessment, authors have often described it in reference to the traditional test methods, the pencil-and-paper tests, to say not only what it is but also what it is not and to stress its advantages as a test method. Kane, Crooks, and Cohen (1999, p. 5) believe that "the direct assessment of some of our most important educational goals seems to require that complex performances be evaluated."

As Baker, O'Neil, and Linn (1993) point out, performance tests in general have also been referred to as "direct tests", "authentic tests" and "performance-based assessment" (e.g. Linn, Baker, Dunbar, 1991; Wiggins 1989) where they were defined as tests that "involve the performance of tasks that are valued in their own right" (Linn et al., 1991, p. 15). Baker et al. (1993, p. 1210) define performance-based assessment as "a type of testing that calls for demonstration of understanding and skill in applied, procedural, or open-ended settings", while McNamara (1996, p. 6) defines it as "an actual performance of relevant tasks (…), rather than more abstract demonstration of knowledge, often by means of pencil-and-paper tests."

According to McNamara (1996, pp. 9 - 10), performance as a test method typically involves the performance process illustrated in the figure below, which is observed and judged using an agreed judging process.

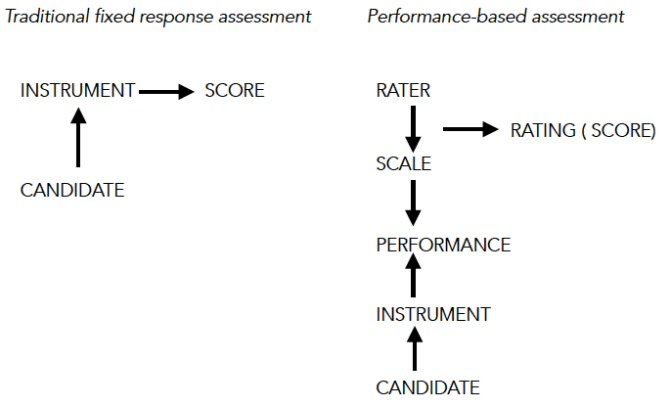

Figure 8. The characteristics of performed assessment. Reprinted from *Measuring Second Language Performance*, (p. 9), by T. F. McNamara, 1996, London: Longman.

Baker et al. (1993, p. 1211) list the following as features of performance-based tests:

1. Use of open-ended tasks.
2. Focus on higher order or complex skills.
3. Employment of context sensitive strategies.
4. Use of complex problems requiring several types of performance and significant student time.
5. Can be group or individual performance.
6. May involve a significant degree of student choice.

Similarly, Slater (1980, pp. 14 - 15) proposes three features of performance tests as opposed to non-performance tests: *stimulus characteristics of tests*, *response characteristics of tests* and *surrounding conditions*. The first feature, the stimulus, can be defined as whatever initiates the examinees behavior, for example, in what McNamara (1996) calls the work sample approach, when a student in a medical technician training course, in a role-play, answers a phone call from a near-hysterical parent whose child has just swallowed a medicine. The stimulus, according to Slater, can "vary in its fidelity or resemblance to naturally occurring, real life stimuli" (p. 14). The response characteristic, the second

feature, encompasses two types of responses or behaviors: respondent behaviors, where the examinee chooses from a set of options; and operant respondents, where there are no preconceived, offered answers. The second behavior again offers two options: whether to observe the *process*, for example, the way the medical trainee deals with the near-hysterical parent, or the *product* of the performance, for example, whether or not the trainee manages or not to save the children's life. Finally, the third feature are the environmental conditions under which a task is performed.

Kane et al. (1999, pp. 6 - 7) believe that the defining feature of performance-based assessment is their similarity with the performance that is of interest in the sense that they either samples of the performance of interest, that is the target domain or simulations of that kind of performance. Another feature, though not necessarily a defining one, is the fact that performance assessment often involves extended-production tasks and not short answers.

McNamara (1996) and Linn, et al. (1991) agree that a significant feature of performance-based tests is their authenticity or simulation of real life, a reason for which many other authors have referred to performance-based tests as authentic tests. This, however, raises the question of the degree to which a test needs to be realistic in order to be considered a performance-based test. The issue of authenticity of performance-based tests will be further addressed in relation to second language performance assessment.

The variety of terms used to refer to performance-based assessment summarizes their features: it is an assessment of a task performance, or performance product, using predetermined judging criteria, in an environment or surrounding which is as realistic or authentic as possible. These same features can be said to characterize the second language performance assessment as well and will be dealt with in more detail in below.

3.2. Second Language Performance Assessment

3.2.1. Defining second language performance assessment

The first instances of second language performance assessment can be traced in as early as the 1950s, at the very beginning of the scientific period in language testing, on occasions where it was felt necessary to complement discrete points testing with productive language skills

testing. However, it was only with the development of communicative language testing, in the 1970s, that performance-based assessment evolved as support for its development was found in the theories of communicative competence (McNamara, 1996). As McNamara (2015, p. 16) points out, one of the main features of communicative language tests was that "they were performance tests, requiring assessment to be carried out when the learner or candidate was engaged in an extended act of communication, either receptive or productive or both."

In terms of test method, McNamara (2015, p. 5) distinguishes two broad test categories: the traditional paper-and-pencil language tests and performance tests. Whereas paper-and-pencil language tests most often test only one or some of the language components or receptive skills, that is, listening and reading, the main feature of performance tests is that performance of tasks is actually expected from candidates (McNamara, 1996, 2015). Similarly, Bachman (1990, p. 77) defines a performance test as one where "the test takers' performance is expected to replicate their language performance in non-test situations" and as an example uses the oral interview.

Two traditions of second language performance assessment can be identified (McNamara, 1996, pp. 6, 25), the first one being the "work sample" approach, the use of the techniques of performance assessment developed in non-language contexts. This tradition developed in response to the need to assess the knowledge of English for selection purposes, for example, students wanting to study at universities in English speaking countries. In this approach, the target of the assessment is the performance itself, with a special focus on Hymes' (1967, 1972) sociolinguistic competence, as described in Chapter One of Part Two. The second one, where performance in a second language is seen as a complex "cognitive" achievement, focuses not only on the quality of performance but also on what the performance reveals about the underlying state of language knowledge.

Messick (1994, p. 13) however distinguishes between performances and products, as well as between the assessment of performance per se, which he calls "task-driven" assessment, and performance assessment of a construct, which he calls "construct-driven" performance assessment. In the first case, the target of assessment is either performance per se or the product of the performance. In the second case, however, the performance is merely a vehicle of assessment and the performance or observed behavior is used to make inferences about the actual target, which are constructs such as knowledge and skills underlying the

performance (Messick, 1994, p. 14). This approach can be traced back to Lado (1961), who argued for a structuralist approach to testing, as discussed in the previous chapter, and which has in time become "the most common approach to general-purpose performance assessment" (McNamara, 1996, p. 26).

McNamara (1996, p. 43) supports this distinction and proposes the terms "strong" and "weak" language performance tests, where "strong" language performance tests are what Messick calls task-driven performance assessment. In such tests, a second language is simply a medium used to perform a task, whereas the actual target is the same as in Messick's task-driven performance – the task performance. McNamara's "weak" language performance tests are actually Messick's construct-driven performance tests, where the target or focus of assessment is language performance, that is to elicit language on the basis of which inferences on the second language knowledge could be made. This distinction will also be addressed in relation to task-based assessment.

We can see two different approaches or traditions in the second language performance testing – one focusing on the performance itself, and the other one, where the focus is what underlies the performance.

3.2.2. Development of second language performance assessment before communicative competence theories

According to McNamara (1996, p. 24), the first linguists to advocate for what is now known as second language performance assessment were Carroll and Davies, two decades before the first theories of communicative competence. Carroll (1961 [1972] as cited in McNamara, 1996), disagreed with Lado's structuralist approach to language testing, which was based on discrete-point testing, and advocated for an "integrative approach" which would involve performance. The reason for this was the need to assess the knowledge of English of foreign students, without knowing anything about their background knowledge and independent of their first language. The purpose was to establish "how well the examinee is functioning in the target language, regardless of what his native language happens to be" (Carroll, 1961 [1972], p. 319 as cited in McNamara, 1996, p. 28). This view implicitly involved performance testing, that is, performance on tasks which required an integrated use of different aspects of language knowledge and skills and pointed out the differences between what is now known as proficiency

test as opposed to an achievement test. Morrow (1981, p. 15) confirms that Carroll's view of language is a complete opposite to Lado's approach and denies Lado's idea of atomistic nature of language as a basis for language testing.

Davies (1968, 1977, as cited in McNamara, 1996, p. 28) on the other hand, made the distinction between proficiency and achievement tests more explicit and advocated for tests on the basis of which inferences about candidates' knowledge and their future performance could be made, that is proficiency tests. McNamara, however, maintains that, from Davies's point of view, a proficiency test does not necessarily need to be a pure performance test and would require "a demonstration of knowledge, but it will also demand demonstration of skill in performance" (p. 28), that is an application of the knowledge. The "predictive" nature of performance tests consequently raised the issue of predictive validity, which will be discussed in Chapter Two of Part Two.

3.2.3. Underpinning of second language performance assessment in communicative competence theories

With the appearance of the first communicative competence theories, in the early 1970s, a new approach to performance testing came into view. After paramount criticism of the tests that preceded the communicative language theories, language assessment followed the recent development in linguistic theories of the time.

One of the proponents of communicative language teaching and consequently testing, Savignon (1972, p. 11) proposed assessment of language skills "in an act of communication," stressing the importance of the context, that should resemble real life as much as possible.

Munby's theory of communicative competence (discussed in Chapter One of Part Two) influenced significantly the development of performance-based testing in the United Kingdom as he distinguished between actual performance and the underlying rules of performance (Munby, 1978), which subsequently became one of the main features of performance-based testing (McNamara's "weak" view of performance and Messick's performance as a vehicle of assessment, as discussed above).

A considerable contribution to the development of performance-based testing was given by Morrow (1981), who, despite the fact that his main interest is the notion of communicative competence, advocates for the use of performance-based tests within communicative language testing. He

suggests that the starting point of a communication test be "the measurement of what the candidate can actually achieve through language" (Morrow, 1981, p. 17) and that what needs to be tested is

> the candidate's ability to actually use the language, to translate the competence (or lack of it) which he is demonstrating into actual performance 'in ordinary situations', ie actually using the language to read, write, speak or listen in ways and contexts which correspond to real life (Morrow, 1981, p 16).

He goes on to list the characteristics of a test measuring communicative ability, as outlined in the previous chapter: a) to be criterion-referenced, that is linked to authentic tasks, b) to have content, construct and predictive validity, c) to rely on qualitative modes of assessment and d) not to focus on objectivity; reliability needs to be subordinate to face validity. The most significant of the features here is the first one as it refers to the actual candidate performance of specific activities or "What can this candidate do?" (Morrow, 1981, p. 18).

Furthermore, in Morrow's opinion (1981, p. 17), one of the features of language use that tests before the communicative approach to testing did not have, or test is performance. He refers to Chomsky's notion of "competence" to claim that it has been the basis of most language tests, without paying any attention to the context and features of candidate's performance.

Although he does discuss potential problems of performance assessment, he does conclude that "it is performance tests which are of most value in a communicative context" (p. 19).

As performance-based assessment found its theoretical underpinning in communicative language theories, the challenges that originate from communicative language testing, in general, are replicated in performance-based testing.

3.3. Task-based Performance Testing

In the 1980s, performance assessment became progressively identified with task-based assessment (Ross, 2011), or what McNamara (1996, p. 43) calls "strong" performance-based tests and Messick (1996, p. 43) task-driven performance assessment. In his *Task and Performance Based Assessment,* Wigglesworth (2008 p. 111) describes the relationship between task and performance-based assessment arguing that there is little agreement on the relationship between task-based assessment to

language assessment in general but maintaining that the role of tasks has nevertheless provoked discussions by different linguists. A clear and comprehensive definition of task-based assessment in relation to language assessment, in general, is given by Brown, Hudson, Norris and Bonk (2002, as cited in Wigglesworth, 2008, p. 111) who define task-based language testing as a

> subset of performance based language testing, clearly distinguishing between performance based testing, in which tasks are merely vehicles for eliciting language samples for rating, and task-based performance assessments in which tasks are used to elicit language to reflect the kind of real world activities learners will be expected to perform, and in which the focus is on interpreting the learners' abilities to use language to perform such tasks in the real world.

This distinction made by Brown, Hudson, Norris and Bonk (2002, as cited in Wigglesworth, 2008, p. 111) is strikingly similar to the ones by Messick and McNamara. What is different however are the terms used. Similarly, Bachman (2002) uses the term "task-based language performance tests" and refers to the same distinction using the terms "task-centered" and "construct-centered" approaches, asserting that what underlies different definitions and discussions of task-based assessment is the inferences about "underlying language ability" or capacity for "language use" or "ability for use" (Bachman 2002, p. 454).

Wigglesworth (2008, p. 112) believes that both second language task and performance assessment developed simultaneously with the theoretical models of language proficiency and as such it evaluates the candidates "on a much greater range of language skills than those traditionally measured by the ore discrete, paper and pencil-based tests."

An issue that in Bachman's opinion needs to be addressed in relation to task-based assessment, which has also been addressed by Weir (1990) and Morrow (1981) is the problem of "justifying inferences from test performance," which will be discussed in more detail in the following chapter.

Finally, it may be useful to define "task" as different interpretations of the notion can be found in the literature written on the topic.

Skehan (1996, p. 38), for example, defines a task as "an activity in which meaning is primary, there is some sort of relationship to the real world, task completion has some priority, and the assessment of task performance is in terms of task outcome." Bachman (2002, p. 458) distinguishes two approaches to defining tasks, the first one being: "as

those activities that people do in everyday life and which require language for their accomplishment" (Norris, Brown, Hudson and Yoshioka, 1998, p. 331, as cited in Bachman 2002, p. 458). The second approach to defining a task that Bachman identifies is the one that takes into account the goal and the setting of the task, and as such is called "language use task": "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation" (Bachman & Palmer, 1996, p. 44).

Chalhoub-Deville (2001, p. 211) however maintains that, due to Bachman's (1990) early definition of "test method" to mean what he in 2002 defines as "task", the term "task" when used in language testing has been confounded with "test method" and encompassed a number of different exam formats. In second language acquisition and instructional domains, however, the term task has been associated with "activities that simulate those in the real-world outside the classroom and promote interlanguage development" (p. 211). In the 1990s, however, linguists working in the field of second language assessment started using the term "task" increasingly to refer to "directly to what the test-taker is actually presented with in a language test, rather than to an abstract entity" (Bachman & Palmer, 1996, p. 60).

The hypothesis underlying all the above definitions according to Bachman (2002, p. 454) is that the inferences to be made from task-based assessment concern the underlying "language ability." This view is summed up in Brindley's (1994, p. 75) definition of task-centered or task-based assessment, which is also the task definition adopted for the purpose of the study:

> task-centred language assessment is the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of a goal-directed, meaning-focused language use requiring the integration of skills and knowledge.

3.4. Some Implications of Performance Assessment

The most frequently discussed implications of a performance-based approach to testing, in addition to validity and reliability issues, which will be addressed in the following subsections, are authenticity and the difficulty of generalization and extrapolation. These issues will also be discussed in relation to validity and reliability in the following chapter.

3.4.1. Authenticity in performance assessment

One of the most significant issues in performance and consequently task-based testing that originates from communicative language testing is the issue of authenticity.

Bachman (1990, p. 301) points out that the "search for authenticity continues to be a major consideration in language testing, and tests described variously as "direct", "performance", "functional", "communicative", and "authentic" have been developed and discussed in recent years." Authenticity as a feature of performance and what is now know as task-based assessment is also discussed in Linn and Burton, (1994), Bond (1995), Morrow (1981), Bachman (1990), Bachman and Palmer (1996), Shohamy and Reves (1985) and Chalhoub-Deville (2001).

Chalhoub-Deville (2001, p. 216) argues that authenticity has been discussed for years, initially as a differentiation between original and adapted tasks to arrive at the current views of authenticity and its discussion in relation to the relationship of the task language to the real-world language.

One of the first linguists to propose authenticity as one of the features of communicative and consequently performance tests was Morrow (1981, p. 17), who claims that "measuring the ability of the candidate to, eg read a simplified text tells us nothing about his actual communicative ability." He proposes "a set of authentic language tasks" in order to assess "whether (or how well) the candidate can perform a set of specified activities."

In his response to Morrow's paper, Alderson (1981, p. 48) argues that this demand for authenticity is problematic, claiming that the fact that the setting of assessment itself makes language tests inauthentic.

Shohamy and Reves (1985) quote Clark (1972, as cited in Shohamy and Reves, 1985, p. 49) who distinguishes indirect from direct, that is authentic tests and address the issue of authenticity of what is considered direct tests. Namely, the first oral tests, just before Clark's publication, were carried out in artificial conditions, where candidates would talk to machines, which is why Clark refers to them as indirect. As a direct test, he proposes a test that would resemble real-life conditions or circumstances as closely as possible and resemble real-life language performance. Shohamy and Reves, however, identify two potential issues, the first being the psychometric issue, that is the impossibility to apply the existing statistical analyses to authentic tests due to the wide range of variability of real language. As discussed in Chapter One of Part

Two, different models of communicative and language competence were being proposed at the time (Hymes, Halliday), and they all involved factors other than linguistic ones in actual language performance. These factors, that account for the variability of real language and cause variations in produced language, could not be examined with the existing statistical analyses. As a solution, Shohamy and Reves (1985, p. 53) propose identifying stable elements of authentic language performance in addition to the variable ones that depend on the specific instance of language use.

The second issue that Shohamy and Reves address is the authenticity itself and go on to analyze the differences between authentic test language and authentic real-life language: the goal of the interaction, the participants, the setting, the topic and the time. They argue that the goal of interaction in real-life is not "to obtain a score for their language performance" (Shohamy and Reves, 1985, p. 54) but to communicate a message, which makes the so-called authentic tests considerably less authentic. Talking about participants as a difference, they maintain that the tester and test taker are not likely to be involved in a similar conversation in real life, which makes the interaction "artificial, awkward and difficult" (p. 55). Another problematic difference is the setting, or the physical environment, which in real life, unlike in testing situations, is not likely to happen in a classroom. Similarly, the test topics are different from the real-life topics as in testing situations topics are imposed on the test takers and consequently influence the authenticity of language. Finally, in a testing situation, the imposed time limits may affect the quality of language produced. These five factors or differences are considered "threats" to the authenticity of language in testing situations and evoke the impossibility of designing a completely authentic test. Shohamy and Reves, however, maintain that despite the fact that we cannot replicate real-life language in testing situations, we can obtain authentic "test language," and if we decide to do so, we need to ensure that all psychometric requirements are met.

The approach to test authenticity discussed by Morrow (1981) and criticized by Shohamy and Reves (1985) is also discussed by Bachman (1990). He differentiates between what Morrow considered "real-life" authenticity from the so-called "interactional-ability" approach to defining authenticity. Namely, the idea behind "real-life" authenticity is to reflect the real world, and a test is considered authentic if it replicates non-test language performance, where non-test language performance is seen as the criterion for authenticity and the test language performance

predictive of non-test language performance. (Bachman, 1990, pp. 301 - 302). Essentially, Bachman criticizes the approach to test authenticity also because the trait or language ability measured and the observed performance or behavior are regarded as being the same.

On the other hand, the "interactional-ability" approach does not consider non-test performance as a criterion but aims at distinguishing characteristics or underlying abilities of communicative language user: it sees authenticity as "the interaction between the language user, the context, and the discourse." This approach clearly distinguishes between the observed performance and the underlying abilities to be measured.

Harding (2014, p. 189) recognizes these as "two separate forms of CLT [communicative language testing] that functioned under the same banner": the real-life approach, which is atheoretical but authentic, and the interactional authenticity approach, which considers the underlying traits of communicative ability.

In their *Language Testing in Practice,* Bachman and Palmer (1996, p. 39-42) use different terms for what Bachman (1990) called "real-life" authenticity and "interactional-ability" approach to authenticity, explaining that what they now call "interactiveness" is actually an elaborated view of Bachman's "interactional-ability" approach to authenticity, while the term "authenticity" now is used to refer to "real-life" authenticity and is defined as "the degree of correspondence of the characteristics of a given language test task to the characteristics of a TLU [target language use] task", while "interactiveness" is defined as "the degree to which the constructs we want to assess are critically involved in accomplishing the test task" (Bachman & Palmer, 1996, p. 39). They conclude that they "recognize the value and usefulness of each in characterizing test tasks" (p. 42) and that, since they are both relative, the degree to which a test is authentic needs to be determined in relation to "the characteristics of test takers, the TLU domain, and the test task" (p. 39).

A much simpler definition of authenticity in relation to task-based language assessment is given by Wigglesworth (2008, p. 117): "a central tenet of task-based language assessments is that the tasks are designed to represent authentic activities which test candidates might be expected to encounter in real world outside the classroom." Similarly, Kane et al. (1999, p. 7) believe that, although the term authentic assessment is sometimes used to refer to performance assessment, these two are not synonymous; however, the term authentic does indicate the relevance of

the assessed performance to the real world, in a highly contextualized manner.

Despite the potential problems that this approach may cause, such as the difficulty in generalizing the results of test takers and extrapolating inferences based on the results, this definition is the most generally accepted one in modern task-based testing.

3.3.2. Generalization and extrapolation

Sampling appropriate tasks and generalizability of results is an essential measurement issue in performance-based assessment and literature frequently discussed as a problem originating in the "authenticity" of performance-based assessment (Bachman, 1990, 2001, 2002; Bachman & Palmer, 1996; Brindly, 1994; and Messick, 1994).

Bachman (1991, as cited in Miyata-Boddy, 2000) addresses the issue of generalizability in language testing in general. The solution to the problem of generalization he proposes here is to make sure that a) the language test corresponds to the language abilities in non-test language use, and b) that the characteristics of the test tasks correspond to the features of a target language use context.

Bachman (2002, p. 458) defines generalizability as "the extent to which our inferences generalize across a set of assessment tasks," and extrapolation as "the extent to which our inferences extend beyond the set of assessment tasks to tasks in real-world domain." These two notions are closely related one to another and often addressed together in the field of language testing. He believes (2001, p. 64) that the main difficulty in the generalizability of performance assessment originates in the fact that the assessment method and the ability we wish to measure, language knowledge, are one and the same, that is "language is both the object and instrument of measurement."

In order to be able to make inferences based on a single task performance and generalize them across a set of assessment tasks, there would need to be quite a high number of tasks and contexts. However, according to Weir (1990), the specificity of context makes it nearly impossible to make generalizations. Similarly, Linn and Burton (1994, p. 5) believe that due to the high degree of task specificity, there would need to be a large number of tasks in order to be able to generalize the results. Considering, however, that task completion takes time, the issue of feasibility of a large number of tasks is raised.

Bachman (2002, p. 455) raises the issue of generalizability in relation to task-based performance assessment as well. He quotes Skehan (1998) to propose a solution: testing procedures that would examine the capacity of language learners to deal with a range of realistic conditions. With regard to the "strong" view of performance assessment, he argues that the proponents of the "strong" view or approach to performance-based testing are mainly interested in making predictions about future performance and illustrates the distinction between what he calls "ability-based" inferences about language ability, or what McNamara calls "weak" view, and "task-based" predictions about future performance as "real-world" tasks.

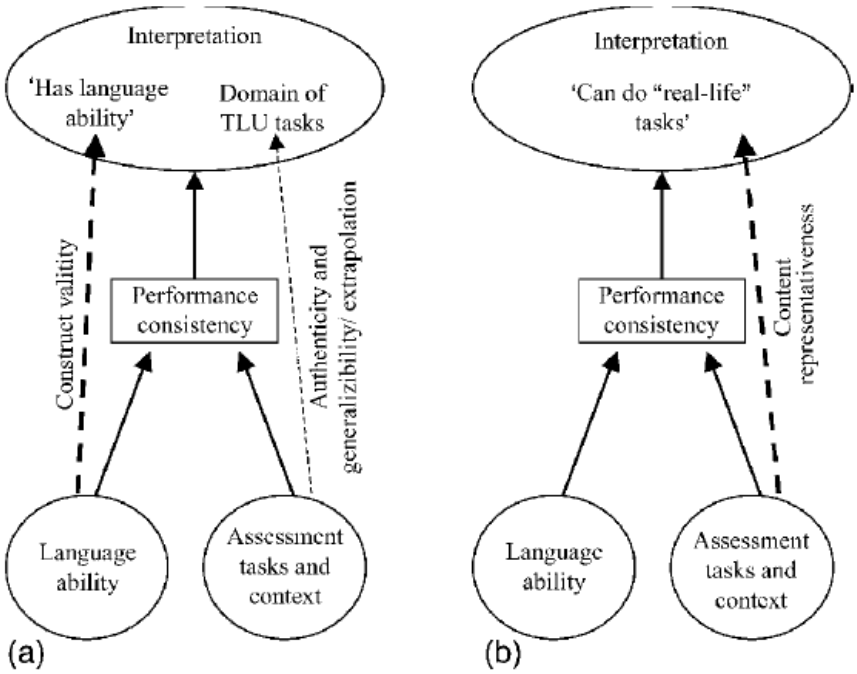

Figure 9. Different interpretations of response consistencies on language assessment tasks: (a) "Ability-based" inferences about language ability and (b) "Task-based" predictions about future performance as "real-world" tasks. Reprinted from Some Reflections on Task-Based Language Performance, (p. 457), by L. F. Bachman, 2002, *Language Testing 19*(4), 453-476

He lists three issues essential for supporting predictions about future performance: 1) defining tasks, or content domain specification; 2) identifying and selecting tasks for use in language assessments, and content relatedness; and 3) the relationship between real-life tasks and assessment tasks (Bachman, 2002, p. 457). With regard to test design, it needs to be both construct-based and task-based as any approach based on only one of the two would be either limited or problematic. Also, a set of task characteristics relative to the particular assessment needs to be designed as a test basis. As a potential basis for the test characteristics, he proposes using an existing framework or the target language use domain. Furthermore, there needs to be a clearly defined construct of what is assessed in the area of language ability, again based on an existing theoretical model of language ability or target language use domain. Finally, information about test performance needs to be collected by means of a number of different procedures in order to try and describe the interactions among specific tasks and specific test-takers (pp. 470 - 471).

Chalhoub-Deville (2001, p. 225) proposes a similar solution: an expansion of test specifications based on the existing linguistic theories, in order to include the language construct underlying knowledge and skills.

Different solutions have been proposed to the problem of generalization or generalizability and extrapolation in performance and task-based assessment. The choice of an appropriate solution, however, depends on the test purpose, that is whether performance is used in its weak or strong sense (McNamara, 1996). If we wish to make inferences about the underlying skills and knowledge (weak sense), we need to define a construct or constructs that we wish to investigate in accordance with linguistic theories (Chalhoub-Deville, 2001, p. 225). If, however, we are interested in performance in its strong sense, that is strictly in performance of the task, we need to consider the characteristics of the tasks and the conditions under which the tasks are administered as well as the extent to which they influence test takers' performance (Wigglesworth, 2008, pp. 113-114) or follow Bachman's procedures to ensure we can make predictions about future performance. Wigglesworth, however, points out that while the "weak" view of task-based assessment

> is likely to assess underlying language skills in ways which are relatively broadly generalizable, the "strong" view is likely to produce judgements which are more

authentic and relevant to the real life situations toward which the candidate may be moving. These judgements may not, however, be replicable in other contexts (Wigglesworth, 2008, p. 118).

The issue of generalizability and extrapolation will be further discussed in the following chapter in relation to approaches to validation.

The approach taken in the study relates to both the "weak" and the "strong" view of performance assessment. The approach will be further discussed in Chapter One of Part Three: Methodology.

3.5. Conclusion: Why Performance-based Assessment?

The main idea behind performance-based and task-based assessment is to "provide information on how well learners are able to mobilize language to achieve meaningful communicative goals" (Brindley, 1994, p. 73). Despite its complexities, the advantages of performance-based assessment are considerable, precisely due to its main goal: to provide information on what learners or test takers are able to do in a second language as well as their underlying knowledge, depending on the approach taken. According to Linn and Burton (1994, p. 5), the main advantages of this type of assessment are its reflection in instructional settings, the fact that they are more engaging for students, and reflect better the criterion performances that are relevant in settings other than the classroom.

Bond (1995, p. 21) similarly believes that this approach to testing is "less stigmatizing, more adaptable to individual student needs, less narrow and more faithful to the richness and complexity of real-world problem-solving, more instructionally relevant, (…) and more reflective of the actual quality of student understanding."

As Messick (1996, p. 1) points out, performance assessment has become popular due to its potential positive consequences for teaching and learning as well as due to the fact that as a result, it has an "extended process or product that can be scored for multiple aspects of quality" (Messick, 1996, p. 2).

In addition, he provides a comprehensive definition of performance-based assessment and its benefits (Messick, 1996, p. 3):

> Prototypical performance assessments occur more toward the unstructured end of the response continuum and include such exemplars as portfolios of student work over time, exhibits or displays of knowledge and skill, open-ended

tasks with no single correct approach or answer, and hands-on experimentation. The openness with respect to response possibilities enables students to exhibit skills that are difficult to tap within the predefined structures of multiple-choice, such as shaping or restructuring a problem, defining and operationalizing variables, manipulating conditions, and developing alternative problem approaches.

Washback or backwash is another potential benefit of performance-based assessment and can be defined as the "impact of tests on the teaching programme that leads up to them" (McNamara, 1996, p. 23). This is due to the fact that teachers are likely to prepare students for the tasks that represent real-life tasks and consequently for non-test language use. Similarly, Wigglesworth maintains that performance-based assessment can have a positive wash-back in the classroom (2008, p. 114). Kane et al. (1999, p. 5) agree that this type of assessment can have "a profound influence on how students study and on how teachers teach."

To sum up, the main advantages of performance-based assessment are its effort to reflect real-life activities and tasks (Wiggins, 1989; Linn, Baker, and Dunbar, 1991), and, in construct-based assessment, that it can provide information about the underlying knowledge and skills (Messick, 1994, 1995, 1996). The outcomes of performance-based assessment, as well as its usefulness for a particular context, will, however, depend on the validation criteria and processes chosen in accordance with the assessment and its particular purpose and use. The validation criteria pertinent to performance-based assessment are discussed in the following chapter.

The previously discussed features of performance-based assessment in conjunction with the discussion of performance-assessment validation are aimed at providing arguments for the employment of performance-based assessment for the purpose of the study.

.

Chapter Four
Validity in Foreign Language Assessment

4.1. The Concept of Validity

The concept of validity has always been a crucial issue in the field of language assessment. Fulcher (2010) distinguishes two traditions of or approaches to language assessment validity: the one before Messick's 1989 publication and the one after. Before Messick's outstanding work concerning the validity of language testing, the approach to validity could be summed up as: "does my test measure what I think it does?" Fulcher (2010, p. 19).

McNamara (1996, p. 15) identifies two issues with regard to the validity of performance-based assessment: "*how* and *how well* we can generalize from the test performance to the criterion behavior" (emphasis in original), where "how" relates to the test design and "how well" to the empirical data obtained from the test administration and their, for example, predictive power. This distinction is similar to Messick's distinction between "evidential" and "consequential" validity criteria (Messick, 1994, p. 13) as well as to Weir's (2005) "a priori" and "a posteriori" validity. Kane (1992, 2011, 2013) as well as Kane, Crooks, and Cohen (1999) talk about two stages in the validation process, the interpretation / use argument and its evaluation, while Bachman and Palmer propose their Assessment Use Argument as a framework for test validation. This chapter addresses the most influential approaches to validation, starting from Messick to the most recent ones, the ones proposed by Bachman and Palmer (2010, 2016), Kane (1992, 2011, 2013) and Weir (2005).

4.2. Messick's integrative approach to performance assessment validity

Messick's most often cited definition of validity "Validity is not a property of the test or assessment as such but rather of the meaning of the test scores" (Messick, 1995, p. 5, 1996, p. 1) can be said to be the underlying principle of his integrative view of validity, where validation implies creating a "network of evidence supporting (or challenging) the intended purpose of the testing" (Messick, 1996, p. 1). He maintains that

performance-based assessment needs to be evaluated by the same validity criteria as any other type of assessment: validity, reliability, comparability, and fairness, as these four are social values relevant to the judgments and decisions to be made on the basis of the test performance.

Messick disagrees with the specialized validity criteria proposed by Linn, Baker, and Dunbar (1991) as he finds them limited despite the fact that they are in line with the general standards of validity. Namely, Linn et al. maintain that although the existing fundamental concepts of validity and reliability, such as efficiency, reliability, and comparability are to be applied to performance-based assessment, they should not be the only criteria used for evaluating any type of assessment. They propose an expansion of the existing or traditional concepts, based on the "theoretical rationale of the modern views of validity" (1991, p. 16). The criteria that they propose, stressing the fact that the presented list is not exhaustive, are: intended and unintended consequences, the degree to which performance on specific assessment task transfers to either the domain or across raters, the fairness of assessment, the cognitive complexity, the meaningfulness of the problems, content quality and comprehensiveness of the content coverage. They conclude, however, that the most significant issue to take into consideration is the "appropriateness and importance of the criteria for the purposes to which assessments are put and the interpretations that are made on the results" (p. 20).

Some of the specialized aspects of validity proposed by Linn, Baker & Dunbar can be mapped to Messick's (1994, 1995, 1996) aspects of construct validity described below. The specialized one, the one of "meaning" that provides a different view from the one proposed by Messick will be discussed afterwards.

Messick (1994, 1995, 1996) proposes an integrative view of validity in performance-assessment, where the notion of validity is seen through six distinct aspects of construct validity or validation. The six aspects that he proposes are:

1) the content aspect,
2) the substantive aspect,
3) the structural aspect,
4) the generalizability aspect,
5) the external aspect, and
6) the consequential aspect.

The content aspect of construct validity involves both content relevance and content representatives and implies defining "the

boundaries of the construct to be assessed" (Messick 1995, p. 6, 1996, p. 7) with regard to both content relevance and content representativeness. This implies a careful analysis of the content domain to be assessed as well as the careful sampling of tasks from the domain, in order to ensure that the domain is well defined and the tasks are representative of the domain which is being assessed. Messick here stresses the significance of assessment specifications, which need to detail the components of the construct domain included in the assessment. This is particularly relevant if not all components of the construct domain are included in the test and, consequently, the score interpretation cannot go beyond what is actually assessed or tested (Messick, 1995, 1996). McNamara (1996, p. 16) sees this aspect as part of the process of test design and agrees that the domain needs to be clearly described and delineated whereas tasks need to be adequately sampled.

The substantive aspect of construct validity is again related to sampling, however, not of tasks but the domain processes involved in the performance of the assessment task. Domain processes need to be not only carefully sampled, but also there needs to be evidence that these processes are actually engaged by task takers in the tasks they perform. This aspect is closely related to the notion of "representativeness," where "representativeness" is seen as whether or not the construct's engagement in the tasks is representative of the domain (Messick, 1995, p. 6, 1996, p. 9).

The structural aspect of construct validity relates to the scoring criteria and rubrics, which need to reflect the task and the domain structure, that is "the internal structure of the assessment should be consistent with what is known about the internal structure of the construct domain" (Messick, 1989, as cited in Messick 1996, p. 10). This is particularly important for comparability of results of assessments that evaluate the same construct domain and use the same scoring criteria (Messick, 1996, p. 10). McNamara's issue of how well we can generalize from test performance (1996, p. 19), encompasses Messick's structural aspect of construct validity as well as the aspect of generalizability. He stresses the importance of carefully designed and clearly defined scoring criteria and rating scales. He identifies his view with the importance that Linn et al. (1991) give to the analysis of the cognitive complexity of the tasks, which needs to be reflected in the validation criteria, that is contained in the rating scales.

The aspect of generalizability is concerned with the interpretation of the scores as not limited to the assessed tasks but generalizable to the

entire construct domain. Whether or not such generalizations can be made depends largely on the correlations between or among the scores on the sampled tasks. Considering that administering and scoring performance-based tests is time-consuming, "a trade-off between the valid description of the specifics of a complex task and the power of construct interpretation" may be necessary (Messick, 1996, p. 11).

The external aspect encompasses convergent and discriminant correlations with external variables. There needs to be convergent evidence that the measure of the construct correlates with the other measures of the same construct and to the other variables "that it should relate to on theoretical grounds" (Messick, 1996, p. 12). In addition, there needs to be discriminant evidence that the measure of the construct is not actually a measure of another distinct construct (Messick, 1996, p. 12).

The final aspect of construct validity, the consequential one, is concerned with the evidence of positive consequences as well as the evidence that there are minimum negative ones. This relates to both intended and unintended consequences of score interpretation of the assessment (Messick, 1995, 1996).

Potential sources of test invalidity that are also a major threat to the concept of validity and therefore to consequential validity as well are construct underrepresentation and construct-irrelevant variance (Messick, 1995, p. 7, 1996, p. 13). Construct underrepresentation occurs when the assessment is too narrow and does not adequately represent the construct, whereas construct-irrelevant variance occurs when the assessment is too broad and contains variance that is not pertinent to the construct (Messick, 1996, p. 5).

The notions of construct underrepresentation and construct-irrelevant variance are closely dependent on the authenticity and directness of assessment, respectively. Performance-assessment is frequently referred to as an authentic and direct type assessment. The main concern regarding authenticity in performance assessment, however, is "that nothing important has been left out of the assessment of the focal construct" whereas the main concern regarding directness is that "nothing irrelevant has been added that distorts or interferes with construct assessment" (Messick, 1996, p. 6). These notions are essential for the criterion validity as "they signal the need for convergent and discriminant evidence that the test is neither unduly narrow because of missing construct variance nor unduly broad because of added method variance" (Messick, 1994, p. 22). For this reason, authenticity and directness are seen as validity standards (Messick, 1994, 1995, 1996).

The aspect of validity, proposed by Linn, Baker, and Dunbar (1991) that according to Messick (1996, p. 13) provides a view different than his is "meaningfulness." Meaningfulness is seen as posing to students or test takers meaningful problems or tasks as to provide worthwhile educational experiences (Messick, 1996, p. 13; Linn et al., 1991, p. 20). This implies not only that the students need to know what exactly of their performance is being assessed but also how it will be scored and what they can do to improve performance (Messick, 1996, p. 13).

What characterizes Messick's approach is his view of the criterion domain as a construct, and as such, it is the criterion domain that needs to be validated (McNamara, 2006).

4.2.1 Performance as a vehicle or target of assessment

Messick (1994, p. 13; 1996, p. 4) however distinguishes two approaches to performance assessment: assessment of performances and products, that is the assessment of performance per se, which he calls "task-driven" assessment, and performance assessment of a construct, which he calls "construct-driven" performance assessment. In the first case, the target of assessment is either performance per se or the product of the performance. The first step, if this approach is adopted, starts by determining the task whose performance we want to assess and then deciding on the constructs to be evaluated. In the second case, however, the performance is merely a vehicle of assessment and the performance or observed behavior is used to make inferences about the actual target, which are constructs such as knowledge and skills underlying the performance (Messick, 1994, p. 14; 1996, p. 4). This approach can be traced back to Lado (1961), who argued for a structuralist approach to testing, as discussed in Chapter Two of Part Two, and which has in time become "the most common approach to general-purpose performance assessment" (McNamara, 1996, p. 26). When adopting this approach, it is the construct that is first identified. The task or tasks to administer and scoring criteria and rating scales or rubrics are determined or designed on the basis of the construct.

These two approaches have different implications for the concept of validity. In task-based performance assessment, "replicability and generalizability are not an issue" (Messick, 1994, p. 14, 1996, p. 4) as assessment of performance is based on a single task which is valued in its own right. This, however, also implies that inferences on the test taker's abilities cannot be made based on the performance on the task. The

opposite is, however, true for construct-based performance testing, where replicability and generalizability cannot be ignored due to the fact that inferences are made on the basis of the performance product and the knowledge and skills considered to be underlying the construct in question. The construct then becomes delineated by the tasks and situations that it can be generalized to (Messick, 1994, pp. 14 - 15), which takes us back to domain representativeness and consequential validity.

Other authors have also supported and discussed Messick's view of this distinction. For example, McNamara (1996, pp. 43 - 44), talks about a general distinction between "weak" and "strong" view of second language performance tests, as discussed in the previous chapter, where the "strong" language performance tests have as its aim or target the performance of the task itself, while "weak" second language performance focus on the language performance on a task in order to make inferences on the language proficiency. Consequently, the criteria used for performance assessment will be different.

Similarly, Chalhoub-Deville (2011) talks about the underlying language abilities in performance assessment, or in Messick's terms, performance as a vehicle of assessment and the difficulties in uncovering these using task-based tests when they are not construct-driven. Skehan (1996, as cited in Bachman, 2002, p. 455) also stresses the importance of inferences to be made about the underlying abilities by means of a task-based approach. Another example is Brindley (1974, p. 75) who, while talking about task-based performance assessment, in his definition of task-based performance assessment refers' to Bachman's (1990) "view of language proficiency as encompassing both knowledge and ability for use." Despite the fact that they talk about task-based performance, they all stress the need for a clearly defined, theory-based construct, which would make it possible to make and justify inferences about the underlying abilities.

An opposing view of the inferences to be made are the one of Brown, Hudson, Norris, and Bonk (2002, as cited in Bachman, 2002, p. 455) are primarily concerned with the inferences about the test-takers' abilities to complete a specific task and not the underlying abilities. They define the construct as what test-takers can do and, in that way, limit the potential inferences to the ones of future performance only (Bachman, 2002, p. 456) as there is no possibility to generalize the performance to other assessment tasks or extrapolate on the tasks in the target language use domain (Bachman, 2002, p. 462).

We can conclude that although different authors use different terms to refer to task-based and performance-based assessment, or performance as a target or vehicle of assessment, they, at the same time, talk about the necessity of justifying the inferences about the underlying abilities, if that is the intended use, which can only be done by means of Messick's construct-driven approach. To this aim, McNamara (1996, p. 17) stresses the necessity of construct validation in any test development process.

The approach taken in the study relates to both the "weak" and "strong" view of performance assessment as the interest of the study is to examine the underlying abilities of the test takers, as well as their performance on the tasks. The approach will be discussed in relation to the study in Chapter One of Part Three: Methodology.

4.3. "Bachman and Palmer, true heirs of Messick"[19]

Messick's validity theory has had a significant influence on the considerations of the notion of validity in language testing, the most obvious example being Lyle F. Bachman and Adrian Palmer.

4.3.1. Bachman (1990)

The influence of Messick on Bachman's work is most evident in his Fundamental Considerations in Language Testing (1990) in the chapter on validity, which was later continued in his works with Palmer (Bachman and Palmer, 1996, 2010). Although Bachman discusses the reliability of language tests in a separate previous chapter, he supports Messick's view of validity as a unitary concept in the chapter on validity (1990, p. 238). He sees reliability as a requisite for validity and the investigation of reliability and validity as "complementary aspects of identifying, estimating and interpreting different sources of variance in test scores" (1990, pp. 160 - 162), where reliability is concerned with the amount of variance that is due to the measurement error and validity with the factors and abilities that contribute to the reliable variance (1990, p. 239). His distinction between reliability and validity is illustrated in the figure below.

[19] McNamara, 2006

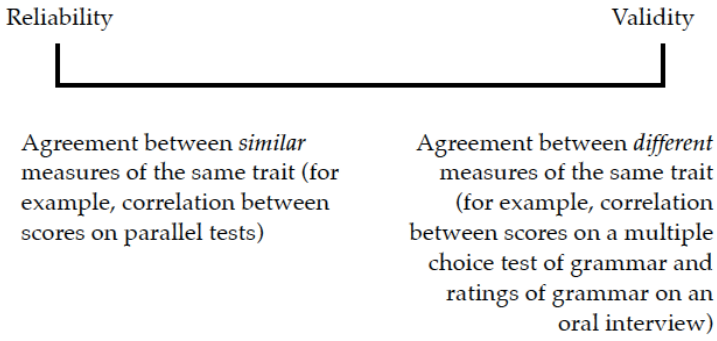

Figure 10. Relationship between reliability and validity. Reprinted from *Fundamental Considerations in Language Testing*, (p. 240), by L.F. Bachman, 1990, Oxford: Oxford University Press.

Finally, he does not find it necessary to draw a clear line between the two but believes that "our concern in the development and use of language tests is to identify and estimate the relative effects of all the factors that influence the test performance" (1990, p. 241). To this aim, he proposes three complementary groups of evidence: content relevance, criterion relatedness, and meaningfulness of construct.

Content relevance concerns the relevance of the test to the specific content or ability and has two aspects: content relevance and content coverage. Similarly to Messick (1995, 1996), Bachman here stresses the importance of the test specifications, that is specifying the ability domain as well as what he calls "test method facets." He provides a framework for describing test method facets (see Figure 11), which in his later works, starting from his Language Testing in Practice (Bachman and Palmer, 1996) he renames to test blueprints (see Figure 12), which include task specifications as well as specifications for the assessment as a whole, procedures for setting cut scores and making decisions, procedures and formats for reporting assessment records, interpretations and decisions, and procedures for administering the assessment.

1 FACETS OF THE TESTING ENVIRONMENT
Familiarity of the place and equipment
Personnel
Time of testing
Physical conditions

2 FACETS OF THE TEST RUBRIC
Test organization
Salience of parts
Sequence of parts
Relative importance of parts
Time allocation
Instructions
Language (native, target)
Channel (aural, visual)
Specification of procedures and tasks
Explicitness of criteria for correctness

2 FACETS OF THE INPUT
Format
Channel (aural, visual)
Mode of presentation (receptive)
Form of presentation (language, nonlanguage, both)
Vehicle of presentation ("live", "canned", both)
Language of presentation (native, target, both)
Identification of problem (specific, general)
Degree of speededness
Nature of language
Length
Prepositional content
Vocabulary (frequency, specialization)
Degree of contextualization (embedded/reduced)
Distribution of new information (compact/diffuse)
Type of information (concrete/abstract, positive/negative, factual/counter-factual)
Topic
Genre
Organizational characteristics
Grammar
Cohesion
Rhetorical organization
Pragmatic characteristics
Illocutionary force
Sociolinguistic characteristics

4 FACETS OF THE EXPECTED RESPONSE
Format
Channel (aural, visual)
Mode (productive)
Type of response (selected, constructed)
Form of response (language, nonlanguage, both)
Language of response (native, target, both)
Nature of language
Length
Prepositional content
Vocabulary (frequency, specialization)
Degree of contextualization (embedded/reduced)
Distribution of new information (compact/diffuse)
Type of information (concrete/abstract, positive/negative, factual/counter-factual)
Topic
Genre
Organizational characteristics
Grammar
Cohesion
Rhetorical organization
Pragmatic characteristics
Illocutionary force
Sociolinguistic characteristics
Restrictions on response
Channel
Format
Organizational characteristics
Propositional and illocutionary characteristics
Time or length of response

5 RELATIONSHIP BETWEEN INPUT AND RESPONSE
Reciprocal
Nonreciprocal
Adaptive

Figure 11. Categories of test method facets. Reprinted from *Fundamental Considerations in Language Testing*, (p. 119), by L.F. Bachman, 1990, Oxford: Oxford University Press.

Blueprint

1 Assessment specifications

(a) Number of parts
(b) Number of tasks per part
(c) Sequence of parts and tasks
(d) Relative importance of parts and tasks
(e) Time allotment: for each part and overall
(f) Instructions

2 Task specifications (for each task type)

(a) Construct to be assessed
(b) Characteristics of the setting
(c) Characteristics of the input, expected response, and relationship between input and response
(d) Recording method
(e) Instructions for responding to the assessment task

3 Procedures for setting cut scores and making decisions
4 Procedures and formats for reporting assessment records, interpretations, and decisions.
5 Procedures for administering the assessment

Figure 12. Components of a blueprint. Reprinted from *Language Assessment in Practice*, (p. 370), by L.F. Bachman and A. S. Palmer, 2010, Oxford: Oxford University Press.

The second aspect of content validity, content coverage, is concerned with "the extent to which the tasks required in the test adequately represent the behavioral domain in question" (1990, p. 245), which Messick (1995, 1996) calls content representativeness. Bachman maintains that demonstrating either content relevance or content coverage is not simple since they are often difficult to specify with precision (Bachman, 2002, p. 460). Furthermore, even if they are specified, they will only support the interpretations that are limited to the specified domain (1990, pp. 245 - 246, 2002, p. 460) and not inferences about abilities (1990, p. 247).

The second type of evidence to support validity that Bachman provides is criterion relatedness or criterion validity, defined as a "relationship between test scores and some criterion which we believe is also an indicator of the ability tested" (1990, p. 248). He considers two aspects of criterion validity: concurrent criterion relatedness or concurrent validity and predictive utility or predictive validity.

Concurrent criterion relatedness can be supported by two types of information: the one coming from the differences in test performance between different groups of individuals with different levels of language ability, and from correlations between or among different measures, such as different types of tests, which is also more common. This implies correlations with, for example, a standardized test, which on its own cannot be sufficient to demonstrate validity as it would imply that the test itself has already been accepted as an indicator of the ability and high correlation with another test may mean a simple transfer of the assumption of validity of the test. The most serious limitation according to Bachman, however, is the fact that it only considers the extent to which the measures of the ability agree and not the extent to which the measures of the ability disagree (Bachman 1990, p. 248 - 250).

Predictive utility or predictive validity, on the other hand, concerns only the tests administered with the purpose of making predictions about the candidate future performance. This is, however, quite problematic as such test would need to cover all the aspects of the criterion ability in question. Considering that we are most often interested in candidate abilities and not making any predictions about their performance, Bachman proposes basing the test on a definition of ability and demonstrating construct validity instead (1990, pp. 250 - 254).

Finally, Bachman defines construct validity as "the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs" (1990, p. 255). He stresses the importance of defining what it is that we want to measure, which then becomes a construct that needs to be further defined in relation to a theory that would relate it to other constructs and observable performance since abilities or constructs themselves are not directly observable.

What Bachman sees as a prerequisite for validity, reliability, is discussed separately in Bachman's *Fundamental Considerations in Language Testing* (1990). This is due to the fact that by identifying potential sources of errors in measurement, we increase the reliability of measures and meet the necessary requirement for validity – reliability of test scores (1990, p. 160). In our investigation of reliability, the question to ask is "How much of an individual test performance is due to measurement error, or to factors other than the language ability we want to measure?" (p. 161) with the aim of minimizing the effects of measurement error and maximizing the effects of the language abilities, we want to measure. In order to be able to do this, we need to identify the potential sources of error, that is the factors that may threaten the reliability of our measures.

Bachman (p. 164) lists three different groups of potential sources of measurement error:

1. test-method facets that is, test characteristics,
2. attributes of test takers which are not related to the language ability in question,
3. random, unsystematic factors that are unpredictable and temporary.

The first two groups are both systematic, in that they remain the same throughout different test administrations (test-method facets) and they would affect a test-taker's performance consistently (personal attributes). On the other hand, the third group comprises factors such as a temporary emotional state, changes in the test environment, etc. These three groups of factors may influence test scores, which Bachman also refers to as "observable attributes" and consequently the unobservable ones to give us an inaccurate interpretation of test-takers' language ability.

For the purpose of achieving reliability of scores, he proposes a number of statistical analyses, among which are Coefficient Alpha for the internal consistency of a test and correlation between raters' marks to evaluate the extent to which ratings by different raters are consistent (inter-rater agreement). Ultimately, the approach to demonstrating reliability of test scores that we will choose will depend on what we consider the potential sources of error in measurement (p. 184).

4.3.2. Bachman and Palmers' Assessment User Argument

Bachman's view of validity, which is consistent with Messick's approach, was transformed in the Assessment Use Argument (AUA) in *Language Assessment in Practice* (Bachman & Palmer, 2010), which they define as "a conceptual framework for guiding the development and use of a particular language assessment, including the interpretations and uses we want to make on the basis of the assessment" (2010, p. 99). The key words in their AUA are claims, data, warrants, backing, and rebuttal backing, while the focus is largely on justifying the use of a particular assessment.

Claims are defined as "statements about the inferences to be made on the basis of data and the qualities of those inferences" (p. 99) and consist of an outcome of the assessment and qualities of the outcome. Data comprise the information used to make a claim. Additionally, warrants are the statements used in the claim in order to detail one of more

qualities of the claim, with the aim of providing justification for the qualities of the intended consequences, decisions, interpretations, and assessment records (p. 101). Rebuttals, on the other hand, are the statements provided in order to challenge the claims. One possible claim, for example, can be that the test-takers' scores are consistent. Accordingly, the internal consistency of the scores and consistency between raters would, in this case, be warrants, while rebuttals would be that the scores are not internally consistent and the ratings between two raters are not consistent either (pp. 101 – 102). Finally, backings are the evidence we provide to support the warrants. The evidence can be gathered before the test development and be in the form of documents, regulations, theory, research or experience, etc. Additionally, it can be gathered during the test design, for example, Design Statement, that is, a document that includes all the necessary information for the test design, Blueprint, that is, the specification of the test and the tasks it includes, and the test. Finally, evidence needs to be provided after the test administration as well, in the form of empirical evidence.

Bachman and Palmer further provide four general types of claims, that need to be made specific for each type of assessment (see Figure 13).

Claim 1 The *consequences* of using an assessment and of the decisions that are made are **beneficial** to stakeholders.

Claim 2 The *decisions* that are made on the basis of the interpretations:

- take into consideration community **values** and relevant legal requirements and
- are **equitable** for those stakeholders who are affected by the decision.

Claim 3 The *interpretations* about the ability to be assessed are:

- **meaningful** with respect to a particular learning syllabus, an analysis of the abilities needed to perform tasks in the TLU domain, a general theory of language ability, or any combination of these,
- **impartial** to all groups of test takers,
- **generalizable** to the TLU domain in which the decision is to be made,
- **relevant** to the decision to be made, and
- **sufficient** for the decision to be made.

Claim 4 The *assessment records* (scores, descriptions) are **consistent** across different assessment tasks, different aspects of the assessment procedure (e.g. forms, occasions, raters), and across different groups of test takers.

Figure 13. Four types of claims in an AUA. Reprinted from *Language Assessment in Practice*, (p. 103), by L.F. Bachman and A. S. Palmer, 2010, Oxford: Oxford University Press.

The Assessment Use Argument provides a general framework that can be used as a guide in assessment developments with a specific focus on the intended use of the assessment. In his Some Construct Validity Issues in Interpreting Scores from Performance Assessments of Language Ability (2001) Bachman provides an account of construct validation for performance-based assessment, which will be discussed in the following section.

As McNamara (2006) points out, Bachman (1990) as well as Bachman and Palmer (1996) employ Messick's construct-driven approach and apply it to the field of language testing not only through their approach to validity but even more through their model of language proficiency and their approach to the test method. McNamara (2006, p. 35) exemplifies his point using three aspects of Bachman's approach. Firstly, the criterion domain is seen as a construct by means of the framework for analysis of the target language use situation. Secondly, the test construct, that is the model of communicative language ability is clearly related to the criterion construct – the criterion domain and the test construct are modeled in the same way. Thirdly, the test method is seen as part of the test content. These three, in McNamara's opinion, establish a clear relationship between the target language use situation, test task and test construct, that is, the communicative language ability, which reflects Messick's approach to test validation as makes it possible to validate hypothesized relationships through by means of test performance data at the same time investigating the possibilities of construct under-representation and construct-irrelevant variance.

Bachman's starting point, as well as the conclusion, are the same as Messick's: validity is to be seen as a unified concept and "none of these [content relevance, predictive utility, and concurrent criterion relatedness] by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores" (1990, p. 237).

4.3.3. Bachman on the validity of performance-based assessment

Although he stresses the importance of reliability as well, Bachman (2001) focuses on the qualities of construct validity, authenticity and interactiveness, which in his opinion provide "a guiding framework for both the design and development of language tests and for the interpretation and use of results from these tests" (p. 87). He defines construct validity "as the extent to which the results of an assessment can be interpreted as an indicator of the ability we intend to measure, with

respect to a specific domain of generalization" (Bachman & Palmer, 1996, p. 21; Bachman, 2001, p. 65). He finds that three elements are necessary for assessment design, development and use, the first one being a definition of language ability that sees the relationship between the components of language ability and other cognitive processes as well as the relationship between these two and characteristics of a language use situation as interactionalist. The second necessary element is a clearly identified domain of a target language use task, while the third is a framework that describes the features of both the assessment task and target language use task (Bachman, 2001, p. 66).

As an example of the first element, an interactionalist definition of the abilities to be assessed, he uses the framework proposed by Bachman and Palmer (1996), where two types of interaction are identified: among attributes of individuals such as language knowledge, topical knowledge, affective schemata, and strategic competence, and, the second one, the interaction between these attributes and the characteristics of the language use situation (Bachman, 2001, p. 67). The framework is consistent with the interactionalist nature of the definition of language knowledge: "a domain of information that is available for use by the metacognitive strategies in creating and interpreting discourse" (Bachman and Palmer 1996, p. 66).

His approach to the second element, defining the target use situation and tasks, is consistent with his previous works on the topic, and he maintains his distinction between "authenticity" and "interactiveness" (as discussed in Chapter Three of Part Two), where authenticity is seen in relation to the correspondence of a language test task to a target language use task, that is, the degree of the similarity between the two (Bachman and Palmer, 1996, p. 39). Consequently, language test tasks become instances of target language use tasks that constitute a target language use domain and to which inferences can be generalized.

The third element, a framework of task characteristics, is necessary in order to demonstrate the correspondence of test tasks to target language use tasks. To this aim, he proposes the Bachman and Palmer (1996) framework of task characteristics (previously called "test method facets"), comprising the setting, the rubric, the input, the expected response and the relationship between input and response. By providing these details about a task, the target language use domain is narrowed and the issues of sampling, content coverage and representativeness can be solved. In addition, it facilitates the comparability of test tasks and

consequently makes the generalization of inferences to the target language use domain feasible.

Furthermore, Bachman (2001, p. 72) proposes using the same framework as a means of addressing the issues of authenticity and interactiveness and illustrates it through Figure 14.

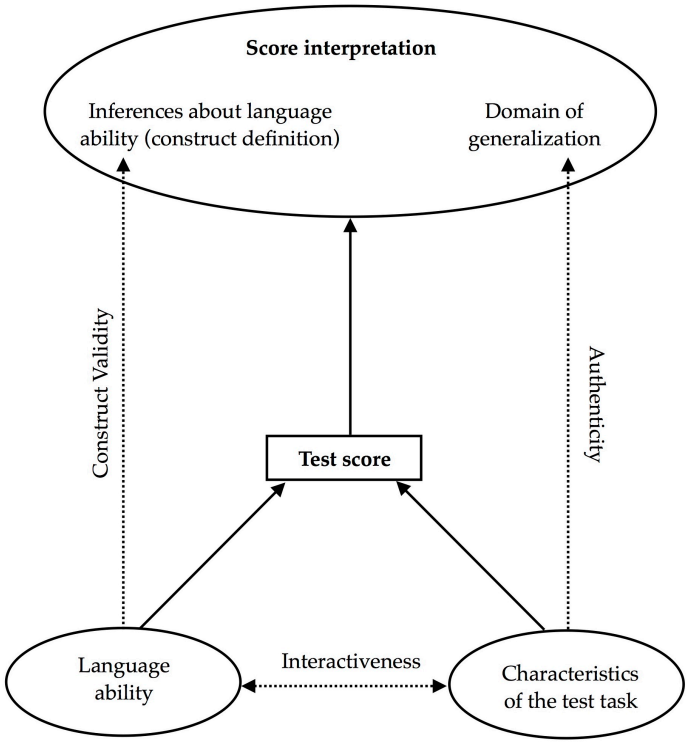

Figure 14. Construct validity of score interpretations, authenticity and interactiveness. Reprinted from *Language Testing in Practice*, (p. 22), by L.F. Bachman and A. S. Palmer, 1996, Oxford: Oxford University Press.

Bachman's approach to construct validity issues in performance-based assessment detailed here are consistent with his previous works (Bachman, 1990; Bachman & Palmer, 1996) where he first talked about the necessity of test method facets, what he later renamed to test characteristics, to finally include in his AUA (Bachman & Palmer, 2010) as task specifications.

4.4. Mislevy's Evidence-Centred Design (ECD)

Another framework for validation in second language testing has been proposed by Mislevy, Steinberg and Almond (2002), which, in their opinion, integrates all the elements of assessment, from assessment purpose to inferences about students and represents "an explicit normative model" (2002, p. 479). The evidence-centered design framework consists of four parts: student models, evidence models, task models and an assembly model.

The student model comprises variables that characterize the students or test-takers, which are unobservable, probability-based and used to gather evidence about students. Four ways for establishing relationships between the student model variables and claims, that is, statements or inferences that we would like to make about student knowledge, are proposed: one-to-one relationship, multiple claims and a single variable with a finite number of levels, multiple claims with a single variable by means of model response probabilities and "the interaction among competences and contexts, as multiple SM [student model] variables are called upon to express evidence for a claim" (Mislevy, et al., 2002, pp. 482 - 484).

The evidence model is to provide evidence about the students' knowledge and skills based on the observable behavior or performance. It consists of the evaluation component and the measurement component, where the evaluation component is concerned with identifying the major notable features of student performance, while the measurement component focuses on the dependency of observable variables, that is features, on student model variables.

The task model is supposed to provide a basis for designing tasks that will elicit the needed evidence through student behavior. In addition, it includes the environment and expected product specifications.

Finally, the assembly model is supposed to provide information on how individual tasks are put together to form assessment and "manages the interplay among student, task and evidence models" (Mislevy et al., 2002, p. 492).

The ECD model is presented as a way of transforming data into evidence about student knowledge. It provides a detailed description and analysis of the stages of assessment validation and "a clear if somewhat abstract blueprint for the design of tests and of associated validation studies" (McNamara, 2006, p. 46). McNamara also stresses the abstract nature of the framework as well as the fact that it relies heavily on

psychometric measures and does not discuss the issue of test use consequences. The framework, however, remains one of the two distinctive examples of the influence of Messick's work in relation to validity, the second one being the one of Kane.

4.5. Kane's Interpretation / Use Argument

Kane's (Kane, 1992, 2011, 2013) interpretative argument for test validation is another validation framework in line with Messick's (McNamara, 2006, p. 47).

While Bachman and Palmer (2010, 2016) use the term AUA, that is, Assessment Use Argument, where the focus is on the use of scores, Kane in his first works uses (Kane, 1992; Kane, Crooks & Cohen, 1999; Kane, 2011) the term interpretative argument to focus on the interpretation of the scores. In his Validating the Interpretations and Uses of Test Scores (2013), he, however, uses the term IUA, that is, interpretation/use argument "in order to recognize the importance of score use in determining score interpretations and to acknowledge the importance of score uses (as well as contexts and test-taker populations) in validation" (Kane, 2013, p. 65).

The primary aim of Kane's IUA is to provide what in his opinion Messick failed to do: "clear guidance for the validation of score interpretations or uses" (Kane, 2011, p. 7). He believes that, despite the fact that Messick's view of construct validity as a unified concept, despite the fact that it was "appealing" and "elegant," was not practical enough nor easy to implement. Kane maintains that an argument-based approach to validation provides a conceptually and operationally clear framework for validation through the intended interpretation and use of the scores (Kane, 2011, p. 8). The distinctive feature of both Bachman and Palmer, and Kane's validity arguments or argument-based approaches is their practicality and ease of use. Kane himself stresses that there is no particular pattern for an IUA to follow but that it will depend on the proposed interpretations and use of the scores (Kane, 2013, p. 10).

Kane maintains that "validity is not a property of the test" and that it is "a matter of degree" (2013, p. 3). The starting point, as well as the underlying principle in Kane's validity argument, are the proposed interpretations and use of scores, which will depend on the purpose of the assessment (Kane, 2013, p. 14) and which need to be supported by appropriate evidence (Kane, 1992; Kane, Crooks and Cohen, 1999; Kane 1999, 2011, 2013). He (1992, p. 527) proposes the following stages in

assessment validation: a) deciding on statements and decisions to be based on the scores, b) specifying the inferences and assumptions leading from the scores to the statements and decisions, c) identifying potential competing interpretations, and d) providing evidence supporting the inferences and assumptions. The process, in his later works (2011, 2013) comprises two steps: an "interpretative argument" or IUA, where proposed score interpretations and uses are detailed, starting from the observed performances, justifications for their use, to arrive at the conclusions and decisions, and the "validity argument", which comprises an evaluation of the interpretative argument or IUA as well as the claims made in the IUA (Kane, 2011, p. 8; Kane 2013, p. 14).

The interpretative argument needs to be assessed against three criteria (Kane 1992, p. 528; Kane 2011, p. 13): clarity of the argument, cohesion of the argument and plausibility of inferences and assumptions. The first criterion, clarity, implies that the inferences need to be stated in detail in order to make the claims obvious. There needs to be a chain of inferences from the observed performance to the conclusions and decision as well as supporting evidence for the inferences. With regard to coherence, it implies that steps leading from the observed performance to the decisions need to be coherent and persuasive and the conclusions reasonable. Finally, the assumptions in the interpretative argument need to be plausible and supported by evidence. In case a single type of evidence cannot make an assumption plausible, there need to be more types of evidence.

In their 1999 paper, Kane, Crooks, and Cohen discuss the validation of performance-based assessment in particular. When talking about disadvantages of performance-based assessment, they quote Messick (1994) to agree that the appearance of validity does not make proposed interpretations valid. As reasons, they list the fact that performance-based tests are marked by human raters, the problem of task specificity and the problem of generalizability (1999, p. 5). These three issues can be resolved by addressing the three types of inferences: scoring, generalizability and extrapolation, and two central assumptions in performance-based assessment: that the interpretation of scores is related to the skills in a performance domain, and that the observations on the basis of which inferences are made are of performance on a task or tasks from the domain of interest (1999, pp. 7 - 9). Kane et al. use the expression "target domain" to refer to the wider or full domain of performances included in the interpretation, while "target score" is used to denote the test-taker's expected score in the target domain. In that way, the score interpretation

extends from the observed score to the type of performance included in the assessment. For this reason, the target domain needs to be broadly defined and not limiting in order to reflect the domain of interest. The tasks included in the assessment will then be representative or random samples of the target domain, even though the broad target domain makes it impossible to cover all the possible samples of tasks, mostly due to practical reasons. For this reason, Kane et al. (1999, p. 8) conclude that "it is generally not plausible to assume that the set of performances included in the assessment is a random or a representative sample from the target domain." To this aim, they use the term "universe of generalization", which has its origin in the generalizability theory, to refer to the subdomain for which "it is plausible to consider the observed performances to be a random or representative sample" (1999, p. 8) and the term "universe score" to denote the test-taker's score in the universe of generalization. In order to make inferences from the observable behavior on a sample task from the universe of generalization and consequently to the target domain, there is a chain of inferences consisting of three types of inferences to follow: scoring, generalization, and extrapolation.

Scoring denotes making inferences from a performance or performances to observed scores by means of scoring criteria that need to be appropriate and consistently applied to the performance that occurs under conditions which are consistent with the proposed interpretations in the sense that "there are no inappropriate impediments" (1999, p. 9). The scoring criteria employed need to clearly distinguish between good and bad performance and provide detailed scoring rubrics and procedures.

The next link in the chain of inferences is generalization from the observed score to the universe of generalization. In order to be able to generalize from the observed score, the sample task or tasks need to be random or representative samples from the universe of generalization. Generalization is then done by means of statistical analyses by evaluating the reliability of scores across samples of observation. A way of improving generalizability can be through standardization of task characteristics and task administration procedures (Kane 1992, p. 529; Kane et al., 1999, p. 10).

The third link or type of inferences is extrapolation, that is transferring inferences from universe scores to target scores, that is, from a quite narrow universe of generalization to a much broader and less defined target domain (Kane et al., 1999, p. 10). The extent to which it is safe to

extrapolate will depend on the proportion of the target domain included in the universe of generalization. Extrapolation can be justified by means of criterion-related validity, where assessment scores are correlated with scores on some criterion, or by arguing that the skills needed for good performance in the universe of generalization are the same as the ones needed for good performance in the target domain (1999, p. 11).

Finally, generalizability can be increased at the expense of extrapolation and vice versa. Namely, if the tasks administered to the test-takers are the ones whose goal is to replicate real-life or non-test situation tasks, they tend to be time-consuming and for that reason, the number of tasks will normally be low, and consequently, the generalizability to the universe of generalization will be low as well. This will, however, strengthen the extrapolation due to the fact that the assessment tasks replicate real-life conditions and situations. If, on the other hand, we decide to administer a larger number of tasks, which reflect real-life situations to a lesser extent, to be able to generalize to the universe of generalization, the extrapolation to the target domain will be much weaker because the tasks do not reflect real-life conditions and situations. This implies that a trade-off needs to be made, depending on whether our focus is on generalizability or extrapolation (Kane et al., 1999, p. 11).

As we can see, the argument-based approach to validation comprises two stages: the interpretation / use argument, where the proposed interpretations and uses of the test scores are detailed, and the evaluation of the plausibility of the proposed interpretation / use argument (Kane, 2011, p. 3). Consequently, the validation framework will depend on the proposed interpretations and uses and the kind of inferences we want to make. As we have seen, the IUA is not prescriptive in any sense but focuses on practicability, while the principal aim is to provide guidance through the validation process of an assessment.

4.6. Weir's Evidence-based Approach to Validity

Weir's works on validity in language assessment are in line with the previously discussed approaches in that it focuses on the test scores as a reliable measure of a trait or construct (2005, p. 12). Accordingly, he agrees that different types of evidence are needed to demonstrate the validity of test scores and sees these different types of evidence as complementary and not as alternatives (p. 13), which is in accordance with Bachman's (1990) approach to validity as well as with Messick's unified view of validity (1994, 1995, 1996). Consequently, he sees

reliability as one type of evidence supporting validity and uses the term "scoring validity" to refer to what is generally known as "reliability" (p. 14). Finally, construct validity is seen as an interaction between theory-based validity and context-validity and not as a superordinate concept encompassing different types of validity. Weir also discusses the criterion-related validity and consequential validity.

Evidence that supports theory-based validity can be collected before or after test administration, and as such can be divided into a priori and a posteriori evidence (Weir, 2005, p. 17), where the a priori hypothesized language theory underlying the test and the defined construct is empirically validated a posteriori through appropriate statistical analyses.

The definition of what Weir sees as the second component of construct validity, context or content validity, is consistent with the one by Bachman: "the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample" (p. 19). In achieving content or context validity, Weir particularly stresses the necessity of ensuring that the abilities or skills for which claims are made in the specifications are actually tested and that the test conforms to the specification.

The third type of validity that Weir proposes is "scoring validity", the term he uses to encompass different types of what is more traditionally known as reliability, and it is concerned with "the extent to which test results are *stable over time*, *consistent in terms of content sampling* and *free from bias*" (Weir, 2005, p. 23, emphasis in original). Similarly to Bachman, Weir claims that scoring validity, or what Bachman refers to as reliability, is necessary but not sufficient evidence of a test's validity (p. 24). He distinguishes four different types of scoring validity: test-retest reliability, parallel forms reliability, internal consistency and marker reliability. The latter two types of scoring validity will be discussed in more detail in relation to the scoring system of the test administered for the purpose of the study.

Weir also addresses the issue of criterion-related validity, which, together with scoring validity belongs to the group of a posteriori type of evidence. Criterion-related validity, or what Bachman refers to as criterion relatedness, comprises concurrent and predictive validity. Weir employs Bachman's definition of concurrent validity: "information … which demonstrates a relationship between test scores and some criterion which we believe is also an indicator of the ability tested" (1990, p. 248). As examples of evidence of concurrent validity, Weir proposes other

measures of performance, teachers' ranking of students or student self-assessment. Predictive validity, on the other hand, is of interest only in case we intend to make predictions about test-takers' future performance. This would, however, imply a different type of test than the one normally designed to collect information on test-takers' knowledge (p. 36).

Finally, a further type of validity, introduced by Messick, consequential validity, encompasses the potential and actual intended and unintended consequences of test interpretation and use (p 37.)

4.7. Conclusion

It is obvious from the contemporary approaches to validity that Messick's legacy is considerable. The unified view of validity, as well as their complementary nature, is present in all the discussed approaches. It is the proposed interpretations and uses of the test scores that will determine the process of validity and evidence collection in order to make the proposal plausible. The types of validity most frequently discussed in second language assessment, construct validity, context or content validity, scoring validity and consequential validity, have been employed to justify the proposed interpretations and uses of the test scores of the test administered for the purpose of the study.

Part Three

Chapter One
Methodology

1.1. Task-based performance assessment in the study

The research methodology has been chosen in accordance with the main research questions:

- How do Italian students, after they have finished high school, perform on written and spoken extended production tasks that reflects everyday real-life activities and situations?
- Are their speaking and writing skills at the CEFR B2 level of English language knowledge (as per the Ministry of Education Guidelines)?
- What is their level of acquisition in different areas of language knowledge such as organizational and pragmatic knowledge and their individual components?

As discussed in Chapter Three of Part Two, there have been different definitions of task-based performance assessment, the main feature of each of them being the inferences we want to make about the test-takers.

In view of the research questions and the information on the test-takers that we are interested in, the definition of task-based performance assessment employed for the purpose of the study is the one of Bachman (1990, p. 77), where performance test is defined as one where "the test takers' performance is expected to replicate their language performance in non-test situations". This definition is also consistent with the one proposed by McNamara (1996, p. 6), who defines performance-based assessment as "an actual performance of relevant tasks (…), rather than a more abstract demonstration of knowledge."

The definition of a task employed in the study is the one by Bachman and Palmer (1996, p. 44): "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation."

With regard to the term "real-life," it is primarily employed as defined by Wigglesworth (2008, p. 117): "a central tenet of task-based language assessments is that the tasks are designed to represent authentic activities which test candidates might be expected to encounter in real world outside the classroom." In addition to the real-life view of authenticity,

Bachman also recognizes "interactiveness" as an approach to defining authenticity, defined as "the degree to which the constructs we want to assess are critically involved in accomplishing the test task" (Bachman & Palmer, 1996, p. 39). This view encompasses the engagement of the areas of language knowledge, metacognitive strategies, topical knowledge and affective schemata, that is the feelings associated with specific kinds of topical knowledge (Bachman & Palmer, p. 42) in performing a task. The aim of the research is the English language competence, as defined by Bachman and Palmer (1996, 2010) and by means of their model of language competence. For this reason, the research does focus on the underlying skills, however only in terms of language competence, not metacognitive strategies, topical knowledge and affective schemata. For this reason, as suggested by Bachman and Palmer (2010, p. 217), the option chosen for defining the construct to be measured, does not encompass topical knowledge, but language competence only.

1.2. Research Constructs

Two distinct constructs are investigated in the research: writing and speaking skills in the real-life public domain by means of Bachman and Palmer's (1996, 2010) framework of English language knowledge, and performance on real-life tasks. These two constructs are in line with the "construct-driven" and "task-driven" performance assessment (Messick, 1994, p. 14) or "weak" and "strong" language performance tests (McNamara, 1996, p. 43) and "task-centered" and "construct-centered" approach (Bachman, 2002, p. 454). The essential difference between these two approaches is in the inferences we want to make about the test-takers' knowledge: the first one is concerned with the underlying language ability, while the second one relates to how well test-takers perform a given task. A discussion of the two different approaches to performance-based assessment is provided in Chapter Three of Part Two, while a detailed description of Bachman and Palmer's model is provided in Chapter One of Part Two. This distinction is also reflected in the assessment criteria: analytic scales are utilized to investigate the underlying language knowledge, in particular, the language knowledge components of Bachman and Palmer's framework. Additionally, holistic rating scales are utilized to gather information on the task achievement. The rating scales will be discussed in more detail in the Assessment criteria subsection of this chapter.

1.3. CEFR Alignment

The Common European Framework of Reference for Languages: Learning, Teaching and Assessment, "the common currency in language education" (Alderson, 2007, p. 660), is discussed in detail in Chapter Two of Part One. As outlined there, major awarding bodies, as well as publishers all, base their exams and course books on the CEFR levels. Similarly, different institutions, such as Ministries of Education, define the required level of English for the purposes of their decrees and public calls in terms of CEFR levels.

The Italian Ministry of Education in their National Guidelines (2010) define the aims and objectives of the fifth-year language curriculum for upper secondary schools in the following way:

> The student acquires linguistic-communicative competences equivalent to the CEFR level B2. The student can produce oral and written texts (in order to report, describe and argue) and reflects on the formal characteristics of texts he/she produces in order to demonstrate an acceptable level of fluency. In particular, the fifth year of the lyceum serves to consolidate the methods of study of the foreign language by learning non-language content, in accordance with the cultural characteristics of each lyceum and the development of personal and professional interests. (National guidelines, 2010)

In original:

> Lo studente acquisisce competenze linguistico-comunicative corrispondenti almeno al livello B2 del Quadro Comune Europeo di Riferimento per le lingue. Produce testi orali e scritti (per riferire, descrivere, argomentare) e riflette sulle caratteristiche formali dei testi prodotti al fine di pervenire ad un accettabile livello di padronanza linguistica. In particolare, il quinto anno del percorso liceale serve a consolidare il metodo di studio della lingua stranieri per l'apprendimento di contenuti non linguistiche, coerentemente con l'asse culturale caratterizzante ciascun liceo e in funzione dello sviluppo di interessi personali o professionali. (Indicazioni nazionali, 2010)

The global scale of the Common European Framework for Languages (Council of Europe, 2000, p. 24) defines level B2 in the following way:

> Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can

interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

Furthermore, the illustrative scales of the CEFR describe the abilities as well as communicative language competences of interest to the research in the following manner.

WRITTEN PRODUCTION	
Overall Written Production	Can write clear detailed texts on a variety of subjects related to his field of interest, synthesizing and evaluating information and arguments from a number of sources.
Creative Writing	Can write clear, detailed descriptions of real and imaginary events and experiences marking the relationships between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest.
Reports and Essays	Can write an essay or report that develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. … Can write an essay or report which develops an argument, giving reasons in support or against a particular point of view and explaining the advantages and disadvantages of various options. …
WRITTEN INTERACTION	
Overall Written Interaction	Can express news and views effectively in writing, and relate to those of others.
Correspondence	Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views.
SPOKEN INTERACTION	

Overall Spoken Interaction	Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational and leisure topics, marking clearly the relationships between ideas. Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the circumstances. Can interact with a degree of fluency and spontaneity that makes regular interaction, and sustained relationships with native speakers quite possible without imposing strain on either party. Can highlight the personal significance of events and experiences, account for and sustain views clearly by providing relevant explanations and arguments.
Understanding a Native Speaker Interlocutor	Can understand in detail what is said to him/her in the standard spoken language even in a noisy environment.
Conversation	Can engage in extended conversation on most general topics in a clearly participatory fashion, even in a noisy environment. Can sustain relationships with native speakers without unintentionally amusing or irritating them ore requiring them to behave other than they would with a native speaker. Can convey degrees of emotion and highlight the personal significance of events and experiences.

Table 1. CEFR Illustrative scales for Written Production, Written Interaction and Spoken Interaction. Adapted from *Common European Framework of Reference for Languages: Learning, teaching, assessment* by Council of Europe, 2001

COMMUNICATIVE LANGUAGE COMPETENCE	
Linguistic	
General Linguistic Range	Can express himself/herself clearly and without much sign of having to restrict what he/she wants to say. Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous looking for words, using some complex sentence forms to do so. Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.
Vocabulary Range	CEFR B2 Has a good range of vocabulary for matters connected to his field and most general topics. Can vary formulation to avoid repetition, but lexical gaps can still cause hesitation and circumlocution.
Grammatical Accuracy	Good grammatical control. Occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect. Shows a relatively high degree of grammatical control. Does not make mistakes that lead to misunderstanding.
Vocabulary Control	CEFR B2 Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.
Phonological Control	Has a clear, natural pronunciation and intonation.
Orthographic Control	Can produce clearly intelligible continuous writing, which follows standard layout and paragraphing conventions. Spelling and punctuation are reasonably accurate but may show signs of mother tongue influence.
Sociolinguistic	
Sociolinguistic Appropriateness	Can express him- or herself confidently, clearly and politely in a formal or informal register, appropriate to the situation and person(s) concerned. …

	Can express him or herself appropriately in situations and avoid crass errors of formulation.
Pragmatic	
Coherence	Can use a variety of linking words efficiently to mark clearly the relationship between ideas. Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution.
Spoken Fluency	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can provide stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party.

Table 2. CEFR Illustrative scale for Communicative Language Competence. Adapted from *Common European Framework of Reference for Languages: Learning, teaching, assessment* by Council of Europe, 2001

1.4. Target domain, tasks and specifications

Council of Europe (2000, p. 10) defines a domain as "the broad sectors of social life in which social agents operate," and identifies four broad domains: the public domain, the personal domain, the educational domain and the occupational domain.

The public domain encompasses all types of ordinary social interactions, e.g., with business and administrative bodies, cultural and leisure activities of a public nature, etc. The personal domain, on the other hand, refers to family relations and individual social practices. The occupational domain covers anything related to one's occupation, while

the educational domain refers to the learning and training context. (pp. 14 – 15).

Bachman and Palmer (p. 60), on the other hand, talk about a target language use domain and define it as "a specific setting outside of the test itself that requires the test taker to perform language use tasks." They distinguish between two types of domain: the language teaching domains, where language is used for the purpose of language learning and teaching, and the real-life domains, where language is used for purposes other than learning or teaching.

Considering the above definitions of the domain, the target language use domain of the research can be defined as the real-life public domain, where language is used in a setting where the goal is actual communication with either institutions or individuals.

As the first step in test design, Bachman (2002, p. 459) proposes the identification of the target language use domain, that is "a set of specific language use tasks that the test-taker is likely to encounter outside the test itself, and to which we want our inferences about language ability to generalize." (Bachman and Palmer, 1996, p. 44) The second step would be selecting appropriate tasks from the target language use domain, as well as designing the specification of the assessment task, which will then support the content relevance and coverage or representativeness part of our validity argument.

Other authors have also stressed the significance of task specifications. Alderson, Clapham, and Wall (1995, p. 9) define it as "the official statement about what the test tests and how it tests it" and maintain that they are essential for the test's construct validity. Brown's (1996, p. 234) view is in line with the one of Alderson, et al., as he talks about the item and test specification as the argument for the content validity of the test. McNamara (2006, p. 36) similarly believes that specifications of criterion behavior can be used as the basis for rating scales.

As discussed in Chapter Four of Part Two, the two steps proposed by Bachman can be problematic for two reasons, the first being that the target use domain is often difficult if not impossible to specify, or it is not clear. Even if the domain is clearly delineated, there remains the second problem: selecting specific tasks (Bachman, 1990, 2002).

Bachman lists three reasons why real-life tasks are not always appropriate. First, not all target language use tasks will engage the areas of ability that we want to assess; second, some of the tasks may not be practical to administer; third, some of the tests may not be appropriate for all the test-takers due to differences in their background knowledge

(Bachman, 2002, p. 460). The first problem is the one that the researcher has encountered to a certain extent. For that reason, the tasks chosen the purpose of the research are the ones that, first of all, belong to the real-life domain and are authentic in the sense that they reflect non-test language use. Another reason why these tasks have been chosen is that, due to the task characteristics and kind of response, they engage different areas of language knowledge that is of interest for the research, and in that way, enable an investigation of the individual components of the Bachman and Palmer framework, namely, one part of the underlying abilities. These individual components of the framework are investigated by means of observable behavior, that is, the performance product.

In addition, each of the test tasks that have been administered belongs to a more specific domain, with the intention of generalizing across a set of assessment tasks by means of Bachman and Palmer's framework. At the same time, because the test tasks are highly specific and reflect non-test performance, extrapolation to beyond the set of tasks, that is, to real-life tasks, is strengthened (Kane et al., 1999, p. 11).

1.5. The test

The test designed for the purpose of the research, aiming at gathering information about the test-takers' English language knowledge is a tailor-made, criterion-referenced performance-based test consisting of two parts, each with two extended production tasks. In addition, a short questionnaire on personal data has been administered.

The test consists of 2 parts: Part 1: written language tasks and Part 2: oral language tasks. There are two written tasks: Task 1 Writing and inquiry email and Task 2 Writing a blog entry, whereas in Part 2 there are two role-plays with an interviewer, randomly chosen from a pool of five.

Each of the test tasks is intended to test the language knowledge at a CEFR B2 level, using Bachman and Palmer's (2010, p. 45) model of language knowledge.

The test and task specifications, based on Bachman and Palmer's specifications as a basis for specifying assessment procedures, are provided on the following pages.

TEST SPECIFICATIONS	
A	Number of parts: Two - Written production, Spoken production
B	Number of tasks per part: • Part 1: 1) writing an inquiry email 2) writing an opinion blog entry/post • Part 2: 1) 2 role-plays
C	Sequence of parts/tasks: Part 1, in the order tasks are listed above, then Part 2, in the order tasks are listed above
D	Relative importance of parts/tasks: All parts of same importance
E	Time allotment: Part 1 80 minutes, Part 2 5 minutes per candidate
F	Instructions: general instructions and task-specific instructions or directive

Table 3. Test Specifications based on Bachman and Palmer's Blueprint

LANGUAGE TASK SPECIFICATIONS - WRITING TASK 1	
A	*Definition of construct to be assessed*: 1) writing skills in the real-life public domain by means of Bachman and Palmer's framework of language; 2) writing an enquiry email
B 1 2 3 4	Setting *Physical characteristics*: Classroom, quiet, comfortable *Equipment*: pencil, pen, eraser *Attributes of participants*: • Test takers: 1st-year university students of English, Italian • Test administrator: The researcher, positive attitude towards the test takers *Time of task*: By appointment within a fixed time period during the day
C	Characteristics of the input, expected response and relationship between input and response
1	Input a Format 1) *Channel*: visual (written paper-based text) 2) *Form*: language including a bullet list plus some non-language parts (illustrations) 3) *Language*: English 4) *Length*: flyer, approximately 45 words 5) *Vehicle*: reproduced 6) *Degree of speededness*: unspeeded/power test

	b Type: • input for interpretation: flyer advertising study holidays in the UK • prompt: directive to write an email with a list of details to include and information to obtain (90 words)
	c Language of input 1. Organizational characteristics: flyer a) *Grammatical* 1. *Syntax*: a few organized structures 2. *Vocabulary*: a range of general vocabulary (related to providing personal information), some topic-specific vocabulary (related to study holidays) 3. *Graphology*: typewritten b) *Textual (cohesion and organization)*: flyer a limited number of cohesive devices (*and*) and rhetorical organizational patterns c) *Pragmatic characteristics* Functional: input for interpretation: ideational function (use of language to inform, to express or exchange information about ideas, knowledge or feelings; descriptions and explanations). Prompt: manipulative (instrumental) Sociolinguistic: standard dialect, slightly formal register, natural but no idiomatic expressions, no figurative language d Topical characteristics: study holidays, travel and accommodation
2	Characteristics of the expected response a Format 1) *Channel*: visual (written paper-based) 2) *Form*: language 3) *Language*: English 4) *Length*: email approximately 200 words 5) *Type*: extended production 6) *Degree of speededness*: unspeeded, 35 minutes
	b Language characteristics 1. Organizational characteristics: flyer a) *Grammatical* 2. *Syntax*: a range of organized structures, standard English morphology, and syntax. 2. *Vocabulary*: a range of general vocabulary (related to providing personal information), some topic-specific vocabulary (related to study holidays, travel and accommodation) 2. *Graphology*: typewritten b) *Textual (cohesion and organization)*: a limited range of cohesive devices and rhetorical organizational patterns; clear paragraphing c) *Pragmatic characteristics* Functional: ideational function (use of language to inform, to express or

	exchange information about ideas, knowledge or feelings; descriptions and explanations). Manipulative (interpersonal functions): e.g., greetings. Sociolinguistic: specific genre (email), standard dialect, slightly formal register, natural but no idiomatic expressions, figurative language or cultural references c Topical characteristics: study holidays, travel and accommodation
3	Relationship between input and expected response and type of interaction a Type of external interactiveness: non-reciprocal b Scope of relationship: narrow c Directness of relationship: indirect

Table 4. Writing Task 1 Specifications

LANGUAGE TASK SPECIFICATIONS - WRITING TASK 2	
A	*Definition of construct to be assessed*: 1) writing skills in the real-life public domain by means of Bachman and Palmer's framework of language; 2) writing an opinion blog entry/post
B 1 2 3 4	Setting *Physical characteristics*: Classroom, quiet, comfortable *Equipment*: pencil, pen, eraser *Attributes of participants*: • Test takers: 1st-year university students of English, Italian • Test administrator: The researcher, positive attitude towards the test takers *Time of task*: By appointment within a fixed time period during the day
C	Characteristics of the input, expected response and relationship between input and response
1	Input a Format 1) *Channel*: visual (written paper-based text) 2) *Form*: language 3) *Language*: English 4) *Length*: n/a; 5) *Vehicle*: reproduced 6) *Degree of speededness*: unspeeded/power test
	b Type: • prompt: directive to write an opinion blog entry/post (40 words)
	c Language of input 1) Organizational characteristics: prompt a) *Grammatical* 1. *Morphology and syntax*: a few organized structures 2. *Vocabulary*: some general vocabulary, some topic-specific vocabulary (related to contemporary technologies)

	3. *Graphology*: typewritten b) *Textual (cohesion and organization)*: prompt a limited number of cohesive devices (*and*) and rhetorical organizational patterns c) *Pragmatic characteristics* Functional: manipulative (instrumental) functions Sociolinguistic: standard dialect, slightly formal register d Topical characteristics: general English, some topic-specific vocabulary (contemporary technologies)
2	Characteristics of the expected response a Format 7) *Channel*: visual (written paper-based) 8) *Form*: language 9) *Language*: English 10) *Length*: email approximately 300 words 11) *Type*: extended production 12) *Degree of speededness*: unspeeded, 45 minutes
	b Language characteristics 1) Organizational characteristics: prompt a) *Grammatical* 1. *Morphology and syntax*: a few organized structures; standard English morphology and syntax 2. *Vocabulary*: a range of general vocabulary, topic-specific vocabulary (contemporary technologies) 3. *Graphology*: handwritten b) *Textual (cohesion and organization)*: a range of cohesive devices and rhetorical organizational patterns; clear paragraphing c) *Pragmatic characteristics* Functional: ideational functions (use of language to inform, to express or exchange information about ideas, knowledge or feelings; descriptions and explanations) Sociolinguistic: specific genre (opinion blog post) standard dialect, slightly formal register, natural expressions, potential idiomatic expressions and figurative language c Topical characteristics: contemporary technologies
3	Relationship between input and expected response and type of interaction a Type of external interactiveness: non-reciprocal b Scope of relationship: narrow c Directness of relationship: indirect

Table 5. Writing Task 2 Specifications

LANGUAGE TASK SPECIFICATIONS - SPEAKING TASKS	
A	*Definition of construct to be assessed*: 1) speaking skills in the real-life public domain by means of Bachman and Palmer's framework of language knowledge; 2) performance of real-life tasks
B 1 2 3 4	Setting *Physical characteristics*: Classroom, quiet, comfortable *Equipment*: N/A *Attributes of participants*: • Test takers: 1st-year university students of English, Italian • Test administrator: The researcher as one of the examiners, plus another examiner; positive attitude towards the test takers *Time of task*: By appointment within a fixed time period during the day
C	Characteristics of the input, expected response and relationship between input and response
1	Input a Format 7) *Channel*: visual (role-cards); audio (conversation with the examiner) 8) *Form*: language 9) *Language*: English 10) *Length*: 30 – 40 words; 11) *Vehicle*: live 12) *Degree of speededness*: unspeeded/power test
	b Type: • input for interpretation: role-cards
	c Language of input 1) Organizational characteristics: 5 role-cards, 2 per student a) *Grammatical* 1. *Morphology and syntax*: a few organized structures 2. *Vocabulary*: a range of general vocabulary, some topic-specific vocabulary 3. *Graphology*: typewritten b) *Textual (cohesion and organization)*: role-card a limited number of cohesive devices (*and*) and rhetorical organizational patterns c) *Pragmatic characteristics* Functional: ideational, manipulative Sociolinguistic: genre: role-play card, standard dialect, informal register, natural but no idiomatic expressions, figurative language or cultural references d Topical characteristics: no specific topical knowledge needed
2	Characteristics of the expected response a Format

	13) *Channel*: audio 14) *Form*: language 15) *Language*: English 16) *Length*: 3 minutes approximately per role-play 17) *Type*: limited and extended production 18) *Degree of speededness*: unspeeded
	b Language characteristics 1) Organizational characteristics: prompt a) *Grammatical* 4. *Morphology and syntax*: a few organized structures; standard English morphology and syntax 5. *Vocabulary*: a range of general vocabulary, topic-specific vocabulary (directions, at the bus station, socializing, eating out) 6. *Graphology*: handwritten b) *Textual (cohesion and organization)*: a range of cohesive devices and conversational organizational patterns; c) *Pragmatic characteristics* Functional: ideational functions (use of language to inform, to express or exchange information about ideas, knowledge or feelings; descriptions and explanations); manipulative (interpersonal) functions (greetings); imaginative functions Sociolinguistic: no specific genre, standard dialect, semi-formal register at times, natural expressions, with potential idiomatic expressions and figurative language c Topical characteristics: some topical knowledge such as traveling, socializing, eating out)
3	Relationship between input and expected response and type of interaction a Type of external interactiveness: adaptive b Scope of relationship: narrow c Directness of relationship: indirect

Table 6. Speaking Task Specifications

The test/assessment specifications, together with task specifications, are one part of the assessment Blueprint (Bachman & Palmer, 2010, p. 370). The remaining three are procedures for setting cut scores and making decisions, procedures and formats for reporting assessment records, interpretations, and decisions; and procedures for administering the assessment. These first two components will be detailed in the following subsection, in light of the purpose of the test administered, while the third one will be addressed in the last subsection of the chapter.

1.6. Assessment criteria: Rating scales

The nature of performance-based second language assessment makes it possible to evaluate productive language skills – writing and speaking – by means of rating scales, which are a particularly significant means of marking in performance-based assessment. The scales can have a double purpose: to guide the rating process and to provide score interpretation. (McNamara, 1996, p. 182). This is because performance-based assessment quite often employs extended production responses.

Different rating scales provide different amounts of information on the test-taker's abilities. One of the main distinctions is the one between "global scales of language ability" also called "holistic" scales and analytic rating scales (Bachman & Palmer, 2010, p. 338; Weigle, 2002, p. 109, Alderson et al., 1995, p. 107), the main difference between them being the amount of information on test takers' abilities they provide. These scales contain several band or level descriptors that illustrate the competence at that level (Alderson at al. 1995).

1.6.1. Global or holistic scales

The defining characteristic of "global" or "holistic" scales is that they provide a single score for a task which is based on the overall impression (Weigle, p. 112), that is, a single general scale is used to give a single global rating (Brown, 1996, p. 61). According to Bachman and Palmer (2010), this type of scales, although they result in a single score, due to the fact that they are based on the view of language ability as a single unitary ability, frequently contain "different 'hidden' components of language ability" (p. 339).

One of the advantages of holistic scales is its practicality as it is enough to read the piece of writing once to give a score. Furthermore, they tend to focus on the test taker's strengths, not weaknesses. Finally, they also reflect a more authentic reaction of the reader, unlike the analytic scales (Weigle, 2002, p. 112).

Bachman and Palmer (2010, p. 339), however, focus on the drawbacks only and list three types of problems with holistic scales: problems of interpretations, difficulties in assigning levels, and differential weighting of components. The problems of interpretations originate in the global or holistic character of the scales, where quite often it is difficult to understand what a score reflects. They can refer to different areas of language knowledge, topical knowledge or target language use domains.

As a result, there is a problem of meaningfulness, that is the impossibility to know what these scales mean. In addition, these scales can be used for multiple tasks or domains and in such cases, it becomes difficult if not impossible to generalize the interpretations of the assessed language ability.

The second difficulty that Bachman and Palmer list is the difficulty that the raters encounter when they need to assess the level.

Finally, the fact that these scales often comprise "hidden" components of language ability implies that different raters, or the same rater on different occasions, may identify and give different amount of importance to different hidden language abilities.

Weigle (2002, p. 114) agrees that holistic scales may be difficult to interpret due to the reasons that Bachman and Palmer provide. As a disadvantage of holistic scales, she also addresses their failure to distinguish between different aspects of language ability.

Despite the obvious drawbacks of holistic scales, the fact that they reflect the authentic reaction of the reader defines their value in performance-based assessment.

1.6.2. Analytic scales

Whereas holistic scales focus on the global performance of the learner and, analytic or multi-trait scales (Weigle, 2002, pp. 114 - 115) use several criteria and provide descriptors for different levels of each criterion or aspect and for that reason are considered to be the most informative ones.

The rating scales will, first of all, be defined and designed according to the construct we intend to measure. After the construct has been defined, the different components of the construct that we intend to measure will be defined, and separate scales for separate components will need to be provided (Bachman & Palmer, 2010, p. 341).

Another issue that Bachman and Palmer address is criterion-referenced scales of language ability. Namely, unlike norm-referenced tests, which are designed to assess global language abilities of students, whose tests scores are then interpreted in relation to the scores of other students, criterion-reference tests intend to measure "well-defined and fairly specific objectives" (Brown, 2006, p. 2). Criterion-referenced tests measure whether a student has reached certain objectives without relating their score to other students' scores and as such provide information on students' mastery of a criterion Brown, 2006, Bachman 1990, Bachman & Palmer, 2010). Consequently, the scale levels are

defined in terms of levels of mastery: from a zero to a mastery level (Bachman & Palmer, 2010, p. 342).

An obvious advantage of analytic scales is that they can help distinguish between students' strengths and weaknesses, and provide useful diagnostic information, due to the fact that they comprise scales for different aspects or criteria of language ability (Weigle, 2002, p. 120; Bachman & Palmer, 2010, p. 342). This type of scales can also be easier to use, especially by inexperienced raters (Weigle, 2002, p. 120), and they actually reflect how raters think when assessing extended production tasks: in terms of individual or specific areas of language ability (Bachman & Palmer, 2010, p. 342).

Finally, a major drawback of analytic rating scales is that they are time-consuming as raters need to score a single extended production task by means of different criteria (Weigle, 2002, p. 120). Wigglesworth (2008, p. 116) also comments on the fact that the scales are "necessarily limited in scope" as they cannot possible encompass every single criterion or aspect of language ability.

Although there have been many discussions on the advantages and disadvantages of global and analytic scales, according to Wigglesworth (2008, p. 116) there has been little empirical research.

1.6.3. The rationale for the use of both holistic and analytic rating scales

The choice of the type of rating scale to use will ultimately depend on the test purpose and the inferences we want to make.

Since performance on the tasks, or task achievement, is one of the assessment constructs, and considering that holistic scales are seen as reflecting the authentic reaction of the reader (Weigle, 2002) the global or holistic rating scale has been used in the research to assess the level at which the task is completed, as well as the construct of performance of real-life tasks.

The analytic rating scales, however, have been designed on the basis of the Bachman and Palmer (2010) framework of language ability, focusing on the writing skills in the real-life public domain, in accordance with one of the assessment constructs.

In assessing the writing and the speaking skills of the students, a combination of multilevel and level-specific approach has been chosen as it is considered the most appropriate one in order to respond to the research questions. As Harsh and Rupp (2011, p. 2) explain, a multilevel

approach is the one in which "opens tasks are scored by trained raters using a scale that covers several bands or levels of proficiency," whereas in a level-specific approach "tasks are targeted at one specific level" and the students are then assessed in terms of fail/pass ratings. The tasks chosen for the test administered in the research were aimed at a specific level, CEFR B2 in particular, in accordance with the research questions; however, the analytic scales created on the basis of the CEFR illustrative descriptors and the Bachman and Palmer model of language knowledge have allowed raters to give marks that report on the students' level of proficiency and not only a pass/fail rating. However, the descriptors for the highest mark (four) have been designed having in mind the CEFR B2 level of proficiency since few if any students were expected to be at a level higher than that one.

The design of the rating scales was based on the adaptation of the existing CEFR scales for levels from A1 to B2, along with the creation of the missing ones.

CEFR DESCRIPTOR	ANALYTIC SCALE DESCRIPTOR WRITING	ANALYTIC SCALE DESCRIPTOR SPEAKING
Vocabulary Range Vocabulary Control	Vocabulary	Vocabulary
Linguistic General Range Linguistic Accuracy	Syntax / Grammar	Syntax / Grammar
Orthographic Control / Phonological Control	Graphology	Phonology
Coherence	Cohesion*	Cohesion*
Coherence / Conversation	Rhetorical knowledge*	Conversational knowledge
n/a	Functional knowledge	Functional knowledge
Sociolinguistic	Genre and Register	Genre and Register
Sociolinguistic	Natural/Idiomatic Expressions; Cultural References and Figures of Speech	Natural/Idiomatic Expressions; Cultural References and Figures of Speech

Table 7. Analytic scale descriptors adapted from the CEFR.

Furthermore, the holistic or global scales have been adapted from the CEFR descriptors for Written Interaction and Written Production, and

Spoken Interaction and Spoken Production, for the writing and speaking holistic scales respectively.

1.7. Rater training and standardization

Since performance-based assessment is based on human, and consequently subjective, judgments, selection, and training of raters are of utmost importance (Alderson at al. 1995; Bachman & Palmer, 2010; McNamara, 1996; Wigglesworth, 2008; Weir, 2005). The exact process of rater selection and training will, however, depend on the specific circumstances. Weigle (2002, p. 130) for example, maintains that the procedures can be less complex where the number of scripts is not high and where there are only two to three raters.

Bachman & Palmer (2010, p. 353) propose the following steps in rater preparation:

1. Read and discuss scales together.
2. Review language samples which have been previously rated by expert raters and discuss the ratings given.
3. Practice rating a different set of language samples. Then compare the ratings with those of experienced raters. Discuss the ratings and how the criteria were applied.
4. Rate additional samples and discuss.
5. Each trainee rates the same set of samples. Check for the amount of time taken to rate and for consistency.
6. Select raters who are able to provide reliable and efficient ratings.

Wigglesworth (2008, p. 117) suggests "double or even multiple ratings."

Similar procedures have been proposed by McNamara (1996, 2015; Alderson et al., 1995).

The proposed procedures have been adapted to the needs of the research. The scales were designed by the researcher, and both the training and standardization process took place during the pilot sample marking due to time constraints. All scripts were marked by two raters: the researcher and an experienced teacher, both with more than ten years of experience as examiners with an awarding body. After the pilot sample marking, the scales and the descriptors, how the criteria were applied and any potential problems with the scales and descriptors were discussed. Based on the outcomes of the discussion, the scales were slightly

modified, to address the problematic scripts, such as the ones with poor handwriting and the too short ones, to produce the final version of the scales (see Appendices A - D).

Since the CEFR levels, A1 to B2 descriptors were used in the design of the scales, aligning each scale level to a CEFR level and descriptors, after the statistical analyses were performed, the average mark between the two raters was aligned to the CEFR levels in the following way:

MARK	CEFR Level
1	CEFR A1 or lower
1.25/1.5 – 2	CEFR A2
2.25/2.5 – 3	CEFR B1
3.25/3.5 - 4	CEFR B2

Table 8. Scores and CEFR alignment

The average holistic marks, between the raters, for the individual tasks, would sometimes result in half-points, whereas the final average holistic mark in quarter-points. Any mark above the highest mark at the previous CEFR level was assigned to the next CEFR level, as shown in Table 8.

1.8. Assessment administration

The writing test was first administered to a pilot sample consisting of 54 second-year students of Educational Sciences of the Faculty of Psychology and Medicine, University of Sapienza, Rome.

The same test was then administered to a total of 189 first-year students, of the same department and university, on four occasions, as shown in Table 9.

Test Administration Date	Number of Students
16 March 2016	72
11 January	20
23 February 2017	8
13 March 2017	89

Table 9. Number of students according to the date of test administration.

A short questionnaire on personal data was administered together with the test (see Appendix E).

On both occasions, it was during regular lesson hours, due to the fact that the test was not mandatory for the students, and invigilated by the researcher. The students did not have access to their mobile phones, and there was a minimum distance of one meter between two students.

Instructions were given in both English and Italian, and any questions on instructions were answered.

The speaking test, however, has been administered to 29 students only. This was due to student unavailability, time constraints and lack of adequate premises. The test was not mandatory for the students, and although attempts were made to schedule the speaking test, the students most often did not appear for it. Nevertheless, a short overview of the speaking skills of the students who did take the speaking test is provided in Chapter Three of Part Three: Results.

Chapter Two
Test Validation

2.1. Introduction

The validation procedures employed in the validation of the test administered in the research are based on the validation frameworks and theories discussed in Chapter Four of Part Two, especially the Messick's principles contained in the validation frameworks by Bachman, (Assessment User Argument), Kane (Interpretation / Use Argument) and Weir. Messick's principle that the validity of a test resides in the test scores and score interpretations is the guiding principle of the validation process for the test administered to the university students. Bachman (1990, p. 238) similarly states that it is not the test content or test scores that we need to validate but the way we interpret or use the information we have gathered. As Kane et al. (1999, p. 6) state, "the interpretations assigned to assessment scores are said to be valid to the extent that these interpretations are supported by appropriate evidence," where there is a chain of interpretations and inferences to create as well as stages to follow. For the chain, that is, the interpretative argument to be convincing, "each of the separate inferences must be convincing" (Kane at al., 1996, p. 9).

Two distinct constructs are investigated in the research: writing and speaking skills in the real-life public domain by means of Bachman and Palmer's (1996, 2010) model of English language knowledge, and performance on real-life tasks. These two constructs are in line with the "construct-driven" and "task-driven" performance assessment (Messick, 1994, p. 14) or "weak" and "strong" language performance tests (McNamara, 1996, p. 43) and "task-centered" and "construct-centered" approach (Bachman, 2002, p. 454). The essential difference between these two approaches is in the inferences we want to make about the test-takers' knowledge: the first one is concerned with the underlying language ability, while the second one relates to how well test-takers perform a given task.

For the purpose of clarity, the evidential basis of the test validity will be addressed in four groups:

1) content relevance and coverage/context,
2) criterion relatedness/criterion validity,

3) scoring validity,
4) construct validity/theory-based validity.

2.2. Content Validity

As outlined in Chapter Four of Part Two, content validity encompasses both content relevance and content representativeness, and is defined by Messick (1995, p. 6) as "the boundaries of the construct to be assessed." The intention of the researcher has been to investigate the writing and speaking skills of the students in a public real-life domain. The domain is quite wide, although well delineated. Furthermore, completion of performance-based tasks is time-consuming since this type of tasks reflects non-test behavior or use of language, and as such is highly specific. As the main intention of the researcher is to investigate the performance on real-life tasks in the public domain, the content of the tasks is considered relevant and the tasks representative of the domain. They certainly are not the only tasks that a student may need to perform in real-life. To avoid having to compromise between authenticity and generalizability, the analytic scales have been employed as a means of generalization.

In addition to what has traditionally been referred to as content validity, Weir (2005, p. 19) adds the social dimensions of language use and for that reason refers to it as "context validity," a notion that encompasses both the traditional content validity and the context in which the tasks are completed. He also sees not only the task but also the administrative setting of the task as part of the context and maintains that efforts need to be made to incorporate real-life conditions into the test (p. 56). During the test administration, every attempt was made to approximate the conditions to real-life conditions, to the extent to which the classroom setting allowed it. Students were given clear instructions in both English and Italian, their native language, and students' questions answered. They were also given a clear purpose on the piece of writing and speaking that they were expected to produce and the task goal was clear. Finally, the conditions were uniform for all test administrations.

Finally, what should be the first step in developing a test (Bachman, 1990, p. 244), a definition of the content or ability domain to be assessed, is Chapter One of Part Three: Methodology. In addition, test as well as task specifications are provided in the same Chapter, while the tests are available in Appendices F and G.

2.3. Criterion-related Validity

Criterion-related validity or what Bachman (1990) calls criterion relatedness, as discussed in Chapter Four of Part Two, encompasses concurrent validity and predictive validity. Concurrent validity can be defined as the relationship between the test scores and another criterion which we believe indicates the tested ability (Bachman, 1990; Weir, 2005), for example, student self-assessment or teacher's rankings (Weir, 2005, p. 36). Furthermore, according to Kane et al., extrapolation to the real-life domain can be justified by means of criterion-related validity (1999, p. 11).

A Kendall's tau-b correlation has been computed to determine the relationship between the students' self-assessment of their English knowledge (competenza inglese), listening competence (competenza ascolto), speaking competence (competenza parlato), reading competence (competenza lettura), writing competence (competenza scrittura) and their English language grade (voto in inglese) at the end of the first semester of the last year of upper secondary school. As can be seen from the table below, there is a moderate, positive correlation, which is statistically significant.

		HOLISTIC1	HOLISTIC2	HOLISTIC AVERAGE
Competenza inglese	Coefficiente di correlazione	.458**	.413**	.437**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza ascolto	Coefficiente di correlazione	.401**	.311**	.351**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza parlato	Coefficiente di correlazione	.362**	.321**	.326**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza lettura	Coefficiente di correlazione	.346**	.265**	.287**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza scrittura	Coefficiente di correlazione	.303**	.257**	.255**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Voto in inglese	Coefficiente di correlazione	.344**	.316**	.323**
	Sign. (a due code)	0,000	0,000	0,000
	N	174	147	144

Table 10. Kendall Tau b correlation between student self-assessments, their grade in English and their Writing test results.

2.4. Scoring Validity

Considering that the approach to test validation taken in the research is the one of Bachman (1990), Bachman and Palmer (1996, 2010), Kane (1992), Messick (1994) and Weir (2005) where validity is seen as residing in the test scores, that is, a unified concept that encompasses the test reliability as well, what is traditionally known as "reliability" will be discussed under the term "scoring validity", as a superordinate concept for the different aspects of reliability (Weir, 2005).

Scoring validity can be defined as the extent to which assessment marks are "free from errors of measurement" (Weir, 2005, p. 23). Consequently, in order to properly address the issue of scoring validity for a specific language test, we need to consider the potential sources of error in the observed score on the particular test. What differentiates performance-based assessment from other types of assessment is the use of rating scales for marking. For that reason, in performance-based assessment, the measurement error most often has its origin in the fact that performance-based assessment, being rated by human raters, necessarily involves subjective judgment (McNamara, 1996, p. 117).

The two types of scoring validity examined for the purpose of validating the test scores are inter-rater reliability and internal test consistency.

2.4.1. Inter-rater reliability

To evaluate the inter-rater reliability, that is, the extent to which the raters' scores are consistent, the paired sample correlation coefficient for both analytic and holistic scales has been calculated for both the pilot and the actual sample: the bivariate Pearson correlation coefficient (with a two-tailed test of significance) for each pair of variables entered: Task 1 Vocabulary, Task 1 Syntax, Task 1 Graphology, Task 1 Cohesion, Task 1 Rhetorical Knowledge, Task 1 Functional Knowledge, Task 1 Genre and Register, Task 1 Natural and Idiomatic Expressions, Task 2 Vocabulary, Task 2 Syntax, Task 2 Graphology, Task 2 Cohesion, Task 2 Rhetorical Knowledge and Task 2 Natural and Idiomatic Expressions.

The pilot sample correlation coefficients range from $r = .828$ to $r = .972$, all at $p < .001$, which indicates a significant positive correlation, meaning that the raters' marks can be considered highly correlated.

		N	Correlazione	Sign.
Coppia 1	Task 1 Vocabulary	54	,898	,000
Coppia 2	Task 1 Syntax / Grammar	54	,931	,000
Coppia 3	Task 1 Graphology	54	,902	,000
Coppia 4	Task 1 Cohesion	54	,904	,000
Coppia 5	Task 1 Rhetorical Knowledge	54	,862	,000
Coppia 6	Task 1 Functional Knowledge	54	,828	,000
Coppia 7	Task 1 Genre & Register	54	,955	,000
Coppia 8	Task 1 Natural / Idiomatic Expressions	54	,972	,000
Coppia 9	Task 2 Vocabulary	54	,947	,000
Coppia 10	Task 2 Syntax / Grammar	54	,874	,000
Coppia 11	Task 2 Graphology	54	,939	,000
Coppia 12	Task 2 Cohesion	54	,949	,000
Coppia 13	Task 2 Rhetorical Knowledge	54	,914	,000
Coppia 14	Task 2 Natural/Idiomatic Ex.	54	,923	,000

Table 11. Pilot sample paired samples correlation coefficients for Writing Test analytic rating scales.

The same can be said for the holistic marks: the correlation coefficient $r = .943$ and $r = .939$ for Task 1 and Task 2 respectively, both at $p < .001$, indicate a strong positive correlation.

		N	Correlazione	Sign.
Coppia 1	Task 1 O Rater 1 & Task 1 O Rater 2	54	,943	,000
Coppia 2	Task 2 O Rater 1 & Task 2 O Rater 2	54	,939	,000

Table 12. Pilot test paired samples correlation coefficients for Writing test holistic rating scales.

With regard to the sample, the first-year students, the correlation coefficients for the analytic scale range from $r = .862$ to $r = .955$, all at $p < .001$, whereas for the holistic scale they are $r = .928$ and $r = .936$ for Task 1 and Task 2 respectively, both at $p < .001$ again indicating a strong positive correlation.

		N	Correlazione	Sign.
Coppia 1	Task 1 Vocabulary	179	0,945	,000
Coppia 2	Task 1 Syntax / Grammar	179	0,922	,000
Coppia 3	Task 1 Graphology	179	0,906	,000
Coppia 4	Task 1 Cohesion	179	0,955	,000
Coppia 5	Task 1 Rhetorical Knowledge	179	0,916	,000
Coppia 6	Task 1 Functional Knowledge	179	0,947	,000
Coppia 7	Task 1 Genre & Register	179	0,900	,000
Coppia 8	Task 1 Natural / Idiomatic Expressions	179	0,907	,000
Coppia 9	Task 2 Vocabulary	152	0,935	,000
Coppia 10	Task 2 Syntax / Grammar	152	0,877	,000
Coppia 11	Task 2 Graphology	152	0,883	,000
Coppia 12	Task 2 Cohesion	152	0,917	,000
Coppia 13	Task 2 Rhetorical Knowledge	152	0,862	,000
Coppia 14	Task 2 Functional Knowledge	152	0,946	,000
Coppia 15	Task 2 Genre & Register	152	0,883	,000
Coppia 16	Task 2 Natural / Idiomatic Expressions	152	0,921	,000

Table 13. Sample paired samples correlation coefficients for Writing test analytic rating scales.

		N	Correlazione	Sign.
Coppia 1	Task 1 O Rater 1 & Task 1 O Rater 2	179	,928	,000
Coppia 2	Task 2 O Rater 1 & Task 2 O Rater 2	152	,936	,000

Table 14. Sample test paired samples correlation coefficients for Writing test holistic rating scales.

As we can see, the inter-rater reliability coefficient is quite high, which may be explained by the fact that the two raters have a considerable teaching experience as well as experience in the area of assessment and CEFR levels.

2.3.2. Internal consistency of the test

One of the most commonly used method of calculating reliability (Bachman, 1990; Brown, 2002; Weir, 2005) for language tests is Cronbach's

Alpha, which estimates the internal consistency of a test as it "estimates the proportion of variance in the test scores that can be attributed to the score variance" (Brown, 2002, p. 17). Even though this type of reliability estimate is more appropriate for norm-referenced tests, where there is a high variance of scores, the performance-based test that has been administered revealed a relatively high variance of scores despite the fact that it is a criterion-referenced test.

The analysis of the pilot sample scores has yielded the following results:

Alfa di Cronbach	N. di elementi	Alfa di Cronbach	N. di elementi
,959	8	,948	6

Table 15. Pilot sample Cronbach's Alpha values for Writing Test Task 1 and Task 2.

	Media scala se viene eliminato l'elemento	Varianza scala se viene eliminato l'elemento	Correlazione elemento-totale corretta	Alfa di Cronbach se viene eliminato l'elemento
T1 Vocabulary	27,519	105,613	,887	,951
T1 Syntax	28,093	107,633	,821	,955
T1 Graphology	25,296	115,684	,694	,962
T1 Cohesion	27,611	105,789	,860	,953
T1 Rhetorical	27,907	107,784	,884	,951
T1 Functional	27,667	111,094	,878	,952
T1 Genre & Reg.	27,426	105,834	,879	,951
T1 Natural Ex.	27,944	110,318	,859	,953

Table 16. Pilot Task 1 Reliability Statistics.

	Media scala se viene eliminato l'elemento	Varianza scala se viene eliminato l'elemento	Correlazione elemento-totale corretta	Alfa di Cronbach se viene eliminato l'elemento
T2 Vocabulary	17,519	75,764	,887	,933
T2 Syntax	18,185	85,022	,820	,943
T2 Graphology	15,759	77,960	,713	,957
T2 Cohesion	17,537	75,046	,923	,929
T2 Rhetorical	17,833	78,858	,870	,935
T2 Natural Ex.	17,981	76,773	,885	,933

Table 17. Pilot Task 2 Reliability Statistics.

Considering that Cronbach's Alpha is $\alpha = .959$ and $\alpha = .948$ for Task 1 and Task 2 respectively, we can say that the pilot test is 95% reliable. The same can be said for the actual sample test, where $\alpha = .960$ for Task 1 and $\alpha = .957$ for Task 2.

Alfa di Cronbach	N. di elementi	Alfa di Cronbach	N. di elementi
,960	8	,957	8

Table 18. First-year sample Cronbach's Alpha values for Writing Test Task 1 and Task 2.

	Media scala se viene eliminato l'elemento	Varianza scala se viene eliminato l'elemento	Correlazione elemento-totale corretta	Alfa di Cronbach se viene eliminato l'elemento
T1 Vocabulary	27,492	84,251	0,889	0,952
T1 Syntax	28,215	86,070	0,893	0,951
T1 Graphology	25,674	101,565	0,588	0,967
T1 Cohesion	27,845	86,409	0,858	0,953
T1 Rhetorical	28,155	89,987	0,846	0,954
T1 Functional	27,917	84,076	0,919	0,949
T1 Genre & Reg.	28,166	89,906	0,852	0,954
T1 Natural Ex.	28,409	87,632	0,922	0,949

Table 19. First-year sample Task 1 Reliability Statistics.

	Media scala se viene eliminato l'elemento	Varianza scala se viene eliminato l'elemento	Correlazione elemento-totale corretta	Alfa di Cronbach se viene eliminato l'elemento
T2 Vocabulary	28,1513	88,169	,890	,948
T2 Syntax	29,1974	90,875	,891	,948
T2 Graphology	26,7829	105,310	,558	,966
T2 Cohesion	28,7632	90,447	,886	,948
T2 Rhetorical	28,9868	93,854	,859	,950
T2 Functional	29,3947	92,810	,893	,948
T2 Genre & Reg.	28,8026	89,895	,879	,949
T2 Natural Ex.	28,6645	92,754	,839	,951

Table 20. First-year sample Task 2 Reliability Statistics.

In addition, an exploratory factor analysis was performed to investigate the dimensionality of the scale. The pilot sample results reveal that 75,2% and 77% of the variance for Task 1 and Task 2 respectively explained for the pilot sample and 73,2% and 74,4% for the first-year university sample. One factor accounts for the total variance explained, which suggests that the scale items are unidimensional, that is, that they measure a single ability.

	Fattore 1	% di varianza
T1 Natural Ex.	,917	73,199
T1 Functional	,916	
T1 Vocabulary	,890	
T1 Syntax	,888	
T1 Cohesion	,871	
T1 Rhetorical	,861	
T1 Genre & Reg.	,860	
T1 Graphology	,597	

	Fattore 1	% di varianza
T2 Syntax	,921	74,365
T2 Natural Ex.	,918	
T2 Cohesion	,904	
T2 Functional	,903	
T2 Vocabulary	,902	
T2 Rhetorical	,880	
T2 Genre & Reg.	,851	
T2 Graphology	,560	

Table 21. First-year student sample factor matrix and explained variance for Writing Test Task 1 and Task 2.

The factor loadings for Task 1 and Task 2 respectively, that is, the correlation coefficient between the extracted factor and individual components are quite high. This is understandable considering that the analytic scale components are based on a framework of language knowledge. It may be justified by two aspects, the first one being that the students' knowledge in different areas is quite balanced, likely because of the approach to learning or their teachers' approach to teaching, where these areas are not taught separately. Secondly, the knowledge of certain areas depends on the knowledge of other areas, such as the knowledge of natural expressions, which depends on the knowledge in other subskills, as to be able to sound natural and be fluent, one needs to have not only the knowledge of grammar but also the knowledge of other areas.

2.4. Construct Validity

The research investigates two distinct constructs: writing and speaking skills in the real-life public domain by means of Bachman and Palmer's framework of language knowledge,) and performance on real-life tasks.

A detailed account of student performance is provided in Chapter Three of Part Three: Results. The model of language knowledge used as a basis for the analytic rating scales makes possible the generalization across the set of writing tasks despite the high task specificity.

Furthermore, because the tasks reflect non-test behavior, an attempt is made to extrapolate to real-life tasks.

The analyses that have been performed to evaluate construct validity are presented in the following subsections.

2.4.1. Correlations between the scores

According to Weir (2005, p. 242), the correlation coefficient between the scores on the various tasks indicate the degree of overlap between the tasks. With regard to the research in question, the correlations between the individual areas or analytic scales below indicate the degree to which these two scales, by means of two different tasks, measure the same areas. As shown in Table 22, there is a moderate to strong, positive, statistically significant relationship (all at $p < .001$)

		T2 Voc	T2 Syn	T2 Graph	T2 Coh	T2 Rhe	T2 Fun	T2 Gen	T2 Nat
T1 Voc	*r* di Pearson	**.738****	.756**	.515**	.733**	.638**	.714**	.653**	.746**
T1 Syn	*r* di Pearson	.679**	**.761****	.421**	.691**	.565**	.641**	.581**	.734**
T1 Graph	*r* di Pearson	.543**	.484**	**.551****	.507**	.387**	.473**	.512**	.522**
T1 Coh	*r* di Pearson	.687**	.755**	.452**	**.762****	.651**	.689**	.600**	.736**
T1 Rhe	*r* di Pearson	.668**	.747**	.475**	.673**	**.670****	.676**	.587**	.736**
T1 Fun	*r* di Pearson	.694**	.773**	.453**	.726**	.662**	**.700****	.599**	.758**
T1 Gen	*r* di Pearson	.705**	.719**	.453**	.699**	.608**	.651**	**.604****	.710**
T1 Nat	*r* di Pearson	.730**	.812**	.462**	.741**	.645**	.705**	.648**	**.796****

Table 22. Correlation between Writing Test Task 1 and Task 2 analytic scale components (N = 149, correlation is significant at the 0.01 level (2-tailed).

The correlation coefficient for Graphology $r = .551$ is the lowest one, which may be explained by the fact that the students were provided with a longer input for Task 1 than for Task 2. Consequently, students made

different types of spelling mistakes – whereas in Task 1 they could base their answer on the input and simply copy the words and correct spelling from the input, they were not able to do so for Task 2, where they needed to provide content and may have done so focusing on the content and not on spelling.

Genre and register were also different for the two tasks. In Task 1, they needed to write an email, while in Task 2, a blog post. Accordingly, in Task 1, apart from the genre, the register was semi-formal, while in Task 2 it was informal. The low correlation coefficient may be due to this difference.

Finally, the low correlation coefficient for rhetorical knowledge may be explained by the fact that the two tasks requested different types of rhetorical knowledge, especially because Task 2 requested a higher degree of rhetorical knowledge and the expected response was longer.

		HOLISTIC T1	HOLISTIC T2
HOLISTIC T1	Correlazione di Pearson	1	.747**
	Sign. (a due code)		0,000
	N	179	149
HOLISTIC T2	Correlazione di Pearson	.747**	1
	Sign. (a due code)	0,000	
	N	149	152
**. La correlazione è significativa a livello 0,01 (a due code).			

Table 23. Correlation between Writing test Task 1 and Task 2 marks.

The correlation coefficients between the scores for the two tasks support the generalization across the set of tasks, as they indicate that the same construct is measured by these two writing tasks (Messick, 1996, p.11).

According to Bachman (1990, p. 260), a high correlation coefficient between two tests may be due to the fact the scores are affected by a common trait, by the method or both. In the case of the tasks administered in the research, this does not challenge the inferences made based on the correlations because the main interest is to investigate the specific trait, in

particular, the writing skills, on performance-based tasks, which is used as a method.

Furthermore, the correlations between the holistic marks and analytic scale marks, tell us about the extent to which each of the analytic scale components or areas of language knowledge correlates with the holistic mark. Taking into account that all the correlations are statistically significant ($p < .001$), we can see from the Table 24 that for Task 1, the correlation coefficient of $r = .876$ for the knowledge of natural expressions, is the highest one, which may be explained by the fact that the knowledge of natural expressions is what makes one sound fluent in a foreign language, independent of the level of knowledge, especially when one needs to produce a piece of writing, such as the one in Task 1, that is, an enquiry email.

		T1 Voc	T1 Syn	T1 Graph	T1 Coh	T1 Fun	T1 Rhe	T1 Gen	T1 Nat
OLI 1	*r* di Pearson	.848**	.838**	.602**	.824**	.826**	.864**	.802**	.876**
	Sign. (a due code)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	N	179	179	179	179	179	179	179	179
**. La correlazione è significativa a livello 0,01 (a due code).									

Table 24. Correlation between Writing test Task 1 holistic mark and analytic scale marks.

For Task 2, however, the highest correlation coefficient is the one for the functional knowledge ($r = .861$), followed by the ones for vocabulary and syntax (both $r = .847$), which may be explained by the nature of Task 2: an argumentative piece of writing, where the knowledge of functions is essential.

		T2 Voc	T2 Syn	T2 Graph	T2 Coh	T2 Fun	T2 Rhe	T2 Gen	T2 Nat
OLI2	*r* di Pearson	.847**	.847**	.492**	.843**	.861**	.838**	.772**	.823**
	Sign. (a due code)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	N	152	152	152	152	152	152	152	152
**. La correlazione è significativa a livello 0,01 (a due code).									

Table 25. Correlation between Writing test Task 2 holistic mark and analytic scale marks.

		Voc	Syn	Graph	Coh	Fun	Rhe	Gen	Nat
OLI	*r* di Pearson	,890**	,889**	,622**	,894**	,913**	,887**	,852**	,889**
	Sign. (a due code)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	N	149	149	149	149	149	149	149	149
**. La correlazione è significativa a livello 0,01 (a due code).									

Table 26. Correlation between the average Writing test holistic mark and analytic scale marks.

The correlations between the average holistic marks and average analytic marks, however, tell us that all the areas of language knowledge, except for graphology, have a strong positive relationship with the overall mark. Considering that a language is normally learned as a language at a certain level, and now its individual areas, we can say that this is an expected result. The overall graphology mark, on the other hand, correlates more moderately with the holistic mark, which could be due to the fact that the students' spelling of the vocabulary and grammar they used was quite good, which will be discussed in more detail in the Chapter Three of Part Three: Results.

Chapter Three
Results

3.1. Test Takers' Characteristics

As outlined in Chapter One of Part Three: Methodology, the test was administered to a pilot sample consisting of 54 second-year students of Educational Sciences of the Faculty of Psychology and Medicine, University of Sapienza, Rome.

The same test was then administered with a total of 189 first-year students, of the same department and university, on four occasions, as detailed in Chapter One of Part Three.

A short questionnaire on personal data was administered together with the test. The questionnaire required the students to provide information on their age, country of origin, the type of upper-secondary school they attended, the language or languages that they speak at home, if they have studied another language except for English, the grade in English at the end of the first semester of the last year of upper secondary school, self-assessment of their English language knowledge, self-assessment of their English language listening, speaking, reading and writing skills, if they have gained a certificate in English and which level, if they have ever studied in an English speaking country, if they have attended and English language course outside school, and if they have passed the university qualifying exam or better known as idoneità in Italian (see Appendix E for the Student Questionnaire).

The speaking test, however, has been administered to 29 students only. This was due to student unavailability, time constraints and lack of adequate premises.

3.1.1. Personal Characteristics

The personal data of the test takers identified by means of the questionnaire are displayed in the following tables.

Most of the students who have completed the test are between 18 and 20 years old, 75.7% of them, while the second largest group in terms of age are the ones from 21 to 26 years old, 20.6%. These two groups account for the majority of the sample, 96.3%.

	Frequenza	Percentuale	Percentuale cumulativa
18 - 20	143	75.7	75.7
21 - 26	39	20.6	96.3
28 - 36	3	1.6	97.9
40 - 53	4	2.1	100.0
Totale	189	100.0	

Table 27. Number of students per age.

The same percentage of students, 96.3% are Italian, while only seven of them were born in another country.

	Frequenza	Percentuale	Percentuale cumulativa
Italia	182	96,3	96,3
Altro paese	7	3,7	100,0
Totale	189	100,0	

Table 28. Number of students per country of origin.

The largest number of students come from either liceo socio-psico-pedagogico (*n* = 52), 27.5% of the total number of students, and from liceo scientifico (*n* = 51), 27% of them.

	Frequenza	Percentuale	Percentuale cumulativa
istituto professionale	14	7,4	7,4
istituto tecnico	16	8,5	15,9
liceo classico	35	18,5	34,4
liceo linguistico	15	7,9	42,3
liceo scientifico	51	27,0	69,3
liceo socio-psico-pedagogico	52	27,5	96,8
altro	6	3,2	100,0
Totale	189	100,0	

Table 29. Number of students per school of origin (upper-secondary school).

Most of the students speak only Italian at home (*n* = 181), 95.8%, while four of them speak Italian and another language, and four only another language.

	Frequenza	Percentuale	Percentuale cumulativa
Albanese	1	0,5	0,5
Italiano	181	95,8	96,3
Italiano e inglese	1	0,5	96,8
Italiano e rumeno	1	0,5	97,4
Italiano e spagnolo	2	1,1	98,4
Rumeno	2	1,1	99,5
Spagnolo	1	0,5	100,0
Totale	189	100,0	

Table 30. Number of students according to the language they speak at home.

For only 5.3%of the students (*n* = 10), English is the only foreign language they have studied, while all other students have studied at least another foreign language, if not two.

	Frequenza	Percentuale	Percentuale cumulativa
	1	0,5	0,5
francese	78	41,3	41,8
francese e arabo	1	0,5	42,3
francese, tedesco	3	1,6	43,9
giapponese	1	0,5	44,4
nessuna	10	5,3	49,7
russo, italiano	1	0,5	50,3
spagnolo	53	28,0	78,3
spagnolo, francese	33	17,5	95,8
spagnolo, greco	1	0,5	96,3
spagnolo, italiano	1	0,5	96,8
spagnolo, tedesco	2	1,1	97,9
tedesco	4	2,1	100,0
Totale	189	100,0	

Table 31. Number of students according to the foreign languages they have studied.

In terms of the English language grade at the end of the first semester of the last year of upper-secondary school, the majority of students had grades 6 or 7 ("discreto"), 42.3% of them, while 29.1% had grade 5 ("sufficiente"). Finally, 24.3% of the students had 8 or 9 ("buono").

		Frequenza	Percentuale	Percentuale cumulativa
Valido	sufficiente	55	29,1	30,4
	discreto	80	42,3	74,6
	buono	46	24,3	100,0
	Totale	181	95,8	
Mancante	Sistema	8	4,2	
Totale		189	100,0	

Table 32. Number of students per grade in the first semester of upper-secondary school.

To get a better insight into students' background knowledge of English, they were asked whether they hold an internationally recognized certificate in English, whether they have ever studied in an English-speaking country, whether they have passed the University qualifying exam (idoneità) and whether they have ever attended a course in English outside of school.

78.8% of the students (n = 149) do not possess an internationally recognized certificate in English, while the rest of them, n = 49, possess a certificate in English at one of the CEFR levels as can be seen from Table 33.

	Frequenza	Percentuale	Percentuale cumulativa
No	149	78,8	78,8
A1	4	2,1	81,0
A2	14	7,4	88,4
B1	8	4,2	92,6
B2	10	5,3	97,9
C1	4	2,1	100,0
Totale	189	100,0	

Table 33. Number of students per internationally recognized certificate in English.

Only 22.8% of the students (n = 43) have studied in an English-speaking country.

	Frequenza	Percentuale	Percentuale cumulativa
No	146	77,2	77,2
Sì	43	22,8	100,0
Totale	189	100,0	

Table 34. Number of students according to whether or not they have studied in an English-speaking country.

Most students have not yet passed the university qualifying exam, 74.1% of them (n = 140).

	Frequenza	Percentuale	Percentuale cumulativa
No	140	74,1	74,1
Sì	49	25,9	100,0
Totale	189	100,0	

Table 35. Number of students according to whether or not they have passed the university qualifying exam (idoneità).

Almost half of the students, however, have taken an English language course, 48.7%, as opposed to the 51.3% of those who have not.

	Frequenza	Percentuale	Percentuale cumulativa
No	97	51,3	51,3
Sì	92	48,7	100,0
Totale	189	100,0	

Table 36. Number of students according to whether or not they have taken a course in English outside of school.

3.1.2. Students' self-assessment

The students were also asked to evaluate their English language knowledge on a scale from 1 to 4, where 1 equals to "low" ("bassa"), 2 "sufficient" ("sufficente"), 3 "moderate" ("discreta"), and 4 "good" ("buona") for general English knowledge, listening comprehension skills, speaking skills, reading comprehension skills and writing skills.

With regard to the general English language knowledge, as can be seen from Figure 15, the largest number of students, 37.6% of them (n = 71) assessed their knowledge as sufficient, 30.2% (n = 57) as low, 25.4% (n = 48) as moderate, and only 6.9% (n = 13) as good.

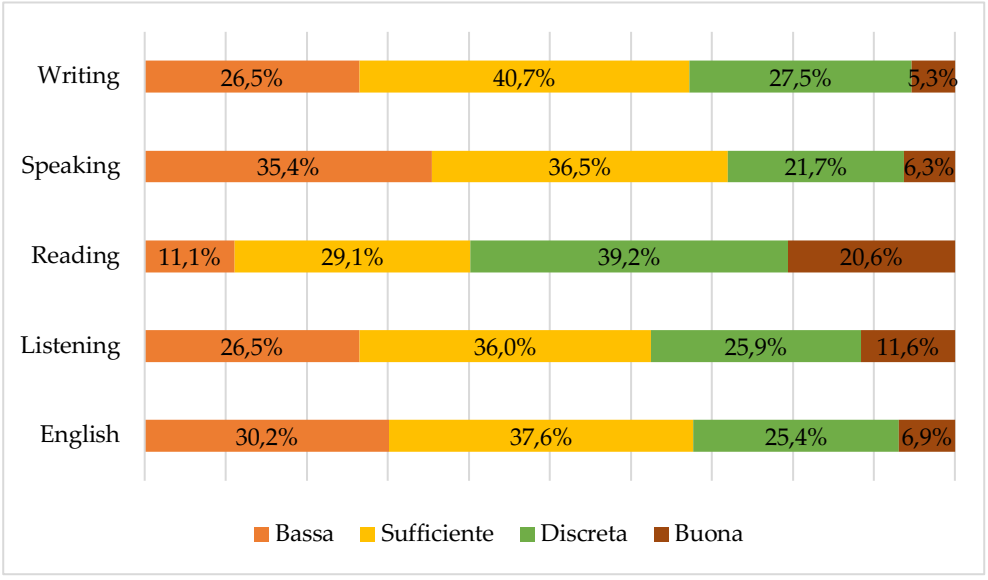

Figure 15. Students' self-assessment of their writing, speaking, reading and listening skills for the total of 189 students.

With regard to students' self-assessment of their individual language skills: writing, speaking, reading and listening, we can see from Figure 15, that the highest number of students, 20.6% believe that their reading comprehension skills are good, and 11.6% that their listening comprehension skills are good. However, only 5.3% and 6.3% of students believe that their writing and speaking skills are good. This could be explained by the fact that writing and speaking skills are the productive ones, which request students to actually produce language. Listening and

reading skills, on the other hand, are receptive ones, and consequently easier to acquire or learn, as students have access to texts and songs in English via the Internet. Most often, students evaluated their skills as either sufficient: 40.7% for writing, 36.5% for speaking, 29.1% for reading and 36% for writing. The students' opinion on their speaking skills is particularly poor: as many as 35.4% of them think that their speaking skills are low.

The extent to which students' self-assessment and personal information reflect on their performance will be detailed in the following subsections. Furthermore, their self-assessment in relation to the CEFR levels of their performance will be discussed in the last subsection of the chapter.

3.2. Holistic Scale Marks

As shown in Table 37, $n = 179$ students completed Task 1, while $n = 152$ have completed Task 2. $n = 10$ and $n = 37$ students failed to complete Task 1 and Task 2 respectively. The total number of students who completed both tasks is $n = 149$.

		Holistic 1	Holistic 2	Average Holistic
N	Valido	179	152	149
	Mancante	10	37	40
Media		1,8408	1,9474	1,9295
Mediana		2,0000	2,0000	2,0000
Deviazione std.		0,82860	0,88990	0,80017
Varianza		0,687	0,792	0,640

Table 37. Number of students who completed the writing test tasks and mean marks.

The average Task 1 and Task 2 marks are shown in the following two tables. The highest number of students, $n = 70$, obtained mark One for Task 1, while the least frequent marks for Task 1 are the highest ones: 3.50 ($n = 1$) and 4 ($n = 4$).

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulativa
Valido	1.00	70	37,0	39,1	39,1
	1.50	11	5,8	6,1	45,3
	2.00	52	27,5	29,1	74,3
	2.50	7	3,7	3,9	78,2
	3.00	34	18,0	19,0	97,2
	3.50	1	0,5	0,6	97,8
	4.00	4	2,1	2,2	100,0
	Totale	179	94,7	100,0	
Mancante	Sistema	10	5,3		
Totale		189	100,0		

Table 38. Distribution of Writing Test Task 1 holistic marks.

Similarly, the most frequent mark for Task 2 is One (n = 54), while the least frequent marks are 2.50 (*n* = 4), 3.50 (*n* = 5) and 4 (*n* = 6).

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulativa
Valido	1.00	54	28,6	35,5	35,5
	1.50	7	3,7	4,6	40,1
	2.00	48	25,4	31,6	71,7
	2.50	4	2,1	2,6	74,3
	3.00	28	14,8	18,4	92,8
	3.50	5	2,6	3,3	96,1
	4.00	6	3,2	3,9	100,0
	Totale	152	80,4	100,0	
Mancante	Sistema	37	19,6		
Totale		189	189		

Table 39. Distribution of T2 holistic marks.

Only 29 students out of the total of 189 completed the speaking test.

	Holistic 1	Holistic 2	Average Holistic
Media	1,9138	1,9310	1,9224
Mediana	2,0000	2,0000	1,7500
Deviazione std.	0,90701	0,95173	0,89169
Varianza	0,823	0,906	0,795

Table 40. Speaking test mean marks (n = 29).

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulativa
Valido	1.00	11	5,8	37,9	37,9
	1.50	2	1,1	6,9	44,8
	2.00	7	3,7	24,1	69,0
	2.50	2	1,1	6,9	75,9
	3.00	5	2,6	17,2	93,1
	3.50	1	0,5	3,4	96,6
	4.00	1	0,5	3,4	100,0
	Totale	29	15,3	100,0	
Mancante	Sistema	160	84,7		
Totale		189	100,0		

Table 41. Distribution of Speaking Test Task 1 marks.

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulativa
Valido	1.00	9	4,8	31,0	31,0
	1.50	5	2,6	17,2	48,3
	2.00	8	4,2	27,6	75,9
	2.50	1	0,5	3,4	79,3
	3.00	3	1,6	10,3	89,7
	4.00	3	1,6	10,3	100,0
	Totale	29	15,3	100,0	
Mancante	Sistema	160	84,7		
Totale		189	100,0		

Table 42. Distribution of Speaking Test Task 2 marks.

3.2.1. Personal characteristics and test holistic scale marks

The data collected through the questionnaire have been used to compare the holistic marks for both tasks of different groups of students for each of the independent variables: the age, country of origin, school of origin, whether they have studied in an English-speaking country, whether they have passed their university qualifying exam (idoneità) and their self-assessments.

For the writing test, analyses have indicated that the mean average holistic mark of the students who hold an internationally recognized certificate in English (M = 2.36 (SD = .928) and M = 2.33 (SD = .955), for Task 1 and Task 2 respectively) is greater than the mean values of the ones who do not (M = 1.66 (SD = .730) and M = 1.84 (SD = .829) for Task 1 and Task 2 respectively). In the same way, it is greater for the students who have studied abroad (M = 2.16 (SD = .898) and M = 2.24 (SD = .991) for Task 1 and Task 2 respectively, against M = 1.75 (SD = .785) and M = 1.85 (SD = .836) who have not) as well as for the ones who have passed the university qualifying exam in English (M = 2.13 (SD = .811) and M = 2.32 (SD = .960) for Task 1 and Task 2 respectively against M = 1.78 (SD = .814) and M = 1.81 (SD = .826) who have not).

Furthermore, as can be seen from Figure 16, where the mean average writing test holistic mark is shown, the mean marks of the students who have studied in an English-speaking country, who have passed the university qualifying exam and who possess an internationally recognized certificate in English are quite higher than of the students who have not studied abroad, who have not passed the university qualifying exam and who do not possess a certificate in English. This may be explained by the additional hours they have spent studying for the exams and practicing their English while traveling and studying.

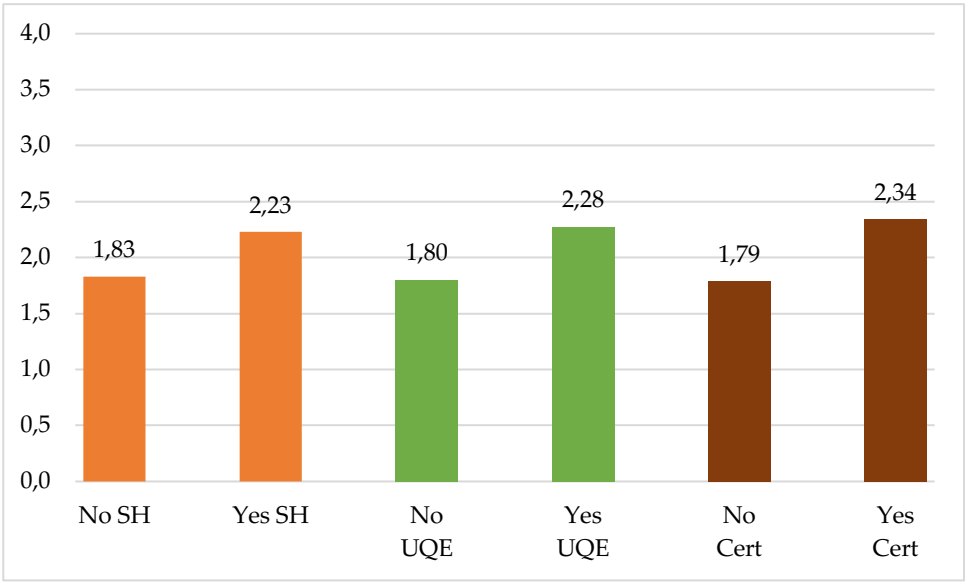

Figure 16. Mean average Writing test holistic marks of the students who have not and who have gone on study holidays (SH), who have not and who have passed the university qualifying exam (UQE) and who do not and who have an internationally recognized certificate in English (Cert).

With regard to the speaking test, the effects of students' experience with English outside the classroom are quite similar to the effects of these factors on the writing test: the mean average holistic mark of the students who have studied in an English-speaking country, have passed the university qualifying exam and possess an internationally recognized certificate in English is higher than of the ones who have not or do not, as shown in Figure 16. This is particularly true for the students who have studied in an English-speaking country as their mean average mark is M = 2.66 (SD = 1.043), while it is M = 1.64 (SD = .066) of the ones who have not.

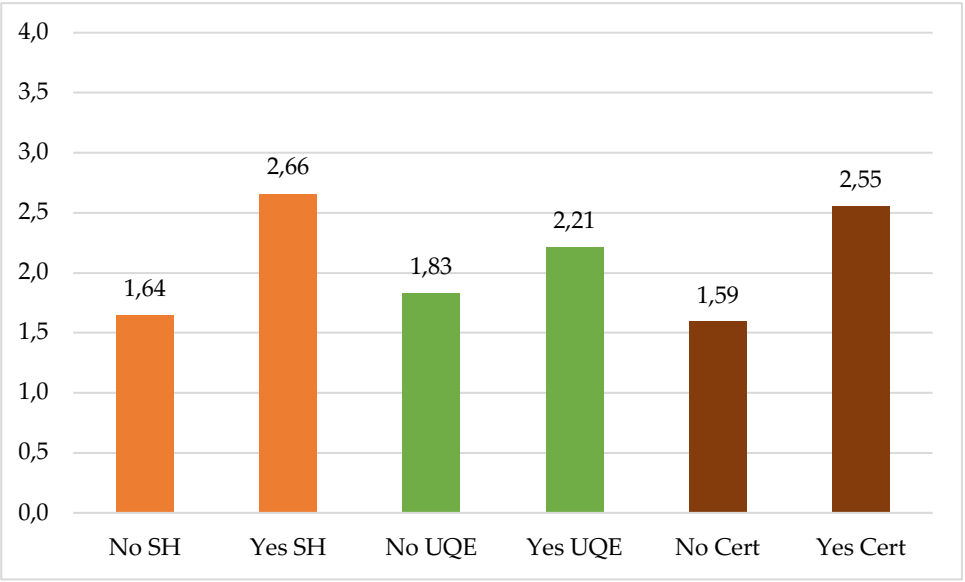

Figure 17. Mean average Speaking Test holistic marks of the students who have not and who have gone on study holidays (SH), who have not and who have passed the university qualifying exam (UQE) and who do not and who have an internationally recognized certificate in English (Cert).

Also, the higher the level of the certificate a student possesses is, the higher their average holistic mark is as shown in Figure 18. The mean holistic mark of those who do not hold a certificate in English is, however, slightly higher than of the ones who hold a CEFR A2 certificate, as is the case with the writing test average marks, where the average mark of the students who possess a CEFR A1 certificate is lower than of the ones who do not. It is obvious there are students who, despite their level of English, never felt or had the need to certify their knowledge, while the ones who have done so at lower levels (such as CEFR A1 or A2) may have stopped studying the language once they obtained the certificate.

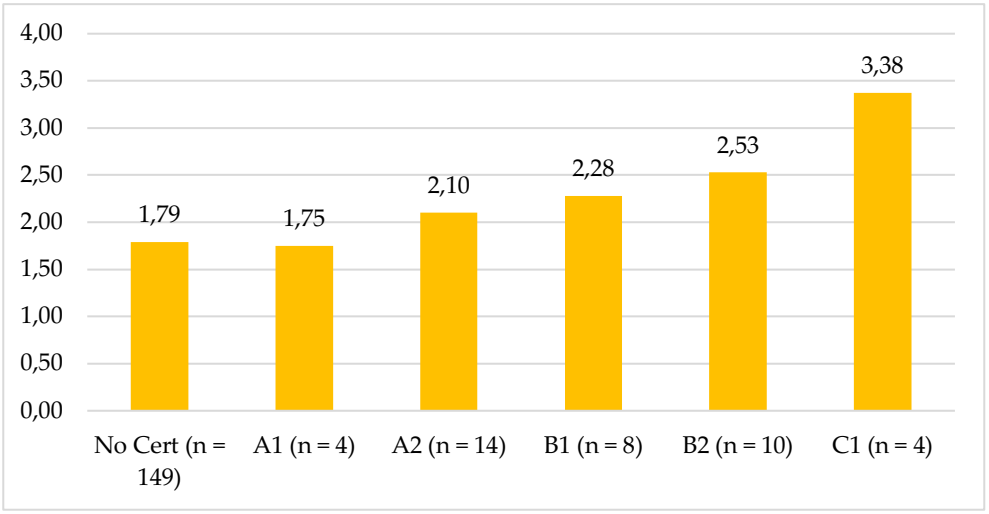

Figure 18. Writing test mean average holistic mark of students who do not and who do hold a certificate in English at levels CEFR A1 to C1.

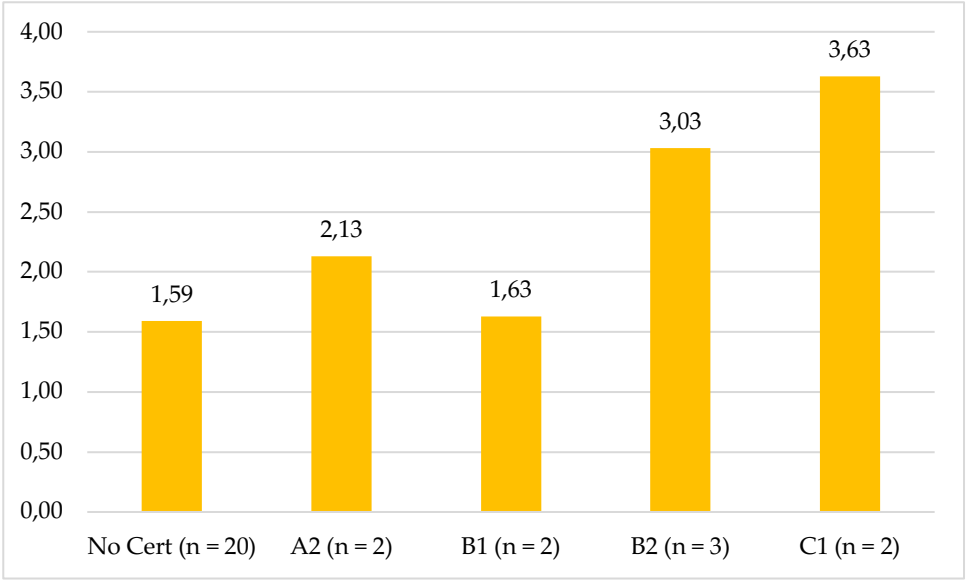

Figure 19. Speaking test mean average holistic mark of students who do not and who do hold a certificate in English at levels CEFR A2 to C1.

Even though some of the students possess a certificate in English, at a CEFR B1 level, for example (n = 2 of the speaking test sample), their level of English is considerably lower, as the mean average mark of these two

students is $M = 1.63$. This is the most striking example though it is only two students. However, the same is also true for the writing test: students in possession of a CEFR B2 certificate have a mean average mark of $M = 2.53$ only, out of four, which equals CEFR level B2. Finally, students who hold a CEFR C1 certificate in English are nowhere near CEFR C1 level when their writing or speaking skills are considered.

3.2.2. Self-assessment and holistic scale marks

Kendall's Tau-b correlation coefficient of $\tau = .437$ ($p < .001$) indicates a moderate positive relationship between the students' self-assessments of English language knowledge and their average holistic mark on the writing test. Similarly, there is a moderate positive correlation between their grade in English at the end of the first semester of the fifth year of upper-secondary school and their performance on the tasks, $\tau = .323$ ($p < .001$) as reported in Table 43.

The correlation is lower, however, for the self-assessment of the writing skills $\tau = .255$ ($p < .001$), although the test that has been administered is a test of writing. The students' self-assessment in relation to their performance in terms of CEFR levels will be discussed in the last subsection of this chapter.

		Holistic 1	Holistic 2	Holistic
Competenza inglese	Coefficiente di correlazione	.458**	.413**	.437**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza ascolto	Coefficiente di correlazione	.401**	.311**	.351**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza parlato	Coefficiente di correlazione	.362**	.321**	.326**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza lettura	Coefficiente di correlazione	.346**	.265**	.287**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Competenza scrittura	Coefficiente di correlazione	.303**	.257**	.255**
	Sign. (a due code)	0,000	0,000	0,000
	N	179	152	149
Voto in inglese	Coefficiente di correlazione	.344**	.316**	.323**
	Sign. (a due code)	0,000	0,000	0,000
	N	174	147	144

Table 43. Writing Test marks and student self-assessments correlation coefficients. Correlation is significant at the 0.01 level (2-tailed).

		Writing Holistic	Speaking Holistic
Competenza inglese	Coefficiente di correlazione	.437**	.415**
	Sign. (a due code)	0,000	0,006
	N	149	29
Competenza ascolto	Coefficiente di correlazione	.351**	.409**
	Sign. (a due code)	0,000	0,007
	N	149	29
Competenza parlato	Coefficiente di correlazione	.326**	.280
	Sign. (a due code)	0,000	0,066
	N	149	29
Competenza lettura	Coefficiente di correlazione	.287**	.427**
	Sign. (a due code)	0,000	0,006
	N	149	29
Competenza scrittura	Coefficiente di correlazione	.255**	.366*
	Sign. (a due code)	0,000	0,016
	N	149	29
Voto in inglese	Coefficiente di correlazione	.323**	.415**
	Sign. (a due code)	0,000	0,006
	N	144	29

Table 44. Writing and Speaking tests holistic marks correlation coefficients with student self-assessments. Correlation is significant at the 0.01 level (2-tailed).

As shown in Table 44, the correlations with the Speaking test average holistic mark are lower and not all significant. The lowest correlation coefficient is the one between students' self-assessment of the speaking skills and their average speaking test holistic marks (r = .280, p = .066). This may mean that the students do not know how they actually know or do not know.

The mentioned independent variables are the ones that positively influence the dependent ones. The rest of the data collected through the questionnaire, such as a course in the English language, the language that they speak at home, other language or languages that they have studied, they did not prove significant for the student performance. Furthermore, the students who have taken a course in English (n = 88) performed worse on the test (M = 1.79, SD = .840 for Task 1 and M = 1.83, SD = .837 for Task

2) than the ones who have not (M = 1.89, SD = 819 for Task 1 and M = 2.07, SD = .932 for Task 2).

3.2.3. School of origin and writing test performance

With regard to the school of origin, the highest mark was achieved by the students coming from the classical lyceum (M = 2.43, SD = .809), despite the fact that the highest number of English language lessons (four hours a week in the last three years, as opposed to three hours a week in other lyceums) is in linguistic lyceums (M = 2.10, SD = .666), not in the classical ones, while the lowest was the one of the students coming from professional institutes (M = 1.15, SD = .242). The average holistic mark on the writing test of the students coming from the scientific lyceum was M = 2.07 (SD = .743), of the students coming from the socio-pedagogic one M = 1.58 (SD = .685), while the average holistic mark of the students coming from the technical institutes was M = 1.46 (SD = .509).

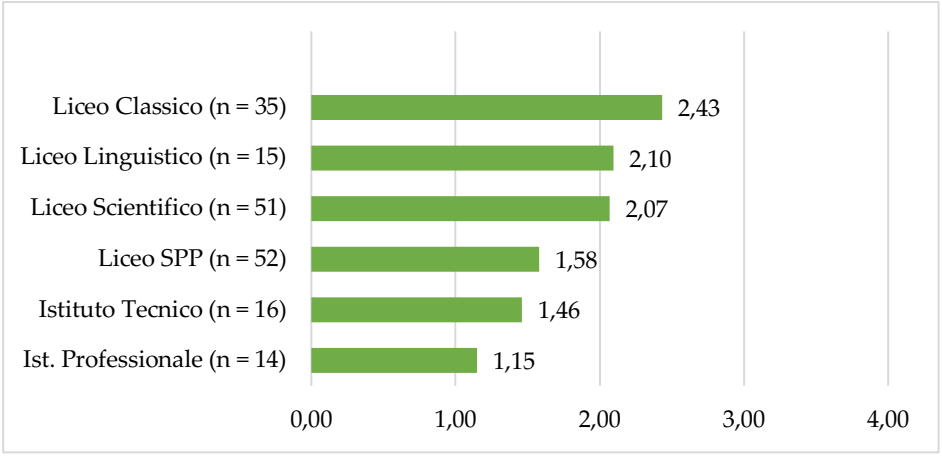

Figure 20. Mean average holistic marks on the Writing test per school of origin.

Post hoc comparisons using the Bonferroni test indicated that the mean scores of the students coming from professional institutes were significantly different from the ones of the students coming from the classical lyceum at $p < .001$, linguistic lyceum at $p = .031$ and the scientific one $p = .005$. Also, the scores of the students coming from the classical lyceum were significantly different from the students coming from the technical institutes at $p = .001$ and socio-pedagogic ones at $p < .001$, while the scores of the students coming from socio-pedagogic lyceums were

also significantly different from the ones coming from the scientific ones, at $p = .044$. No other differences between the scores of the students coming from different schools have proven to be statistically significant.

3.3. Analytic Rating Scale Results

In terms of individual components of Bachman and Palmer's framework, employed to assess the individual components of language knowledge, Vocabulary, Syntax, Graphology, Cohesion, Rhetorical knowledge, Functional knowledge, Genre and Register, and Knowledge of natural expressions, we can see from Figure 21, the differences in the mean analytic mark per component per task, with the standard deviation of *SD* = .860 for Task 1 and *SD* = .882 Task 2 Vocabulary, *SD* = .804 for Task 1 and *SD* = .803 Task 2 Syntax, *SD* = .533 Task 1 and *SD* = .609 Task 2 Graphology, *SD* = .818 for Task 1 and *SD* = .820 Task 2 Cohesion, *SD* = .723 for Task 1 and *SD* = .742 for Task 2 Rhetorical knowledge, *SD* = .840 for both Task 1 and Task 2 Functional knowledge, *SD* = .719 for Task 1 and *SD* = .789 for Task 2 Genre and Register, and *SD* = .738 Task 1 and *SD* = .747 Task 2 Knowledge of natural expressions. There do not seem to be major differences between the Task 1 and Task 2 marks.

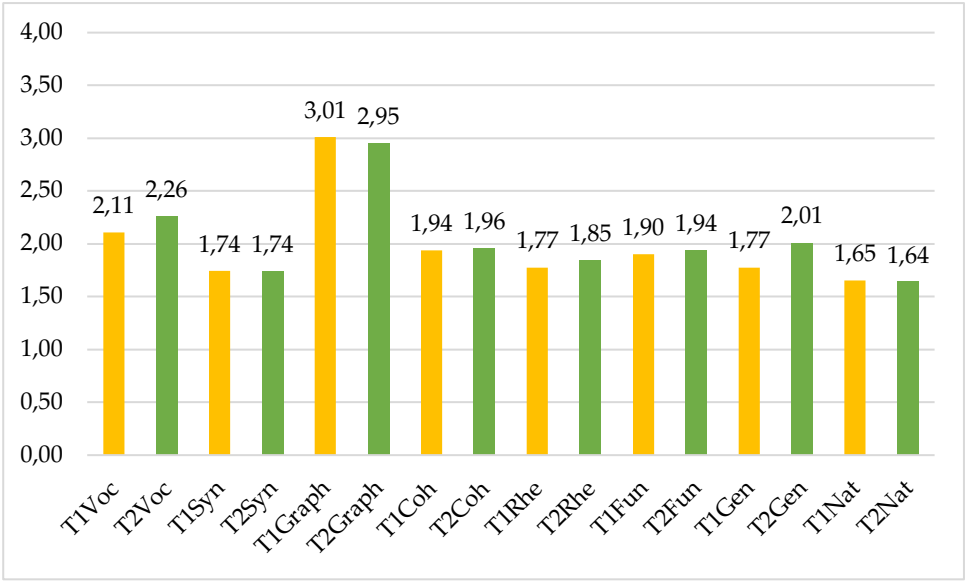

Figure 21. Writing test mean analytic scale marks for Task 1 (n = 179) and Task 2 (n = 152)

Furthermore, the highest average component mark on the writing test is the one in Graphology (*M* = 2.99), followed by Vocabulary (*M* = 2.21), while the lowest ones are in Syntax (*M* = 1.76) and Knowledge of Natural expressions (*M* = 1.68), as shown in Figure 22. An analysis of student performance in terms of both holistic and analytic marks is provided in the following section, addressing students' knowledge in each of the areas of language knowledge.

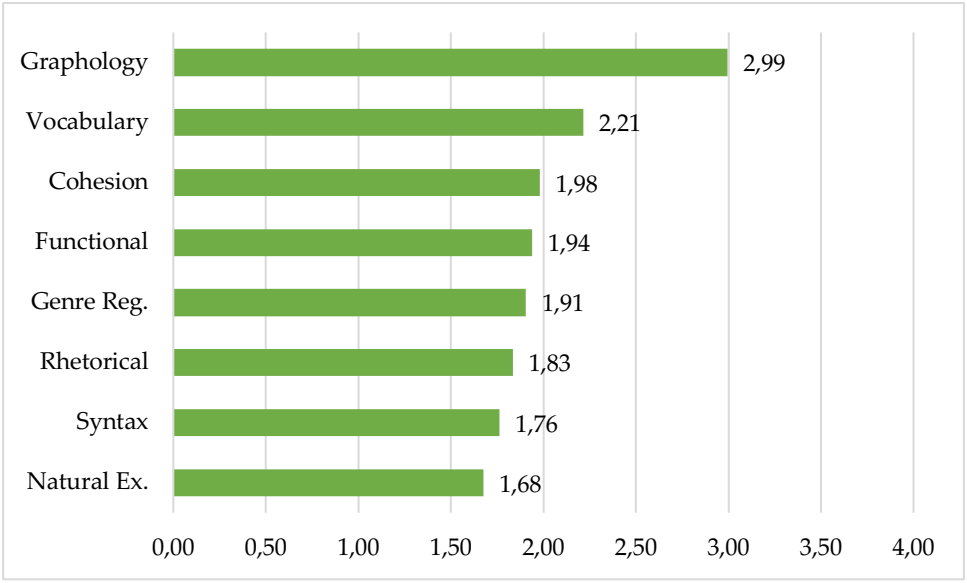

Figure 22. Writing test average mark per framework component.

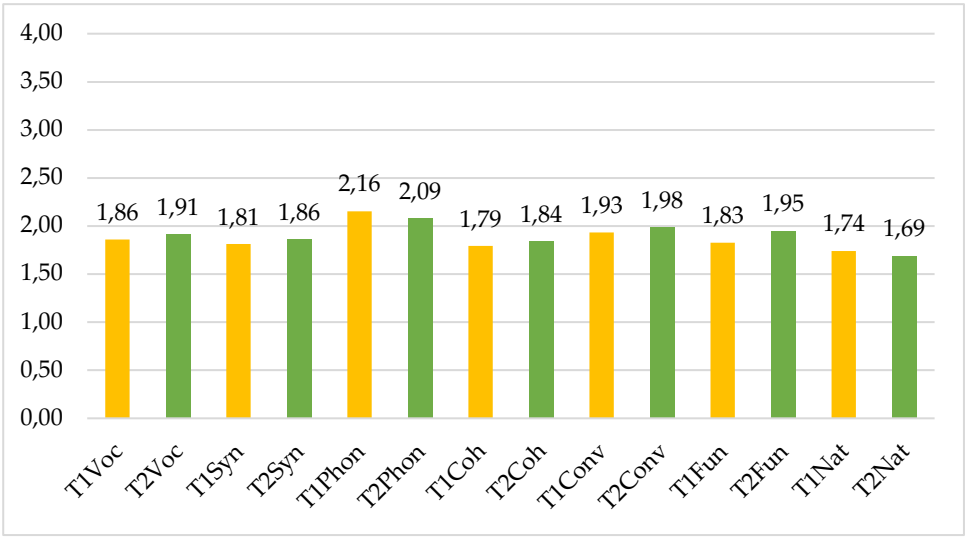

Figure 23. Speaking test mean analytic scale marks for Task 1 and (*n* = and Task 2 (*n* = 29).

The speaking test analytic marks are not much different from the writing test analytic marks, as shown in Figure 23. The highest mean mark is in phonology, that is pronunciation, where for Task 1 it is *M* = 2.16 (*SD* = .983), and *M* = 2.09 (*SD* = .955) for Task 2. The fact that students were easily understandable and that their pronunciation did not impede the communication may explain their marks in phonology. The lowest mark, in the knowledge of natural expressions, for Task 1 *M* = 1.74 (*SD* = .997), and *M* = 1.69 (*SD* = .870) for Task 2, may be due to their negative transfer from Italian and word for word translation of some phrases. This is also reflected in the average mean mark across the two tasks, as shown in Figure 24.

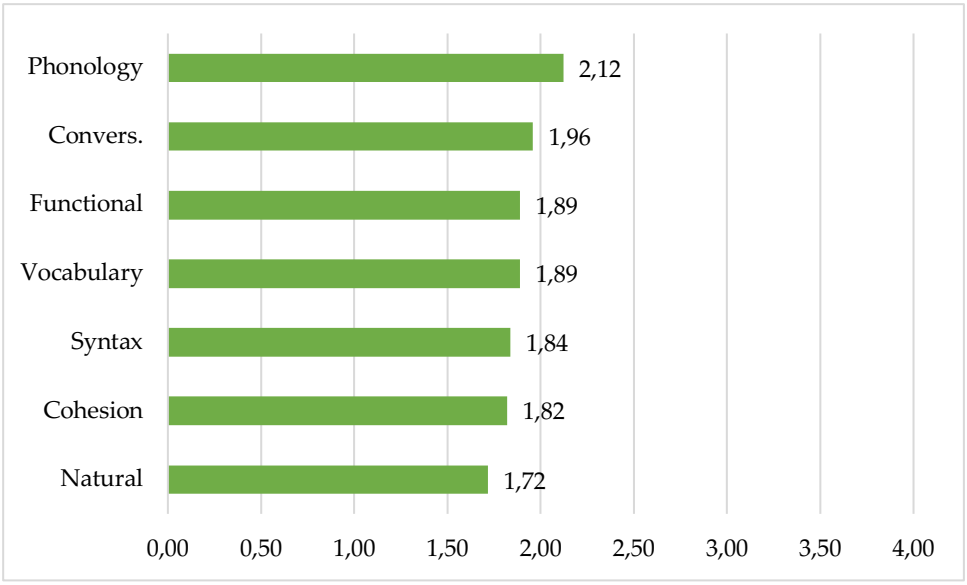

Figure 24. Speaking test mean average marks per framework component.

3.4. Analysis of Student Performance on the Writing Test

As outlined in Chapter One of Part Three, the test consisted of two written tasks: in Task 1 they were required to write an enquiry email and in Task 2 a blog entry. For the writing test, please see Appendix F.

The level of English expected from the students was CEF B2.

The tasks have been marked by two raters/examiners using the same holistic and analytic scales.

The holistic scale focuses on the task achievement, that is completion: to what extent the candidate has managed to complete the task considering all the individual language areas or components included in the analytic scales. A mark from 0 to 4 has been awarded to each of the tasks.

The Task 1 holistic mark is based on how likely the student would be to receive a response with all the information he or she required. Since an email is normally typed and not hand-written, the handwriting was not considered unless it interfered with the meaning or made understanding the text considerably more difficult. The following was considered:

- how many of the points the student mentioned and/or addressed,
- how the content was organized and whether it was easy to follow and read,
- the individual analytic scale areas of language knowledge or components, such as syntax, vocabulary, and graphology, were considered only to the extent to which they impeded with the meaning or message that the student tried to communicate,

Spelling was not considered for two reasons: in a real-life situation, students would normally type an email and would use a spell-checker. The second reason is that there were very few spelling mistakes across all students and these were penalized when marking the graphology with the analytic scale.

The Task 2 holistic mark is based on the degree to which students completed the task. Since the main purpose of a blog post is to arouse interest in a topic and provide a specific point of view that would be of interest to the readers, the mark was based on whether the student managed to achieve that. No specific format was requested for the post. Again, the individual analytic scale components were considered only if they interfered with the meaning or the message the student was trying to communicate.

The analytic scale covers a set of eight clearly defined assessment criteria based on Bachman and Palmer's model of language knowledge (Bachman and Palmer, 2010, p. 45): vocabulary, syntax, graphology, cohesion, rhetorical knowledge, functional knowledge, knowledge of genre and register, and natural expressions. A mark from 0 to 4 is awarded for each of the criteria.

Differences between Task 1 and Task 2 Achievement. There are some obvious differences in task achievement between the two tasks. The input for Task 1 was considerably longer and provided language for the students to rely on. Most of the students relied heavily on the input, using most if not all of the phrases from the input / instructions. Quite often, no additional information or details were added.

Holistic marks awarded for Task 1 are considerably higher - independent of the mastery of the areas of language knowledge, most of the students managed to communicate the message. This is most likely due to the specific instructions that the students relied on.

Holistic marks awarded for Task 2 are considerably lower. This is mostly due to the fact that very few students actually wrote a well-organized and convincing blog entry with original ideas and a specific point of view. The input contained very short specific instructions, and for that reason, the students needed to provide the content themselves.

Criteria. The analytic scales have been designed using Bachman and Palmer's model of language knowledge. The set of criteria included in the analytic scales are vocabulary, syntax, graphology, cohesion, rhetorical knowledge, functional knowledge, genre and register, and natural expressions. For each of the criteria, a brief definition or description is provided below.

Vocabulary: the body of words the students were expected to use, based on the task and the level of English knowledge (CEF B2).

Syntax: grammar and morphology appropriate to the level.

Graphology: spelling, punctuation, and capitalization.

Cohesion: marking relationships among sentences using, for example, connecting words such as *however, because*, etc.

Knowledge of rhetorical organization: appropriate sequencing of parts and units of written text.

Functional knowledge: understanding and producing meaningful relationships between sentences and intentions of the writer.

Genre and register: recognizing the type of written communication (email or article) and using the appropriate register that is, the appropriate level of formality.

Knowledge of natural expressions: the extent to which the text produced sounds natural and not only grammatically correct.

Analysis of student responses

A close analysis of the tests revealed several significant characteristics of the test takers.

General Observations: Most of the mistakes that the test takers made were a consequence of negative transfer from Italian into English. Odlin (1989, p. 27) defines transfer as "the influence resulting from similarities and differences between the target language and any other language that has been previously acquired." Therefore, negative transfer, or native language influence, can be defined as the use of native language structures and vocabulary where a student lacks the same in the target language.

Vocabulary: in Task 1, the students relied heavily on the input, and not much additional vocabulary is present apart from some exceptions where students added information that was not included in the instructions. Some of the vocabulary that Task 1 elicited was:

locations, accommodation, family, differences, abroad, organize, opportunity, range of courses, activities, attention, choose, attend, etc.

Most of the vocabulary was either appropriate to the level or lower level than the level required with only few instances of higher level vocabulary.

Some of the instances of students relying on the input are (Task 1):

the advertisement does not provide all the information I need… (1 student)

**How much it costs the study holiday? And what it includes?* (1 student)

…I want to spend two weeks in the UK. (1 student)

…but I don't know how much it costs and what it includes. (2 students)

I have just seen an advertisement for a study holiday in the UK… (4 students)

In a number of responses, the students failed to change the article "an" to "your" but simply copied the instructions.

In addition, a major part of the students did not understand the word "apply" (Task 1) and very few students actually asked for more information on how to apply for the summer holiday. The students that

did use the word either copied the phrase from the instructions word for word or used it incorrectly. Some of the examples are:

**So, I would like to know some information like how to apply.* (1 student)
**How can I apply my request?* (1 student)
**I do not know how apply…* (1 student)
**How I have to do for send you my inseriction.* (1 student)
**How to apply?* (1 student)
**…how to apply to this experience.* (1 student)
**How applies it?* (1 student)
**Then I want to know how to apply this study holiday.* (2 students)
**I would like to know what I have to do to send my apply.* (1 student)
**…I don't know how to apply this advertisement.* (1 student)
**I want to say how to apply.* (1 student)
**I need to know how the project apply.* (1 student)
**I don't know how to apply this study holiday.* (1 student)

Another evident problem in the area of vocabulary were the collocations. Some instances are:

Task 1:

*…I want <u>to do an experience</u>… ⟵ *fare l'esperienza*
*I want to <u>know some informations for</u> a study holiday. ⟵ *sapere delle informazioni*
*I need to <u>know more information</u> about it. ⟵ *sapere delle informazioni*
**…would like to do a study holiday…⟵ fare vacanza studio*
**I'd like <u>to do this experience</u>… ⟵ fare l'esperienza*
**I have <u>to do new friends</u>. ⟵ fare amici*
-**When I was fourteen I went <u>in</u> London. ⟵ a Londra*
**I <u>made the PET exam</u> in june 2011. ⟵ Ho fatto l'esame…*
**I studied English <u>to</u> high-school.*
**I'm a student <u>in</u> university…*

Task 2:

**…so we must <u>do attention</u>. ⟵ …dobbiamo fare l'attenzione.*
**But we have to <u>do a difference about</u>…*

Some other mistakes include "journal" referring to a "newspaper" (3 students) and *"quotidianity" referring to "daily life".

It is evident from the examples that some of the collocation mistakes are most likely a result of the negative transfer from Italian. The

remaining ones are most likely a result of the approach to language teaching and studying.

Below are some more instances of negative transfer.

**I want ? know what the offert includes…* ⟵ *offertà* in Italian

**I'm a university student, frequent the University…* ⟵ *frequentare l'Università* in Italian

**Now I frequent the University* ⟵ *frequentare l'Università* in Italian

*I believe in your *disponibility. (Student 16)* ⟵ *disponibilità* in Italian

*Thank you for the *disponibility. (Student 46, 43)* ⟵ *disponibilità* in Italian

*…and eventually cost for the *abitation. (Student 8)* ⟵ *abitazione* in Italian

*English is a fundamental *strument to do this. (Student 13)* ⟵ *strumento* in Italian. The word *"strument" or "instrument" referring to "a tool" was wrongly used by at least 10 students.

Grammar: The level of grammar was surprisingly low. Most of the students used only basic grammar structures such as simple sentences and the present simple tense. There were very few instances of more complex structures used correctly.

One of the most frequent mistakes is the noun *information* used in plural form, again as a result of negative transfer from Italian.

Other mistakes include the failure to use the question form where appropriate. Even when using the present simple tense, the students failed to use the auxiliary verb *do/does.* Where it was used, it was used incorrectly, together with the present simple third person singular *s.* Most of the students however, used inversion only, or simply a question mark at the end of the sentence, relying on the Italian language grammar forms.

Task 1:

**…the course will prepare me to the exam….?* ⟵ No inversion in Italian: *Il corso mi preparerà per l'esame?*

**…There is the possibility to live where the courses are?* ⟵ No inversion in Italian: *C'è la possibilità di…*

**What courses I can choose for my problem?* ⟵ No personal pronoun in Italian, hence no inversion: *Quali corsi posso scegliere…*

**You know a beautiful home where I can live…* ⟵ No auxiliary in Italian: *Conosce una casa carina dove posso abitare…?*

**How much it costs the study holiday? And what it includes?* ⟵ Literal translation from Italian into English in the first question; no auxiliary in

Italian in the second question: *Quanto costa la vacanza studio? E che cosa include?*

**Where I can eat?* ⟵ No personal pronoun in Italian, hence no inversion: *Dove posso mangiare?*

**How much it costs and what it includes?* ⟵ No auxiliary in Italian: *Quanto costa e che cosa include?*

**How much costs it?* ⟵ No auxiliary in Italian: *Quanto costa?*

**How much this holiday costs?* ⟵ No auxiliary in Italian.

**Plus, how many hours last the lessons?* ⟵ No auxiliary in Italian: *Quante ore durano le lezioni?*

**There is the possibility to study English culture too?* ⟵ *C'è la possibilità di studiare anche la cultura inglese?*

Task 2:

**Where we will arrive?*

Below are some examples of incorrect use of do/does:

Task 1:

**Do the price include other experiences?*

**Does is it possible choose locations?*

**How much does it costs?*

Task 2:

**This doesn't is a progress.*

There are also instances of omission of the auxiliary "do" in the negative form:

Task 1:

**I have not a great background knowledge…*

Task 2:

**But if we haven't a good capacity of…*

…person which not live with you in your city.

…of 'social' haven't nothing

The disadvantages are that the person haven't the contact and relationship…

Below are some examples of failure to use the Past Simple Tense:

Task 1:

**Three years ago I study English at private school.*

Task 2:

**…but long time ago, this struments don't existed.*

**Before a lot of thing are impossible…*

Some examples of failure to use Present Perfect Tense are:

Task 1:

**I don't know about my english current level, because I never do exams that certifies this.*

**I never go to UK, but I know some friend that they went to London for studying.*

**…because I never visit the city…*

**Because I never attended specific courses in .*

Task 2:

**…and he already underline the points…*

**…but we have lose the sense of …*

**There was an important development of things…*

**…and other devices are become a fundamental part of our homes.*

**In the last year, the new technologies are more and more improved and widened…*

**In the last years the technology is entered in our life.*

**…because many things are become simpler than the past…*

**In the last time… are increased a lot.*

**…and old friend that you don't see or meet from long time.*

**You can speak with people that you don't see of long time.*

On occasions, where the past simple tense and present perfect tense were used, they were used incorrectly.

**Did they include an exam of certification at the end? (Student 48)* Past Simple Tense used instead of Present Simple Tense;

**I saw your advertisement for a study holiday in the UK. (Student 22)* without an adverb of time;

**I've studied English at school for 12 years, I've followed a course in Rome at British Council for three years. (Student 14)* Present Perfect Tense used with an adverb of time.

**…but I didn't understand yet. (Student 15)* Past Simple Tense used with 'yet';

**I seen an advertisement… (Student 19)*

**I don't know the UK because there have never been. (Student 3)*

There were very few attempts to use the second conditional - they were however incorrect:

**…how much it cost if I would spend two weeks in the UK. (Student 49)*

**If I arrived with my friend or with my boyfriend we lived and study in the same city, about spent the same money?*

Several students failed to make distinction between *this* and *these*:
**…in this two weeks of my study holiday. (Student 7).*
Similar mistakes were made by five other students.

Age:
**I have 20 age.*
**I have 20 years old.*
**I have 22 years old.*

Additional Examples of Negative Transfer

Most of the errors are a consequence of negative transfer from Italian into English.

Task 1:

**I think that _ is very important have a total control of the situation. … for don't make mistake. ⟵ …è molto importante… per non fare errori*

**I want _ study for another level of English. ⟵ Voglio studiare per un'altro livello d'inglese.*

I want _ know what the offert includes… ⟵ Voglio sapere che include l'offertà.*

**I would like to know how much it costs this study holiday. ⟵ Vorrei sapere quanto costa questa vacanza studio.*

**I would _ to know the english language. ⟵ Vorrei conoscere l'inglese.*

**I wish _ know how much is this course…) ⟵ Vorrei sapere…*

**In fact, I need of a basic course… ⟵ Infatti, ho bisogno di un corso di base.*

**I want to do an holiday of 2 weeks. ⟵ Vorrei fare una vacanza studio di due settimane.*

**How much it costs the study holiday? And what it includes? ⟵ Quanto costa questa vacanza studio? E che cosa include?*

**I would like to come in one of your… ⟵ Vorrei venire in uno dei vostri…*

**How much it costs and what it includes? ⟵ Quanto costa e che include?*

**I would _ know also what we can pass the day. ⟵ Vorrei sapere come possiamo passage la giornata.*

**I have need about some information. ⟵ Ho bisogno delle informazioni.*

**And add to the hours of lessons, are also organised sport activities, excursions and visits…*

**…I am writing just for ask more information. ⟵ …per chiedere delle informazioni.*

I'd like study this language for increasing competences and ability in English.*

**In the specific,…*

**So I can to organize me for the bus's tickets. ⟵ …per poter organizarmi per…*

**I studied English during the period of the school… ⟵ Ho studiato l'inglese durante il periodo della scuola.*

**What is there to visit in the UK of very interesting? ⟵ Che c'è da visitare in Inghilterra di interesante?*

**I want to spend two weeks in the UK for to improve my level of English… ⟵ …per migliorare il mio livello d'inglese.*

**I think to spend two weeks… ⟵ …penso di passage due settimane…*

I'm born in Milan ⟵ Sono nata a Milano.*

Task 2:

**…we'd make it better to do others more important things.*

**All we know, … (Student 49) ⟵ Tutti noi sappiamo…*

**…for permit at the person … ⟵ …per permettere alla persona…*

**…so we don't go to the hospital for to do this.*

**So internet is important and don't exist today a life without it.*

**For not talking about…)*

**…people uses Facebook ⟵…la gente usa Facebook…*

**For example many children spend too time to surf the internet or on social networks instead than to read a book or go out with their friends.*

**Everythings have advantages and disadvantages…*

**…but is true that with these technologies…*

Graphology: There are no major problems in this area. The students made very few spelling mistakes, which may be due to the fact that the vocabulary had been learnt at school and not through other means such as media.

There are very few mistakes in punctuation and capitalization. As regards capitalization, the mistakes were all due to the fact that these words are not capitalized in Italian, for example the months, adjectives *English* and *Italian,* University, etc. Very few students failed to capitalize the personal pronoun 'I'.

Negative transfer in the area of graphology is also present. As a result of the fact that one of the most frequent problems of Italians when speaking English is that they do not manage to pronounce "h" at the beginning of the word but simply omit it, there were also examples of students using the indefinite article "an" instead of "a" before nouns beginning with an "h":

**…an house…*

**…an holiday…*

Cohesion and knowledge of rhetorical organization: The instructions for Task 1 were very clear and just by relying on the instructions, the students had a good chance at achieving appropriate rhetorical organization. However, in most of the responses, there was no evident planning nor paragraphing: information was mostly given or asked in the order in which it appeared in the instructions.

With regard to cohesion, in most cases, the relationships between sentences were clear although not often indicated by linkers or connecting words.

Functional knowledge: The functional knowledge that the students were expected to demonstrate includes the knowledge of ideational functions (description and explanation) and of manipulative functions (instrumental: requesting information and interpersonal: appropriate greetings).

The functional knowledge of appropriate greetings that the students were expected to demonstrate is discussed in the Genre and Register section). With regard to requesting information, describing and explaining, most of the students lacked the basic grammar to effectively use the functions in the target language.

Genre and Register: All the students recognized the genre, as it was clearly indicated in the instructions for Task 1, where it was assessed. The students however had some problems with the register and choosing and using an appropriate level of formality.

Task 1 required an appropriate degree of formality - even though it was not stated in the instructions, it was expected from the students as they did not know the person they were addressing. However, In Task 1, most of the responses started with *Hi* and *Hello,* which was considered inappropriate as it is too informal. There were some instances of *Good morning*, which was considered appropriate for the purpose of an email. There were very few instances of a completely appropriate level of formality - if an email started in an appropriate way, that is the opening greeting was appropriate, the closing one was inappropriate.

Opening greetings:
Hi, (15 students)
Hello (11 students)
Good morning (16 students)

Dear Your English Summer (1 student)
Dear Sir / Madam of… (1 student)

Closing greetings:
I wait your answer. (2 students)
Thanks for the attention. (2 students)
Thank you for attention (1 student)
Thank you for the attention (1 student)
Thank you (1 student)
Thanks for informations (1 student)
Thanks for the information (1 student)
Thanks for the informations, have a nice day (1 student)
Waiting yours answer (1 student)
Good bye (1 student)
Let me know. See you (1 student)
Best regards, thanks (1 student)
Your faithfully (1 student)
Sinceraly (1 student)
Thank you so much (1 student)
Thank you for your informations (1 student)
Best regards (1 student)
Best regard (1 student)
Thank you! (1 student)
Thanks you (1 student)
I waiting your information (1 student)
Hello (1 student)
Have a nice day (1 student)
Thanks, soon (1 student)
Thanks to all informations (1 student)
Bye (1 student)

There were a few instances of completely inappropriate closing greeting, such as:
Bye bye (3 students)
Thank you for the information, you are very dear. Bye bye (1 student)
and an opening one:
- Dear students (1 student)

Content: The main purpose of the email was to ask and get the information needed and most of the students completed the task at least

to a certain degree. There are also a couple of instances of completely inappropriate or off-topic content.

Some other responses included statements that were either irrelevant to the task or both irrelevant and inappropriate:

**My background knowledge of English is not so good, because the teachers of my school don't teach me anything.*

**I'm trying to work hard for earn enough money to make this experience fearlessly…*

**Before I want know who are you?*

**…but I have many economic problems… for the student come from Italy with economic problems the study holidays including economic help?*

**My knowledge of English is short.*

**Thank you for the information, you are very dear.*

Is there a place where I can sleep with other students?

Knowledge of natural expressions: A simple way to define the knowledge of natural expressions is the extent to which a piece of text or discourse sounds natural. It is evident from the student responses that this is a major problem of Italian students. Even though the content was easily understandable in most cases, it often sounded unnatural and at times even awkward. A most obvious explanation is that where a student was lacking grammatical structures to express themselves, they used the vocabulary they had and relied on Italian grammatical structures to create what they thought were meaningful sentences. Although these sentences are understandable in most cases, negative transfer from Italian is evident:

In your advertisement there is the possibility to do this experience two weeks….

I wish these informations, because to travell and to learn a good english is my great dream

In the end I want to know if finished this holiday study I'l recived an a certification of English's level because now I haven't it.

The gust of the person are amalgated.

Comments / Conclusion: It is evident from the instances of mistake types that most of them are the influence of negative transfer in the areas where the students were lacking in the target language. The mistakes appear in different areas: vocabulary, syntax and graphology and as a result they

influence the areas of cohesion, functional knowledge as well as the one of natural expressions: a majority of the responses sound unnatural.

In addition, most of the content was produced using basic grammar structures, for example, the Present Simple Tense, and simple sentences without appropriate connecting words. Furthermore, in the majority of responses there was no evident planning or content organization. Despite the low-level language, the majority of the students managed to communicate the message and achieve the task; for that reason, the holistic marks are on the whole higher than the individual criteria marks. To conclude, a real strength of the students is the ability to communicate the message with a limited vocabulary and/or syntax.

3.5. CEFR B2: An elusive goal

Keeping in mind that scales are based on the CEFR levels and descriptors, converted into CEFR levels (where 1 is CEFR A1 or lower, 1.25 – 2 CEFR A2, 2.25 – 3 CEFR B1 and 3.25 - 4 CEFR B2) and based on the average holistic mark across the two tasks of the 149 students that completed both tasks, the students' marks mostly fall under CEFR A2, 40%, while the level of English of 28% of the students in the sample demonstrated a CEFR B1 level, 7% CEFR B2 level and 25% A1 or lower.

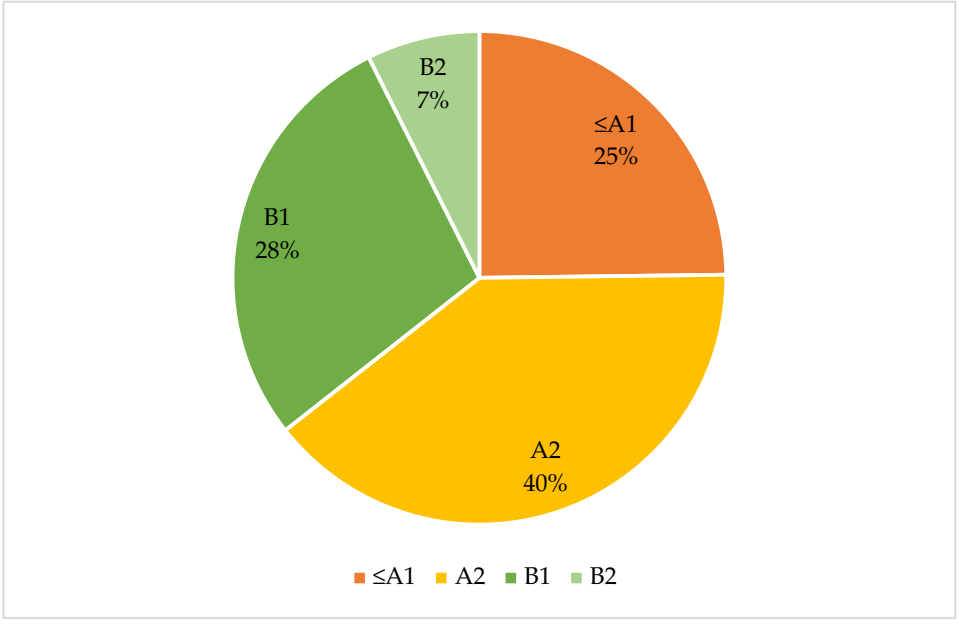

Figure 25. CEFR levels based on the average holistic Writing test mark.

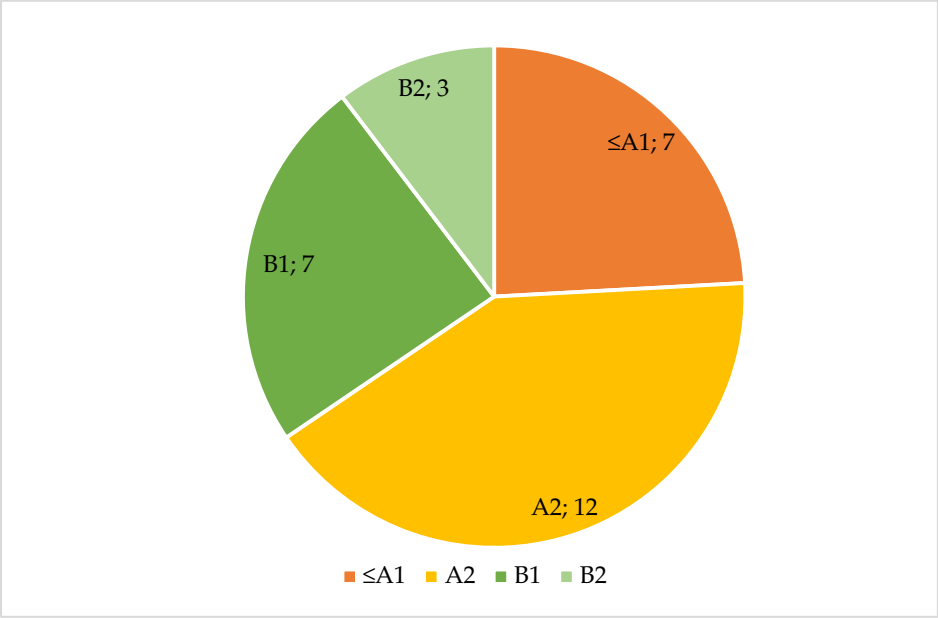

Figure 26. CEFR levels based on the average holistic Speaking test mark (n = 29).

The speaking test results are not much better either: again, most of the students' results are at a CEFR level A2, n = 12, while only n = 3 are at a CEFR B2 level.

When confronted to the students' self-assessment, these results tell us even more: students tend to either underestimate or overestimate their skills and do not know how much they do or do not know. As shown in the Table 45 and Table 46 below, where students' holistic marks are confronted with their self-assessment, some students who believe their knowledge is sufficient (n = 17) or moderate (n = 12) actually demonstrated an A1 or lower level of English. In the same way, some students underestimated their knowledge of English, thinking that their level of English is low (n = 4) or sufficient (n = 2) but demonstrated a CEFR level B1 and B1, respectively. Some students were nearly right, for example, n = 32, who evaluated their knowledge as sufficient and demonstrated a CEFR A2 level of knowledge.

			Bassa	Sufficiente	Discreta	Buona
CEFR Levels based on the average holistic Writing mark	A1 or less	Competenza scrittura	8	17	12	0
	A2	Competenza scrittura	12	32	14	1
	B1	Competenza scrittura	4	18	17	3
	B2	Competenza scrittura	0	2	4	5

Table 45. CEFR levels based on the average holistic Writing test mark and students' self-assessment.

Similar is true for the speaking test marks: some students who believe that their level of English is low (n = 3) demonstrated a CEFR B1 level of English, while some students who find their knowledge sufficient (n = 3) and moderate (n = 1), demonstrated a CEFR A1 or lower level of English. Finally, the students who communicated at a CEFR B2 level were all right about their speaking skills when they evaluated them as good.

			Bassa	Sufficiente	Discreta	Buona
CEFR Levels based on the average holistic Speaking mark	A1 or less	Competenza parlato	3	3	1	0
	A2	Competenza parlato	6	5	1	0
	B1	Competenza parlato	3	2	1	1
	B2	Competenza parlato	0	0	0	3

Table 46. CEFR levels based on the average holistic Speaking test mark and students' self-assessment.

3.6. Conclusion

From the analyses presented in the previous sections, we can conclude that CEFR B2 remains an elusive goal for the group of students who completed the test. Their level of knowledge in some areas of language knowledge is higher than in the others, for example graphology or phonology and vocabulary, as explained in the analysis section, while it remains at a low level for syntax and knowledge of natural expressions.

Considering that the knowledge syntax is necessary for producing grammatically correct sentences and the knowledge of natural expressions is what makes one sound natural in a foreign language, student responses quite often sound "Italian", awkward and inelegant. However, due to the knowledge of vocabulary, at a slightly higher level, students most often do manage to communicate the message despite the obvious difficulties and shortcomings.

The group of students who completed the test is quite heterogeneous. This is understandable considering that they come from different upper-secondary schools and have different background. This, however, shows us that, at least with the group that completed the tests, the objectives of the curriculum were not met. The fact remains that, when they reach the University, where their level is supposed to be CEFR B2, it is difficult, if not possible, to compensate for what they did not learn in secondary school.

Generalization across the set of tasks can be justified by means of the framework of language knowledge on which the analytic scales are based and the high correlation coefficient between the two. Similarly, reasons for extrapolation can be found in the concurrent validity and the fact that the language knowledge employed in the task performance would be the one needed for good performance in real life, due to the high task specificity and fidelity.

The fact remains that only a small number of students has reached the CEFR B2 level while for the others it remains an elusive goal.

Conclusion

Summary of findings

As outlined in the last section of the previous chapter, CEFR B2 has remained an elusive goal for most of the students who participated in this research. This is true for both the writing and the speaking test and skills.

Staying in contact with the language has shown to be significant: the students who have had experience with the English language got better average marks than the ones who have not. This is especially true if we consider how much Italians are or are not exposed to the English language in every-day life: all movies shown in cinemas in Italy are dubbed and not subtitled. This is also true for movies on TV, TV programs and shows. The only other potential exposure to the language would be through music or books, which is a matter of taste of each of the students and we do not know the extent to which they listen to British and American music and actually try and understand the lyrics, or read in English. Certainly, the Internet has helped get closer to the English-speaking world and texts of different kinds in English. Consequently, what is left is traveling to English-speaking countries or studying the language through because one understands that they will eventually need it in life or that it can be something they add to their curriculum.

We have also seen that having passed an internationally recognized exam in English does not necessarily mean that the students' productive skills that they would need in every-day life are at that level. This may be explained by the fact that most of the internationally recognized exams are mostly multiple-choice questions, gap-fills, and similar, while only one part of the writing test is free writing. These tests have a fixed format, which makes it easier for students to prepare for the exam. It is similar to the speaking test – students prepare for a specific format and have sample questions for which to prepare. This enables them to rehearse and even memorize some responses, which is often enough for a pass grade.

In the group of students who completed the test, the ones coming from the classical lyceum have a higher mean average mark than the ones coming from the linguistic lyceum. Although the numbers of students coming from these two types of school are not high (n = 35 from the classical lyceum, n = 15 from the linguistic lyceum), these results are difficult to explain. One foreign language is taught in classical lyceums, English, three hours a week for five years, while in linguistic lyceums,

English is one of the three languages taught, four hours a year for the first two years, and then three hours a week for the last three years. Despite the higher number of hours of English in linguistic lyceums, the students coming from the classical one are better at English, maybe because the students coming from the linguistic one study three languages while the ones of the classical lyceum focus on English only.

The students who completed the tests are quite good at some areas of language knowledge, such as vocabulary but so much at natural expressions. The knowledge of natural expressions is where these students lack – their utterances often sound foreign and inelegant, and so do their writings. This happens when one is not exposed to a language and only learns it through formal education. The good knowledge of vocabulary may be explained by the exposure to the English through social networks such as Facebook and Twitter, which are quite popular among the students in their twenties. This presupposition is also supported by the fact that their spelling is good, which implies that they may have learned their vocabulary through formal education or by reading in English, and not by listening to the language.

The students do not know how much they know. They tend to either underestimate or overestimate themselves. However, they do manage to communicate the message, especially in writing, despite their shortcomings, which is their most significant strength. This, of course, is valid for the types of tasks they completed as part of the exam, which is, every-day life, and where they are likely to need the English language in the future.

The fact remains that only a small number of the students who completed the tests are at a CEFR level B2, which means that the upper-secondary school English language curriculum objectives were not met. The reasons for this are certainly manifold and difficult if not impossible to ascertain. Some of the possible reasons may be the fact that the Ministry of Education does not provide a syllabus, only the level to be reached. Or the fact that the quality of teaching in a classroom of 20 to 30 students cannot be outstanding. The fact remains that, when these students reach the university, they do not meet the minimum requirements, which is most often the B2 level. Consequently, the groups at universities are quite heterogeneous and frequently numerous, and it becomes impossible to recuperate what was lost in upper secondary school. At the same time, it remains with the students for the whole life.

Advantages of the approach

The most evident beneficial traits of the approach are that it has potential washback effect in small-scale assessment, such as university context as well as that it allows for the identification of student strengths and weaknesses. Extended-production responses provide valuable detailed information about student knowledge. In the case of first-year university students of the Sapienza University, a significant weakness is the negative transfer from Italian and consequently appropriacy of some utterances. The fact that the weaknesses are easily identified with this approach would enable professors of English to focus on what the students are lacking.

Finally, the analytic approach based on analytic scales grants assessment of each of the model components or areas of language knowledge. A positive holistic approach to marking, based on Can Do statements, on the other hand, evaluates student knowledge based on what they can do not what they cannot do and prioritizes their strengths over their weaknesses. The use of the two types of scales together provides more information about the student knowledge than the use of a single scale or standardized tests.

Limitations of the research

Since the model used for the scale design and assessment is the Bachman and Palmer (2013) one, its analytical nature implies the assessment of each individual component of the model. The most obvious disadvantage of the model is that not each of the model components can be evaluated by a single task. With regard to the tasks administered, an inquiry email and a blog entry, designing appropriate descriptors for some of the components has proven to be a challenge. For example, online communication (an enquiry email) does not necessarily require a high level of formality. In the same way, there is no fixed format for blog entries, which again made the evaluation of genre and register difficult. It is evident that framework proposed by Bachman and Palmer is not universally applicable and that it requires certain modifications when operationalized, depending on the task to which it needs to be applied and the context in which it is employed.

Furthermore, unlike standardized language tests, with multiple-choice questions, where marking is done automatically, performance-based assessment requires an analytic evaluation of language knowledge.

This, of course, requires more time and assets such as trained raters, who need to go through a standardization process.

The test was administered at one of the university faculties, during regular lessons and was not mandatory for the students. For this reason, as well as due to the time constraints, it was impossible to administer the test with a larger sample.

Despite the approach disadvantages, such as cost-effectiveness and difficulties with the design of some of the descriptors for the analytic scales, the advantages of this kind of approach, especially the potential washback effect are quite significant. Whether the advantages outweigh the disadvantages would depend on a number of factors; however, in the small-scale assessment, this kind of approach is certainly feasible and beneficial.

Appendix A – Writing Test Analytic Rating Scales

ORGANIZATIONAL KNOWLEGE			
	Grammatical Knowledge		
	Vocabulary	Syntax	Graphology
0	Not enough to evaluate	Not enough to evaluate	Not enough to evaluate
1	Limited, a few words or phrases used correctly. Not enough to express himself/herself clearly.	Limited range of morphological and syntactic structures, most often incorrectly used and/or basic structures used correctly.	Frequent errors of spelling, punctuation and capitalization. Parts of the text impossible to understand.
2	Moderate vocabulary, mostly simple everyday English for basic communication, no topic specific vocabulary. Some more complex vocabulary, often incorrectly used.	Moderate range of structures, most often used correctly with occasional misses to mark agreement. Uses basic sentence patterns and phrases correctly.	Occasional errors of spelling, punctuation and capitalization spelling. Most of the text easy to understand.
3	Large vocabulary, most often appropriate to the level and the topic, expresses himself/herself clearly with only occasional errors when expressing more complex ideas.	Appropriate range of structures, with only occasional and not systematic errors in their accuracy. Occasional repetition or difficulty with formulation.	Few non-systematic errors of spelling, punctuation and capitalization.
4	Extensive vocabulary, always uses appropriate word and does it accurately. Few mistakes that do not hinder communication.	Extensive range of structures, always correct, including the complex ones appropriate to the level. No structures causing misunderstanding.	Excellent mastery of conventions. No errors of spelling, punctuation and capitalization.

N.B. 0 (not enough to evaluate): the learner did not produce enough to be evaluated or the response was inappropriate to the task, that is off-topic.

ORGANIZATIONAL KNOWLEDGE		
	Textual Knowledge	
	Cohesion	Rhetorical Knowledge
0	Not enough to evaluate	Not enough to evaluate
1	Little cohesion. Relationships between sentences not marked or few attempts to mark them with very basic connectors such as “and”, “so”, “then”.	Little rhetorical knowledge. Little evidence of planning and organization.
2	Moderate cohesion. Relationships between sentences generally marked but not always clear or clear but not appropriately marked. Uses simple connectors to link simple sentences.	Moderate rhetorical knowledge, some evidence of planning and organization, relatively clear sequencing of text parts.
3	Appropriate cohesion; relationships between sentences always marked, only few misses. Uses a series of linking words and/or cohesive devices to indicate relationships.	Appropriate rhetorical knowledge, evidence of planning and organization, clear sequencing of text parts.
4	Excellent cohesion; a variety of linking devices used correctly. A variety of linking words and cohesive devices used correctly and efficiently.	Extensive rhetorical knowledge showing unity; strong organization appropriate to the content.

Appendix A – Writing Test Analytic Scales

	PRAGMATIC KNOWLEDGE		
	Functional Knowledge	Sociolinguistic Knowledge	
		Genre and Register	Natural/Idiomatic Expressions; Cultural References and Figures of Speech
0	Not enough to evaluate.	Not enough to evaluate.	Not enough to evaluate.
1	No evident knowledge of expected functions, few attempted but inappropriate. Correct use of only basic social forms such as greetings (Task 1).	No evident knowledge of the genre, evidence of only one, inappropriate (informal) register.	Mostly unnatural expressions, no idiomatic expressions and or figures of speech.
2	Some but at times inappropriate functional language. Use of the interpersonal functions (greetings) and simple manipulative functions (requests) (Task 1). Use of basic ideational functions such as descriptions. (Task 2).	Recognizes the genre, evidence of both registers but not much control.	Some natural expressions, idiomatic expressions or figures of speech.
3	Evident knowledge of both ideational (descriptions, explanations) (Task 2) and manipulative (instrumental: requests, suggestions; interpersonal: greetings) (Task 1& Task 2). Occasional inappropriacy or lack of control.	Evident knowledge of the genre, evidence of the appropriate register (formal T1, semi-formal T2) and moderate control.	Language mostly natural; idiomatic expressions or figures of speech present.
4	Excellent knowledge of both ideational (descriptions, explanations) (Task 2) and manipulative (instrumental: requests, suggestions; interpersonal: greetings) (Task 1& Task 2). Well-controlled and appropriate to express himself/herself in a polite manner (Task 1).	Excellent knowledge of the genre, well-controlled correct register (formal T1, semi-formal T2).	Language completely natural; appropriate use of idiomatic expressions or figures of speech.

N.B. 0 (not enough to evaluate): the learner did not produce enough to be evaluated or the response was inappropriate to the task, that is off-topic.

Appendix B – Writing Test Holistic Rating Scales

	TASK 1 HOLISTIC SCALE
0	Not enough to evaluate; Almost no content or content completely inadequate for the task or too confusing and chaotic; difficult if not impossible to understand due to low level grammar. Would not receive a response to the email.
1	None or only one of the points addressed; few points mentioned but not addressed. Major gaps in communicating the message. May receive a response to the email but would not get the information he/she needs. Possible irrelevant information.
2	Content present but obvious problems in communicating the message. Only some of the points mentioned and addressed; all points mentioned but only some addressed. Possible irrelevant and/or redundant information.
3	Most of the content relevant and adequate. All points mentioned and most of them addressed. Communicates most of what is required but there are some gaps.
4	Relevant and adequate content. All requested points addressed. Successfully and with ease communicates the message despite some grammar points acceptable at this level.

Appendix B – Writing Test Holistic Scales

	TASK 2 HOLISTIC SCALE
0	Not enough to evaluate; Almost no content or content completely inadequate for the task or too confusing and chaotic; difficult if not impossible to understand due to low level grammar.
1	Some appropriate content present. None or only one of the points addressed (among advantages and disadvantages); points mentioned but only one addressed. No evident organization, no coherence). Frequent errors in all areas. Major gaps in communicating the message.
2	Only some of the points mentioned and addressed (among advantages and disadvantages). Mostly poorly organized and not coherent; or well somewhat organized but short and not developed. Gaps in communicating the message due to low level grammar although there may be some occasional complex structures and vocabulary. Possible irrelevant and redundant information.
3	Most of the content relevant and adequate. All points mentioned and some addressed (among advantages and disadvantages). Clear organization; coherent. Communicates most of what is required but there are some gaps or redundant information.
4	Relevant and adequate content. All requested points addressed. Excellent organization and supporting ideas and arguments of all parts. Successfully and with ease communicates the message. Few if any errors.

N.B. 0 (not enough to evaluate): the learner did not produce enough to be evaluated or the response was inappropriate to the task, that is off-topic.

Appendix C – Speaking Test Analytic Rating Scales

ORGANIZATIONAL KNOWLEGE			
	Grammatical Knowledge		
	Vocabulary	Syntax	Phonology
0	Not enough to evaluate	Not enough to evaluate	Not enough to evaluate
1	Limited, a few words or phrases used correctly. Not enough to express himself/herself clearly.	Limited range of morphological and syntactic structures, most often incorrectly used and/or basic structures used correctly.	A clear pronunciation of only few basic words. Phrases can be understood with effort.
2	Moderate vocabulary, mostly simple everyday English for basic communication, no topic specific vocabulary. Frequent errors.	Moderate range of structures, most often used correctly with occasional misses to mark agreement. Uses basic sentence patterns and phrases.	A noticeable foreign accent though generally understandable.
3	Large vocabulary, most often appropriate to the level and the topic, expresses himself/herself clearly with only occasional errors when expressing more complex ideas.	Appropriate range of structures, with only occasional and not systematic errors in their accuracy. Occasional repetition or difficulty with formulation.	Most often a clear pronunciation with occasional foreign accent or mispronunciation.
4	Extensive vocabulary, always uses appropriate word and does it accurately. Few mistakes that do not hinder communication.	Extensive range of structures, always correct, including the complex ones appropriate to the level. No structures causing misunderstanding.	A clear and natural pronunciation and intonation. Does not impose strain on the interlocutor.

Appendix C – Speaking Test Analytic Rating Scales

ORGANIZATIONAL KNOWLEDGE		
	Textual Knowledge	
	Cohesion	Conversational Knowledge
0	Not enough to evaluate	Not enough to evaluate
1	Little cohesion. Relationships between sentences not marked or few attempts to mark them with very basic connectors such as “and”, “so”, “then”.	Little conversational knowledge, evidence of knowledge of basic and everyday expressions, no evidence of planning or organization. Understands only if addressed in clear, slow speech.
2	Moderate cohesion. Relationships between sentences generally marked but not always clear or clear but not appropriately marked. Uses simple connectors to link simple sentences.	Moderate conversational knowledge, some evidence of planning and organization, clear sequencing, delivers short turns or longer ones with some pauses and hesitation.
3	Appropriate cohesion; relationships between sentences always marked, only few misses. Uses a series of linking words and/or cohesive devices to indicate relationships.	Appropriate conversational knowledge, evidence of planning and organization, clear sequencing; maintains conversation although may be difficult to follow at moments.
4	Excellent cohesion; a variety of linking devices used correctly. A variety of linking words and cohesive devices used correctly and efficiently.	Extensive conversational knowledge showing planning and strong organization; easily participates in the conversation.

N.B. 0 (not enough to evaluate): the learner did not produce enough to be evaluated or the response was inappropriate to the task, that is off-topic.

	PRAGMATIC KNOWLEDGE		
	Functional Knowledge	Sociolinguistic Knowledge	
		Genre and Register	Natural/Idiomatic Expressions; Cultural References and Figures of Speech
0	Not enough to evaluate.	Not enough to evaluate.	Not enough to evaluate.
1	No evident knowledge of expected functions, few attempted but inappropriate. Correct use of only basic social forms such as greetings.	No evident knowledge of the genre, evidence of only one, inappropriate* register (too formal or too informal).	Mostly unnatural expressions, no idiomatic expressions and or figures of speech.
2	Some but at times inappropriate functional language. Use of the interpersonal functions (greetings) and simple manipulative functions (requests).	Recognizes the genre, evidence of both registers but not much control.	Some natural expressions, idiomatic expressions or figures of speech.
3	Evident knowledge of both ideational (descriptions, explanations) and manipulative (instrumental: requests, suggestions; interpersonal: greetings). Occasional inappropriacy or lack of control.	Evident knowledge of the genre, evidence of the appropriate register (informal and semi-formal) and moderate control.	Language mostly natural; idiomatic expressions or figures of speech present.
4	Excellent knowledge of both ideational (descriptions, explanations) and manipulative (instrumental: requests, suggestions; interpersonal: greetings). Well-controlled and appropriate to express himself/herself in a polite manner.	Excellent knowledge of the genre, well-controlled correct register (informal and semi-formal).	Language completely natural; appropriate use of idiomatic expressions or figures of speech.

** Role-plays 1 and 2: appropriate register informal; role-plays 3-5: appropriate register semi-formal*

N.B. 0 (not enough to assess): the learner did not produce enough to be assessed or the response was inappropriate to the task, that is off-topic.

Appendix D – Speaking Test Holistic Rating Scales

	ROLE-PLAY HOLISTIC SCALE
0	Not enough to assess. No awareness of the role-play whatsoever or the content communicated is inappropriate to the task.
1	Little awareness of the role-play. Limited communication: one-word or very simple questions and answers. Needs to be addressed slowly and carefully. No planning or no evident purpose of what is communicated. The role would be achieved thanks to the intervention or involvement of the interlocutor.
2	Moderate awareness of the role-play. Manages to ask and answer simple questions correctly. Little effort to understand the speaker. Purpose is clear but there are occasional difficulties due to inability to express himself/herself clearly. The role would be achieved with some help by the interlocutor.
3	Aware of the role-play and the role. Exchanges information with ease, asks/gives for more detailed information and gives clear instructions though occasionally pauses in search for the appropriate expression. Purpose is clear and generally developed.
4	Complete awareness of the role-play and the role. Expresses himself/herself with ease, asking and giving detailed information and clear instructions. Purpose is clear and well developed and sustained throughout the role-play.

N.B. 0 (not enough to assess): the learner did not produce enough to be assessed or the response was inappropriate to the task, that is off-topic.

Appendix E – Student Questionnaire

QUESTIONARIO STUDENTE

In questo questionario ci sono alcune domande su di te. Devi compilarlo in ogni sua parte. Non ci sono risposte giuste o sbagliate. Grazie per la collaborazione.

Nome e Cognome _____

Quanti anni hai? _____

In che paese sei nato/a?
- Italia
- un paese europeo
- un paese extraeuropeo

Quale scuola secondaria di secondo grado hai frequentato?
- liceo (scrivi che tipo di liceo hai frequentato) _____
- istituto tecnico
- istituto professionale

Quale lingua/e parli a casa? _____

Quale lingua/e hai studiato oltre l'inglese? _____

Che voto hai ottenuto in inglese alla fine del primo quadrimestre dell'ultimo anno? _____

Da 1 (bassa) a 4 (buona) come giudichi la tua competenza in lingua inglese?
1 - bassa
2 - sufficiente
3 - discreta
4 – buona

Da 1 (bassa) a 4 (buona) come giudichi la tua competenza in inglese nelle seguenti abilità:

	1 - bassa	2 - sufficiente	3 - discreta	4 - buona
ascolto				
parlato				
lettura				
scrittura				

Hai una certificazione di lingua inglese?

a) Sì
b) No

Quale e che livello? _____

Hai mai fatto vacanze studio in Inghilterra o un altro paese dove si parla inglese?

a) Sì
b) No

Hai mai frequentato un corso di inglese fuori dalla scuola?

a) Sì
b) No

Hai sostenuto l'idoneità d'inglese?

a) Sì
b) No

Appendix F – Writing Test

TASK 1

You have just seen an advertisement for a study holiday in the UK. However, it does not provide all the information you need. Write an email asking for the missing information. Consider the following:

- you don't know how to apply,
- you want to spend two weeks in the UK,
- you don't know how much it costs or what it includes,
- any other details that you want to add.
-

The information you need to provide is:

- who you are, what you do,
- your background knowledge of English and your current level.

You have 35 minutes to write about 200 words.

Your English Summer

study holidays in the UK

- a wide range of courses to choose from,
- for students of all ages,
- length: 2, 4, 6 or 8 weeks,
- superb centres in great locations ,

such as Cambridge, Brighton, Stratford and London.

Get in touch: info@yourenglishsummer.co.uk

TASK 2

You have been asked to write a post, to be published on the University blog, on the advantages and disadvantages of new technologies such as smart phones, internet, social networks, etc.

You have 45 minutes to write about 300 words.

Appendix G – Speaking Test Role-plays

Role-play 1: "In a restaurant"
Topic: Services - Travel - Food and Drink

Student role-card:
You are in a restaurant and you need to order your dinner. Have a look at the menu, choose what you would like to eat and order. Consider the following:

- You are vegetarian.
- You have no cash.

The examiner starts the role-play with: *Good evening. Welcome to Savannah. My name is Jane and I will be your waiter tonight.*

Role-play 2: "Socializing"
Topic: Daily life – Relations with other people

Student role-play:
You are in a foreign country and you have just met a person. You would like to find out more about the person. Ask questions to get the following information:

- About their age and their family,
- About their hobbies,
- What they do in life.

Role-play 3: "At the train station"
Topic: Travel – Daily life

Student role-card:
You are traveling with two friends from London to Paris. You are at the train station info desk and you need the following information before you buy tickets:

- The train timetable,
- The cost of the tickets.

Role-play 4: “Giving directions”
Topic: Daily life

Student role-card:
You are in your home town. A foreigner stops you on the street and asks you hot to get to the supermarket. Tell him/her:

- How to get there.
- How long it will take.

Role-play 5: “At the airport”
Topic: Traveling

Student role-card:
You have just landed at the New York airport. The immigration officer needs to know the following:

- The reason for your visit.
- How long you are staying.
- Whether you are staying in a hotel or somewhere else.

Appendix H – Speaking Test Student Responses

1. Student 1

Understands the interlocutor and responds using one-word or one-phrase answers. Both tasks completed with the help of the interlocutor.

Longer stretches of language are:

Role-play "At the airport"

I want to visit some monuments, some places. I want to see new places and go to Aquarium.

Thank you very much.

I'm in holiday - repeated twice.

Role-play "In a restaurant"

Yes, I'm vegetarian. But I can eat fish.

But I have the card, I don't have cash.

Can I have a scallops and zucchini and onions.

2. Student 2

Seems to understand the questions and responds with vary basic grammar and a lot of hesitation. Very difficult to follow, would be understood by a very patient and sympathetic native speaker. Considering that the purpose of the role-play is small talk and socializing, completion of such a task would imply considerable effort on the side of the interlocutor. Strong Italian accent impedes understanding.

Role-play "At the train station"

What you do information for the travel.

Two-ticket.

Role-play "At the airport"

Have … information for … (inaudible)*… with my parents…, my family and … what to do … travel with the, country with my family.*

animal, airplane, age, pronounced with a strong Italian accent, that is an "h" at the beginning of the word: [ˈhænɪməl], [ˈhɛrˌpleɪn], [eɪʤ].

my sister 29 and our mother the baby…

her have a mouse in the house…

I dunno work… I have 23 year.

3. Student 4

Strong Italian accent. Hesitation and long pauses. When asked questions has problems answering. Manages to answer with one-word or one-phrase answers. Whole sentences that the student produces are very basic in terms of syntax and vocabulary. As the second role-play the student is supposed to give instructions, it is obvious that the task would not be completed.

Role-play "Socializing"
I'm 19 years old.
You study …
I have one brother and one sister.

Role-play 2 "Giving directions"
The supermarket is in front of the street. And it is…

4. Student 8

Manages to communicate using very basic syntax and vocabulary. Mistakes in vocabulary. Still puts in effort to communicate the message. In role-play 2, small-talk, goes straight to the point, lacks structures to express herself.

Role-play "Giving Directions"
Yes, you go to the street and then you ……. get right and then get left and on right there's supermarket.

Role-play 2 "Socializing"
You have brothers or sisters?
I have three sister and two brothers.
I like read book. I like walk in a park.
What do you do in your life?
I like children and I am very active when I stay with them.

5. Student 10

Understands most of the examiner's questions though some get misunderstood. Most often responds with one-word answers or simply "Yes". Long pauses, strong Italian accent.

Role-play "At the airport"
I visit New York because I'd like New York and I visit, visit it.

Role-play 2 "In the restaurant"
I'd take the vegetables.
I'd pay to credit card.

6. Student 11

Misses the most basic vocabulary. Gets more confident and conversational towards the end of the role-play.

Role-play "At the train station"
I want to go to Paris.
Can you repeat please?

Role-play "Giving directions"

You cross market and turn right and go … (inaudible). *You try a coop, that is a supermarket and you would buy everything.*

7. Student 12

Takes initiative and is aware of the role-play and role. Asks questions to complete the task. Purpose is clear. Some difficulties with vocabulary and at times difficulty to express himself. Somewhat confused when giving instructions.

Role-play "Giving directions"
There is one at five meters that way. Is small but efficient.

Role-play "At the train station"
I can ask a question?
Where is it a train for Paris?

Where is the time, … train timetable?
For the twelve.
Thank you so much.

8. Student 14

Communicates quite efficiently. Has occasional problems with pronunciation.

Role-play "At the airport"
I came in the US because I will visit. For holiday.
I stay for three months.
She work in Burger King.
I stay in hotel.

Role-play "In a restaurant"
I am ...
I will pay with credit card.
I would like soft drinks please.
I would like mushroom.

9. Student 15

Communicates fluently and efficiently with occasional non-native like language. Completely aware of the role-plays.

Role-play "At the airport"
Happy to be here.
I plan to stay for in New York one week, then I will come back to Miami and spend another week there.
Yes, already have it, it's for 27th.
It's a pleasure, thank you.

Role-play "In a restaurant"
It's for two, 7pm.
Red wine would be perfect for us.
Can I have the menu?
Do you have any vegetarian menu?
One zucchini and onion soup will be perfect.
Maybe later. If you have tiramisu that would be perfect.

Can I pay with a card because I don't have cash.
Do you accept visa?

10. Student 17

Aware of the role-play and puts some effort into the role. Communicates with some ease though with frequent errors in grammar.

Role-play "Socializing"
I'm 19 years old.
Yes, I have one dogs.
I like swim. I like read book. I like x spend my week with my family and with my boyfriend, my friends.
Yes, but I prefer books.
What do you do in your life?
I study and I work. I work in a xxx. And I study…
I work only two day in a weeks.

Role-play "Giving directions"
Yes, there is a supermarket.
Is it near the bookshop.
…and turn left and you are arrived.

11. Student 19

Quite aware of the role-play, puts in some effort but lacks the vocabulary and grammar to express herself.

Role-play "At the train station"
I need some information of the timetable of the train from London to Paris.
What time the train from London to Paris.
And how much this train… the ticket.
Three tickets.

Role-play "Giving directions"
The supermarket is near my home.
50 minutes on foot to the supermarket.
I turn on left and I go right and I stay at the supermarket.

12. Student 20

Aware of the role-play but lacks the language needed. Puts in some effort. Tasks achieved with help of the interlocutor.

Role-play "Giving directions"
The supermarket is near the hotel. Cross the street and turn left to Via … and go…
Straight and turn right to a Starbucks and the supermarket is here.
It's a lot long. 20 minutes.

Role-play "Socializing"
Nice to meet you.
What's your age.
I am 22 years.
I'm single son.
Where is your hobbies?
I love music and like play the piano and I love swim.
One day of week.
What do you do in the life.
I go to the school.

13. Student 21

The role completed thanks to the interlocutor to some part. The student is aware of the role and puts in some effort. Misunderstands the questions in the second role-play.

Role-play "At the airport"
The reason for my visit is to…
I would like to visit New York because is a big city and … and to … learn and speak English.
I'm stay in New York for three weeks and…
In hotel.

Role-play "In a restaurant"
And I would like some salad.
I want to drink tea.

14. Student 23

Quite strong Italian accent. The student misinterprets the task in the first role-play and produces only few phrases although puts in some effort into the role.

Role-play "At the train station"
I…. ask [ˈhɑːsk] *timetable that the timetable of the train.*

Role-play "Giving directions"
Yes, we are a supermarket. You go to the … right the street before the…

15. Student 24

Some questions not understood. Too direct in the socializing role-play.

Role-play "Giving directions"
Yes, is near.
In this street turn to the left and this supermarket.

Role-play "Socializing"
What's your age? How old are you?
And your family?
I'm sister, 2 cats, 10 tortoise, mother and father.
What are your hobbies?
My favorite hobby is sing and play piano.
I think yes.
I sing from when I was a child.
What they do in life?
I study and sing.

16. Student 26

No functional knowledge (asking for information, greetings). Some syntactically correct but inappropriate utterances. Little awareness of the role. Has problems understanding the questions. Not much produced.

Role-play "At the train station"
Where is the timetable of the train?
For the train depends.

All the time.
How much it cost, the ticket cost?

Role-play "Giving directions"
You need to go … to your right. Go and … all the street and at the end you find it.

17. Student 28

The student understands only if asked slowly using very simple structures. Responds with one-word answers. The task would be achieved only with plenty of help by the interlocutor.

Role-play "At the train station"
The… Some information about the … where can I buy the tickets to Paris.
Three tickets.
To Paris.

Role-play "Giving directions"
Yes, how long xxx.
Four miles.
You take the left.
About ten minutes.

18. Student 31

The student is hesitant while responding, seems to be searching for the appropriate language. Produces long stretches of language.

Role-play "At the airport"
I'm flying from Rome.
Yes, of course.
I was born in Ukraine.
I'm living in Rome actually.
Try to have a better life.
I hope to find xxx work here.
I have some friends here. I hope they will have me.

Role-play "In a restaurant"

I'm alone right now.
I have also some specific requests.
The fact is that I'm vegetarian.
Maybe something like salad.
Just a question…

19. Student 33

Aware of the role and puts in effort though there are frequent errors. Maintains the conversation. Has a strong accent but is understandable.

Role-play "At the train station"
I need to know at what time do the train from London to Paris leave.
Can you do a ticket for this afternoon?
And how much this ticket? How much do this ticket cost?
No, three ticket, for me and two my friends.
Thank you.

Role-play "Socializing"
Hello, how are you?
How old are you?
Thank you.
Do you live here, in this country?
Do you live with your family?
I have two daughter also. One, the older is 31 and the second the younger is 25.
I'm only.
What do you do here, what your hobby?
Your job, what is your job?
When don't you work, what do you do?
Do you play tennis?
I don't like sports, I like chair.

20. Student 44

Produces long stretches of language though with some errors and hesitantly. Aware of the role-play although does not take initiative.

Role-play "At the airport"
I'm from Italy, from Rome.

Yes, I would like to stay in the USA because I'd like to visit the country and because I've got some friends and so I would like to stay here for holidays.
10 days.
I will stay in their apartment.
The address is…
Her name is Sara Rossi.

Role-play "In a restaurant"
No, I haven't.
Yes, it's okay.
Yes, I would like to drink tea.
Only sugar.
Yes, I am vegetarian.
I just want vegetables.
No, no, no, rice is okay.
Onion, I don't like onion.
I have got a credit card.

21. Student 48

The student is quite fluent and participates in the conversation with ease. Does not take much initiative though.

Role-play "At the airport"
Good morning.
Yes, sure.
It's a nice place.
I've been here once, two years ago.
I think I'm gonna stay two or thee weeks.

Role-play "In a restaurant"
I think I don't.
You don't have no table…
I don't have cash, do you accept credit card.
That would be perfect.
What about side dishes?
I'd like zucchini and onions.
Maybe as a starter, it's better.
Can I order a beer while waiting, maybe draft?

22. Student 52

Communicates with some difficulties and help by the examiner. Occasionally difficult to understand.

Role-play "At the airport"
Good morning.
Yes, you can.
It's very nice, it's a beautiful city.
I … two days ago so I stay here for informations.
Two or three weeks.
Yes, I have a …
About two weeks.

Role-play "In a restaurant"
I don't have a reservation.
Only me.
No any special requests.
I would … a tea.
Lemon, please.
To eat a scallops and what is house special?
One of these and a salad.
I don't have cash.
Okay, thank you.

23. Student 53

Has some difficulties understanding the examiner. Very basic expressions. Would achieve the role thanks to the interlocutor only.

Role-play "At the airport"
Thank you.
For studying. Psychology in a university.
Two months.
At home of university.
What his name?

Role-play "In a restaurant"
Only me.
A bottle of water.

No bubbles yes.
I'm vegetarian.
Vegetables and rice.
And a salad.
And carrot in the salad.
What do you have?
Ok, I try.
Can I pay with the card.

24. Student 57

Communicates with ease though does not take much initiative. Produces longer stretches of language with few mistakes. Missed the part of the role where she was supposed to ask to pay with a credit card.

Role-play "At the airport"
I came here because my parents live here and I want to visit them.
Yes, they… My mother doesn't work but my father works in an office.
Yes, in their house.
For 2 weeks.
I think that I am going to visit the town and the most beautiful place there are in this town.
I don't know anyone, just my parents.

Role-play "In a restaurant"
Two peoples.
I'd like some salad and rice.
No, thank you.
With cash.

25. Student 59

Misunderstands a question, hesitates a lot. It would take a very sympathetic native speaker. Puts effort though.

Role-play "At the airport"
At a city … south of Italy.
My origins is in … at a city of south of Italy.
Because I would take a course, a course of English.
No, I stay with … in … with my parents in your home.

And your boyfriends.
They work.
Three weeks.
I would visit…

Role-play "In a restaurant
I am vegetarian.
Vegetables. Zucchini and onion.
Sweet carrots.
Just water.
I have a card.

26. Student 61

Very hesitant. Only one-word answers. Often does not understand the question. Orders chicken although should be vegetarian.

Role-play "At the airport"
Yes. (in reply to Welcome to New York*)*
In a city… a little town.
I visit … New York and work, … English.
Two weeks.
Yes, two friends.
I have a room beautiful and I can see…

Role-play "In a restaurant"
Hello.
I have a reservation.
Thank you.
I have… I are… I'm vegetarian and I have not cash and pay with a card.
I have vegetables and I…
And steak and chicken.

27. Student 63

Communicates with easy. Makes some grammar mistakes in longer stretches of language. Takes initiative and puts in effort.

Role-play "At the airport"
Thank you very much.

I come from Rome in Italy.

I'm here for study for few months. About like 6 months. I will stay here in apartment in New York.

No, because I don't know yet if I want stay other months. Maybe New York or other places in US. To improve my English. I've been studying it for a long time.

I come from Sapienza and I am going to Columbia University.

I received some information from the University.

That kind of stuff.

Role-play "In a restaurant"

Good evening.

Very sad :).

Yes, because I'm vegetarian.

I don't have any allergies, I just don't need any animals.

Exactly.

Some fruit, maybe later, I will order later.

28. Student 66

Puts in effort though makes frequent mistakes in grammar in longer stretches of language.

Role-play "At the airport"

Thank you.

No, there is my aunt in the city.

Yes, I am with my dog.

She is a teacher in … high school and I visit her for the Christmas holiday.

She teach literature.

For three weeks.

I think I stay with my aunt for the first week and she took me around the city and I … then I visit Times Square and I am going to shop.

Role-play "In a restaurant"

I take some vegetables for two person and … a salad and zucchini and onions.

Yes, tea please.

No, thanks.

I pay with credit card.

29. STUDENT 71

Very hesitant. Understands some questions if addressed slowly and clearly.

Role-play "At the airport"
With friends.
Yes.
Two month.
Study.

Role-play "In a restaurant"
I would like the vegetables and … salad.
Card.

Appendix I – Glossary

The terms of major importance for the research are explained in more detail throughout the thesis. Below is the glossary of foreign language teaching methodology and foreign language assessment.

Achievement test	The type of test used to evaluate whether the students have mastered a specific content, normally administered at the end of a course or a period of studying.
Authentic, authenticity	Unless defined differently in the research report, the degree to which a test replicates real-life language and situations.
Communicative teaching approach	An approach to teaching that stresses the importance of communication as a means and goal of learning a language.
Competency	While "competence" is a more general term, competency refers to the ability to do a specific thing or perform a specific task.
Content and Language Integrated Learning (CLIL)	Teaching subjects such as history, science, physics, etc. through a foreign language.
Criterion-referenced test	For this type of test, the results are interpreted in relation to a criterion, e.g. whether the test takers have mastered the course contents.
Domain, Target domain	The situations of language use to which we wish to generalize the assessment.
Extended-production response / task	The type of response that is longer than a sentence in writing or utterance in speaking and ranges from two sentences or utterances to longer stretches of language, such as essays, explanation, etc. (Bachman & Palmer, 2010)
Framework	"A selection of skills and abilities from a model that are relevant to a specific assessment context" (Fulcher & Davidson, 2007).
Generalizability theory	Also known as G-theory, a statistical theory used for evaluating the reliability of measures

	and identify different sources of measurement error simultaneously (Bachman 1990).
Language skills	Traditionally divided into reading, listening, writing and speaking.
Model	"An over-arching and relatively abstract theoretical description of what it means to be able to communicate in a second language" (Fulcher & Davidson, 2007).
Multiple-choice tests	The type of test questions that have three to four offered answers.
Multitrait-multimethod	Multitrait-multimethod design is an approach to designing correlation studies for construct validation, where each measure is considered to be a combination of trait and method, and tests are included in the design to combine multiple traits with multiple methods (Bachman, 1990).
Norm-referenced tests	For this type of tests, results are interpreted in relation to the performance, "norm", of the whole group of test takers.
Notional-functional syllabus	An approach where study materials are based on the ideas that students are expected to be able to express and the functions that they need to learn or use.
Parallel forms reliability	A measure of reliability where a test is divided into two parts, supposedly measuring the same construct and administered to the same group of test takers. The scores are then correlated to evaluate the consistency of the test.
Productive skills	Writing and speaking are traditionally known as productive skills, as they imply actual production of structures and forms of a language.
Receptive skills	Reading and listening skills are traditionally known as receptive skills, as they only imply understanding a language.
Role-play	A type of activity used in language teaching and assessment. Students assume a role of another person or character to perform the role-play.

Test-retest reliability	A measure of reliability used to evaluate the test stability over time. The same test is administered for a second time, after a period of time and then the scores are correlated to evaluate the test stability.

References

ALDERSON, J. C., (1978). *A study of the cloze procedure with native and non-native speakers of English.* PhD thesis, University of Edinburgh.

ALDERSON, J. C., Clapham, C. and Wall, D., (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

ALDERSON, J. C. (1997). Models of language? Whose? What for? What use? In: Ryan, A., Wray, A. (Eds.), *Evolving Models of Language: papers from the annual Meeting of the British Association for Applied Linguistics.* Clevedon, England: British Association for Applied Linguistics in association with Multilingual Matters.

ALDERSON, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663. doi:10.1111/j.1540-4781.2007.00627_4.x

ALTE. *Manual for Language Test Development and Examining*, For use with the CEFR, Produced by ALTE on behalf of the Language Policy Division, Council of Europe. Retrieved from: http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

BACHMAN, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

BACHMAN, L. F. (2001). Some construct validity issues in interpreting scores from performance assessments of language ability. *New Perspectives and Issues in Educational Language Policy*, 63 – 90. doi: 10.1075/z.104.07bac

BACHMAN, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476. doi:10.1191/0265532202lt240oa

BACHMAN, L. F., & PALMER, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-65. doi:10.2307/3586464

BACHMAN, L. F. and PALMER, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

BACHMAN, L. F. and PALMER, A. S. (2010). *Language assessment in practice.* Oxford: Oxford University Press.

BAKER, E. L., O NEIL, H. E., & LINN, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*: 1210 - 1218.

BOND, L. (1995). Unintended consequences of performance assessment: issues of bias and fairness. *Educational Measurement: Issues and Practice, 14*(4), 21-24. DOI: 10.1111/j.1745-3992.1995.tb00885.x

BRINDLEY, G. (1994). Task-centred assessment in language learning: the promise and the challenge. In Bird, N., Falvey, P., Tsui, A., Allison, D. and McNeill, A., (Eds.), *Language and learning papers presented at the Annual International Language in Education Conference (Hong Kong, 1993).* Hong Kong: Hong Kong Education Department, 73 – 94.

BROWN, J. D. (1996). *Testing in Language Programs.* Upper Saddle River, NJ: Prentice Hall Regents.

BROWN, J. D. (2002). The Cronbach alpha reliability estimate. *Shiken: JALT Testing & Evaluation SIG Newsletter, 6*(1): 17-18.

CANALE, M. (1983). On some dimensions of language proficiency. In Oller J. W. Jr. (Ed.), *Issues in Language Testing Research*. Rowley, Mass.: Newbury House: 333-42.

CANALE, M., & SWAIN, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, I*(1), 1-47. doi:10.1093/applin/i.1.1

CHALHOUB-DEVILLE, M. (2001). Task-based assessment: Characteristics and validity evidence. In P. Skehan, M. Swain, & M. Bygate (Eds.), *Applied language studies: Task-based research*, NY: Longham, 210-228.

CHOI, I. (1989). Past, present and future of language testing. *English Teaching,* 38: 95-135.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: The MIT Press.

COUNCIL OF EUROPE (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Retrieved from: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

COUNCIL OF EUROPE. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), A Manual.* Retrieved from: http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

CRYSTAL, D. (2006). English worldwide. In R. Hogg and D. Denison (Eds.), *A history of English language,* (430-439). Cambridge: Cambridge University Press.

DAVIES, A. (2003). Three heresies of language testing research. *Language Testing*, 20(4), 355-368. doi:10.1191/0265532203lt263oa

FULCHER, G. (1998). Widdowson's model of communicative competence and the testing of reading: An exploratory study. *System, 26*(3), 281-302. doi: 10.1016/s0346-251x(98)00020-7

FULCHER, G. (2000). The 'communicative' legacy in language testing. *System 28*(4), 483-497. doi:10.1016/s0346-251x(00)00033-6

FULCHER, G. (2004a, March 18). Are Europe's tests being built on an

"unsafe" framework? *The Guardian*. Retrieved from: https://www.theguardian.com/education/2004/mar/18/tefl2

FULCHER, G. (2004b). Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly, 1*(4), 253-266. doi:10.1207/s15434311laq0104_4

FULCHER, G. (2010). *Practical language testing*. London: Hodder Education.

FULCHER, G. (2012). Scoring performance tests. In: G. Fulcher. & F. Davidson (Eds.), *The Routledge handbook of language testing* (378-392). London and New York: Routledge.

FULCHER, G. & Davidson, F. (2007). *Language testing and assessment, An advanced resource book.* London: Routledge.

HALLIDAY, M.A.K. (1978). *Language as social semiotic.* London: Edward Arnold.

HARDING, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly, 11*(2), 186-197. doi: 10.1080/15434303.2014.895829

HARSCH, C., & RUPP, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1-33. doi:10.1080/15434303.2010.535575

HYMES, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmes (Eds.), *Sociolinguistics* (269-93). Harmondsworth: Penguin.

HYMES, D. H. (1973). Towards linguistic competence. *Texas Working Papers in Sociolinguistics, Working Paper No. 16.* Austin, Tex.: Center for Intercultural Studies in Communication, and Department of Anthropology, University of Texas.

KANE, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535

KANE, M. T. (2011). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17. doi:10.1177/0265532211417210

KANE, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000

KANE, M., CROOKS, T. & COHEN, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5 - 17.

LADO, R. (1961). *Language testing: the construction and use of foreign language tests.* London: Longman.

LINN, R. L., BAKER, E. L., & DUNBAR, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8): 15-21. doi:10.2307/1176232

LINN, R. L., & BURTON, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and*

Practice, 13(1), 5-8. doi:10.1111/j.1745-3992.1994.tb00778.x
LITTLE, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645-655. doi:10.1111/j.1540-4781.2007.00627_2.x
MCNAMARA, T. F. (1996). *Measuring second language performance*. London: Longman.
MCNAMARA, T. F. (2003). Looking back, looking forward: rethinking Bachman. *Language Testing, 20*(4), 466-473.
MCNAMARA, T. F. (2015). *Language testing*. Oxford: Oxford University Press.
MESSICK, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. doi:10.2307/1176219
MESSICK, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8. doi:10.1111/j.1745-3992.1995.tb00881.x
MESSICK, S. (1996). Validity of performance-based assessments, in Phillips G. W. (Ed.), *Issues in Large-Scale Performance Assessment* (1-18).
MISLEVY, R. J., STEINBERG, L. S., & ALMOND, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477-496. doi:10.1191/0265532202lt241oa
MIYATA-BODDY, N., LANGHAM C. S., (2000). Communicative language testing - an attainable goal? Retrieved from: www.tsukuba-g.ac.jp/library/kiyou/2000/5.LANGHAM.pdf
MORROW, K. (1981). Communicative language testing: revolution or evolution? In Brumgit, C. J. and Johnson, K. (Eds), *The communicative approach to language teaching* (143-57). Oxford: Oxford University Press.
MORROW, K. (2004). Background to the CEF, In K. Morrow (Ed), *Insights from the Common European Framework* (3-11). Oxford: Oxford University Press.
MUNBY, J. (1978). *Communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes.* Cambridge: Cambridge University Press.
NORTH, B. (2000). Linking language assessments: an example in a low stakes context. *System 28*(4). 555–77. doi:10.1016/s0346-251x(00)00038-5
NORTH, B. (2004a, April 15). Europe's framework promotes language discussion, not directives. *The Guardian*. Retrieved from: https://www.theguardian.com/education/2004/apr/15/tefl6
NORTH, B. (2004b). Relating assessments, examinations, and courses to the CEF, In K. Morrow (Ed), *Insights from the Common European*

Framework (77-90). Oxford: Oxford University Press.
NORTH, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal, 91*(4), 656-659. doi:10.1111/j.1540-4781.2007.00627_3.x
OLLER, W. John. (1979). *Language tests at school: a pragmatic approach.* London: Longman.
PAPAGEORGIOU, S., Xi, X., MORGAN, R., & SO, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly, 12*(2), 153-177. doi:10.1080/15434303.2015.1008480
ROSS, S. J. (2011), Claims, evidence, and inference in performance assessment. In G. Fulcher and F. Davidson (Eds.), Handbook of Language Testing. London: Routledge.
SAVIGNON, S. J. (1972). *Communicative competence: An experiment in foreign language teaching*. The Center for Curriculum Development, Philadelphia, PA
SAVIGNON, S. J. (1983). *Communicative competence: Theory and classroom practice*. Reading, Mass.: Addison-Wesley.
SHOHAMY, E., & REVES, T. (1985). Authentic language tests: where from and where to? *Language Testing, 2*(1), 48-59. doi:10.1177/026553228500200106
SKEHAN, P. (1999). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.
SLATER, S. J. (1980), Introduction to performance testing. In Spirer J. E. (Ed.) *Performance testing: issues facing vocational education*. National Center for Research in Vocational Education, Columbus OH, 3-17
SPOLSKY, B. (1978). Introduction: linguists and language testers. In Spolsky, B. (Ed.), *Advances in language testing* Series 2. Arlington, Virginia: Center for Applied Linguistics.
SPOLSKY, B. (1989). Communicative competence, language proficiency and beyond. *Applied Linguistics, 10*(2), 138-135. doi:10.1093/applin/10.2.138
VAN EK, J. A., (1975). *The threshold level in a European unit/credit system for modern language learning by adults*. Strasbourg: Council for Cultural Co-operation of the Council of Europe.
VAN EK, J. A. & TRIM, J. L. M. (1990a). *Threshold 1990*. Cambridge: Cambridge University Press.
VAN EK, J. A. & TRIM, J. L. M. (1990b). *Waystage 1990*. Cambridge: Cambridge University Press.
VAN EK, J. A., & TRIM, J. L. M. (2001). *Vantage level*. Cambridge: Cambridge University Press.
WEIGLE, S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

WEIR, C. J. (1990). *Communicative language testing*. New York, NY u.a.: Prentice Hall.

WEIR, C. J. (2005a). *Language testing and validation: An Evidence-Based Approach.* Basingstoke: Palgrave Macmillan.

WEIR, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing, 22(3), 281-300. doi:10.1191/0265532205lt309oa

WIDDOWSON, H. G. (1979). *Explorations in applied linguistics.* Oxford University Press: Oxford.

WIDDOWSON, H. G. (1978). *Teaching language as communication.* Oxford University Press: Oxford.

WIDDOWSON, H. G. (2001). Communicative language testing: the art of the possible. In: Elder, C., A. Brown, E. Grove, K . Hill, N. Iwashita, T. Lumley, T. McNamara, K. O'Loughlin (Eds.*), Experimenting with uncertainty. Essays in honour of Alan Davies*. Cambridge: Cambridge University Press, 12-21.

WIGGINS, G. (1989). A true test: Towards more authentic and equitable assessment. *The Phi Delta Kappan, 70*(9), 703-713.

WIGGLESWORTH, G. (2008). Task and performance based assessment. In: Shohamy, E., Hornberger, N. H. (Eds.), *Encyclopedia of Language and Education*, New York: Springer, 111 - 122.

WILKINS, D. (1978). Proposal for levels definition. In: J.L.M. Trim (Ed.), *Some possible lines of development of an overall structure for a European unit / credit scheme for foreign language learning by adults (71–78).* Strasbourg, France: Council of Europe.

Websites

Cifre chiave sull'insegnamento delle lingue a scuola in Europa 2012, Eurydice. Retrieved: 17 October 2017, from: http://eacea.ec.europa.eu/education/eurydice/documents/key_data_series/143IT_HI.pdf

EF English Proficiency Index. Retrieved: 17 October 2017, from: http://www.ef.com/__/~/media/centralefcom/epi/downloads/full-reports/v6/ef-epi-2016-english.pdf

EF English Proficiency Index for Schools, Retrieved: 17 October 2017, from: http://www.ef.edu/epi/reports/epi-s/

Europeans and their languages, European Commission, 2012. Retrieved: 17 October 2017, from: http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/yearFrom/1974/yearTo/2012/surveyKy/1049

Higher Education and Research, Council of Europe. (n.d.). Retrieved: 17 October 2017, from: https://www.coe.int/t/dg4/highereducation/EHEA2010/BolognaPedestrians_en.asp

Multilingualism – Education and training – European Commission. (n.d.). Retrieved 17 October 2017, from: https://ec.europa.eu/education/policy/multilingualism_en

Ofqual. Retrieved: 17 October 2017, from: https://www.gov.uk/government/organisations/ofqual

Percentage of the population able to hold a conversation in English (self-reported). Based on Eurobarometer 365, European Commission. Retrieved: 17 October 2017, from: https://jakubmarian.com/map-of-the-percentage-of-people-speaking-english-in-the-eu-by-country/

Public Opinion – European Commission. (n.d.). Retrieved: 17 October 2017, from: http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/General/index

Ministerial Decisions

Decree 3 November 1999, no. 509 (D.M. 3 novembre 1999, n. 509). Regolamento recante norme concernenti l'autonomia didattica degli atenei. Retrieved on 17 October 2017 from http://www.miur.it/0006Menu_C/0012Docume/0098Normat/2088Regola.htm

Decree 24 February 2000, no. 49 (D.M. 24 febbraio 2000, n. 49). Retrieved on 17 October 2017 from https://archivio.pubblica.istruzione.it/argomenti/esamedistato/secondo_ciclo/quadro/dm49_00.htm

Guidelines 4 February 2010 (Regolamento 4 febbraio 2010). Retrieved on 17 October 2017, from http://www.edscuola.it/archivio/norme/programmi/licei_2010.pdf

Law 28 March 2005, no. 53 (L. 28 marzo 2003, no. 53). Retrieved on 17 October 2017 from https://archivio.pubblica.istruzione.it/mpi/progettoscuola/allegati/legge53_03.pdf

Indicazioni nazionali riguardanti gli obiettivi specifici di apprendimento concernenti le attività e gli insegnamenti compresi nei piani degli studi previsti per i percorsi liceali. (2010). Retrieved on 17 October 2017 from http://www.indire.it/lucabas/lkmw_file/licei2010/indicazioni_nuovo_impaginato/_decreto_indicazioni_nazionali.pdf

Il regolamento degli istituti professionali. (2010). Retrieved on 17 October 2017 from http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_professionali_04_02_2010.pdf

Il regolamento degli istituti tecnici. (2010). Retrieved on 17 October 2017 from http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_tecnici_def_04_02_10.pdf

Progetto Lingue 2000. Retrieved on 17 October 2017 from http://www.edscuola.it/archivio/norme/programmi/progettolingue.pdf