University of Jaen

Department of English Studies

DOCTORAL DISSERTATION

WASHBACK TO THE LEARNER: ARE EXAM-ORIENTED COURSES EFFECTIVE?

Victoria Peña Jaenes

SUPERVISOR

Dr Antonio Bueno González

Jaén

Gracias a mis padres, a Manuel y a mi hermano.

Gracias, Manuel, por animarme a seguir mejorando en todos los sentidos, por ayudarme a ver la vida desde otra perspectiva y por hacer mis días más felices.

Gracias, papás, por vuestro Excel infinito. Gracias, mamá, por estar siempre dispuesta a ayudar. Gracias, en especial, papá, por ser siempre un ejemplo de esfuerzo y honestidad y por tu apoyo constante e incansable para hacer esta Tesis.

Gracias, Kike, por tu calidez y tu apoyo. Gracias por esas llamadas reparadoras.

ACKNOWLEDGEMENTS

I am most grateful to Dr Antonio Bueno González, my supervisor, for understanding my timeframes, for supporting my moments of inspiration and, above all, for offering his guidance and motivation in the difficult times. I would like to express my gratitude to statistician Valentina Cuevas for her technical support with statistical analyses. Thank you very much to Magdalena Martos Puelma and the Centro de Estudios Británicos team for opening your language school to this research project and for giving me the opportunity to teach. Thank you very much to Dr Antonio Bueno González and the Centro de Estudios Avanzados en Lenguas Modernas for opening your language school to this research project and for introducing me to the fascinating world of testing and assessment. Special thanks to Dr Ventura Salazar for offering quick and reliable support.

I am thankful to my family and friends for listening to me, for supporting me, and for understanding the endless hours in front of the computer.

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

INDEX OF ABBREVIATIONS AND ACRONYMS

ACLES	Asociación de Centros de Lenguas en la Enseñanza Superior
ALTE	Association of Language Testers in Europe
BICS	Basic Interpersonal Communication Skills
CALP	Cognitive Academic Language Proficiency
CAP	Certificado de Aptitud Pedagógica
CEALM	Centro de Estudios Avanzados en Lenguas Modernas
CEB	Centro de Estudios Británicos
CEFR	Common European Framework of Reference for Languages
CELST	Computer-based English Listening and Speaking Test
COLT	Communicative Orientation to Language Teaching
CPE	Certificate of Proficiency in English
EAP	English for Academic Purposes
ECT	Entry Cambridge Test
EFL	English as a Foreign Language
EGR	EFL proficiency Graduation Requirements
EGT	Entry Grammar Test
ELT	Entry Listening Test
EoCCT	End-of-Course Cambridge Test
EoCGT	End-of-Course Grammar Test
EoCLT	End-of-Course Listening Test
EoCQ	End-of-Course Questionnaire
EoCRT	End-of-Course Reading Test
EoCST	End-of-Course Speaking Test
EoCUoET	End-of-Course Use of English Test
EoCVT	End-of-Course Vocabulary Test
EoCWT	End-of-Course Writing Test
EQ	Entry Questionnaire
ERT	Entry Reading Test

ESL	English as a Second Language
EST	Entry Speaking Test
ETS	Educational Testing Service
EUoET	Entry Use of English Test
EVT	Entry Vocabulary Test
EWT	Entry Writing Test
FCE	First Certificate in English
FIAC	Flanders' Interaction Analysis Categories
FLINT	Foreign Language Interaction
GEPT	General English Proficiency Test
HKCEE	Hong Kong Certificate of Education Examination in English
IELTS	International English Language Testing System
IIS	IELTS Impact Study
ILTA	International Language Testing Association
L2	Second Language
LOA	Learning Oriented Assessment
MA	Master of Arts
n	number
NVLT	New Vocabulary Levels Test
RELANG	Relating language curricula, tests and examinations to the Common European Framework of Reference for Languages
TKT	Teaching Knowledge Test
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
UCLES	University of Cambridge Local Examinations Syndicate
VLT	Vocabulary Levels Test

INDEX OF APPENDICES

1. INTRODUCTION

Tell me and I forget. Teach me and I remember. Involve me and I learn.
(Benjamin Franklin)

1.1. Learning

Learning is crucial for development. It is the foundation of society and progress. Learning and how best to learn are the cornerstone of this PhD Thesis. But what exactly do we mean by learning?

Learning is a process that leads to change in the way learners see the world. This change can take place at the level of knowledge, attitude or behaviour (Queen's University, n.d.). It is *active* because learners need to explore the world around them, observe and interact with phenomena, manipulate objects, engage in conversations and with others given its social component. The degree of success of the learning process is based on three main factors: cognitive processes, engagement and learning behaviours (Knight, 2020).

The *cognitive aspects of learning* have to do with the thinking processes and mental procedures. They are much deeper than memorisation and information recall. In fact, they involve *constructive learning*, that is, building knowledge and skills based on prior knowledge and understanding, which is considered as the foundation for all future learning, and being able to engage in independent and critical thinking to transfer knowledge to new and different contexts (Queen's University, n.d.). In addition, there is a number of factors that impact on learning and which are crucial for teaching. They range from the *cognitive load*, i.e. how much new information learners can process; *consolidation*, which includes techniques such as spaced repetition, elaboration, rehearsal and relating to personal experience; and *dual coding*, which has to do with reinforcing information by giving visual and verbal input.

The active nature of learning makes *engagement* a basic ingredient of the process and here a number of aspects comes into play. First, *self-efficacy* (Knight, 2020) or *expectation of attainment of a goal* (Svinicki, 2004) because learners engage better if they believe they can be successful in a task that is *relevant* to them because it meets their interests. There must a balance between the challenge the task represents and the beliefs of succeeding. Moreover, learners learn better when they feel they can control what they learn, when they do it and how. This is what Knight (2020) calls

agency. The social component of learning increases learners' willingness to engage in learning, that is, if learners feel comfortable interacting with others, trying out language and engaging in peer-assessment, they are likely to learn better. Also, *enjoyment* and *curiosity* are, without any doubt, crucial elements of successful learning because they influence the direction, intensity, persistence, and quality of the learning behaviours in which students engage (Ambrose et al., 2010:69).

Finally, it must be taken into account that learning is not something done to the students, but rather something that students do, hence the importance of *learning behaviours*. *Goal setting* is crucial because research has consistently suggested that learners achieve most and gain most when they direct their attention to a specific objective that is attainable (Crooks, 1988; cited by Green, 2007:23), manageable and work towards it. If learners perceive the goal as high value they will be more willing to invest time and effort. For that they will need to use specific strategies – what Knight (2020) calls *learning management*, and also reflect on them, evaluating what they are doing well and the aspects that need to be improved.

Learning is the basis for development and evolution and even more in the world where we live, which is in constant change. One of the main challenges learning and teaching face at present is to make lifelong learning a reality.

1.2. Lifelong learning, languages and English as a *lingua franca*

Lifelong learning has been identified by leading institutions such as the United Nations and the European Commission as a critical component in cultural, economic, and environmental prosperity. The United Nations, for instance, believe lifelong learning can "help eradicate poverty, protect the planet, secure human rights, build peaceful, inclusive and equal societies, and promote social, economic, cultural and technological progress" (UNESCO, 2016). In a similar vein, the European Commission has listed eight key competences within the framework of lifelong learning to better prepare people for today's societies (European Commission, 2017). One of these key competences is increasing language ability and the number of languages learned. The European Council on 14th December 2017 concluded that languages play a social role because they help people to have a more active role as citizens and be better prepared to cope with the challenges of today's multilingual and diverse societies. Besides, languages can be said to strengthen cultural understanding and peace because they unite people and render other countries and their cultures

accessible. From an economic perspective, the ability to speak a foreign language enhances employability and mobility, hence making countries more competitive.

English has gained momentum as the *lingua franca* thanks to globalisation. More and more people learn English because there is an increasing number of people who need English for employment, either to find a job or to obtain better working conditions (Chávez Zambano, Saltos Vivas & Saltos Dueñas, 2017:761). The importance of English is such that learning and speaking English is no longer a luxury but a necessity no matter where you live or what your field of expertise is (Jaimechango, 2009; cited by Chávez Zambano, Saltos Vivas & Saltos Dueñas, 2017:761).

Given the relevance of languages in general and English in particular, the teaching, learning and assessment of languages has been in the spotlight. Governments and leading experts in the field of language education collaborate to modernise language teaching and to make it more efficient by promoting innovative teaching methods and fostering common methods of assessment because it is understood that language teaching needs to go hand in hand with assessment methodologies. An example of this endeavour is the "Relating language curricula, tests and examinations to the Common European Framework of Reference" (RELANG) initiative, which helps educational authorities to align language examinations to the proficiency levels of the *Common European Framework of Reference for Languages* (European Commission, n. d.).

1.3. Learning and assesssment

The role of testing and examinations has evolved in time. In some cases, they have been used as an instrument of power and control. In Imperial China, a thousand years or more ago, the highest officials were selected using probably the first civil services examinations ever developed (Lai, 1970; Hu, 1984; Arnove, Altback, & Kelly, 1992). In doing so, authorities were not only selecting staff, but were also establishing and controlling the education system with a very high-stakes exam that would have an impact on the candidates' lives and on the future of the Empire (Spolsky, 1995a, 1995b). However, the very early notion of some of the principles of assessment were already present in these examinations in the sense that measures were implemented to try to produce a fair test and avoid corruption. This use of examinations as a way to avoid corruption and foster excellence and talent is also documented by Eckstein and Noah (1992) and Bray and Steward (1998) and it is the spirit behind the current use of tests so it is equally powerful today as centuries ago.

Examinations have also been used as a way to foster educational reform (Linn, 2000; cited by Cheng, Watanabe & Curtis, 2008:6). This use has brought about a change in the direction of the learning, teaching and assessment process. Traditionally, tests came at the end of the teaching and learning process. Nevertheless, some of them were so important for candidates that they influenced the attitudes, behaviours, and motivation of teachers, learners and parents. This influence is perceived as working in a backward direction, hence the term *washback* (Pearson, 1988:98; cited by Cheng, Watanabe & Curtis, 2008:7). However, like Davies (1985), Pearson believed that the direction in which washback actually works is forward because when a new high-stakes exam was introduced the teaching materials were aligned with the new test, and the stakeholders were required to adapt, and often work harder, to achieve high scores on the test (Cheng, Watanabe & Curtis, 2008:12). This power to transform and change attitudes and behaviours has been the main topic of washback research, e.g. Pearson (1988), Shohamy (1992), Cheng (1997), Andrews (2002), Read & Hayes (2003) Saif (2006), all cited by Tsagari (2007:13). Some scholars saw it as an opportunity to innovate and to promote the evolution of teaching and learning. For others, however, it was perceived as a way to impoverish learning in the sense that teaching and learning were based mainly on past examination papers (Davies, 1968:125) and coaching classes were limited to preparing students for exams rather than teaching language (Wiseman, 1961:159; cited by Cheng, Watanabe & Curtis, 2008:9).

1.4. Assessment as an opportunity for learning

> Tests and exams have been seen as an occasional necessary evil, a dose of unpleasant medicine, the taste of which should be washed away as quickly as possible.(Cheng, Watanabe & Curtis, 2008:14)

The value of exams for selection purposes and as levers for innovation is well established. The use of exams in the day-to-day classroom practice is nothing new. However, their value in this context is not very clear as teachers often complain that there are too many exams and too little time to teach, and students see exams as a way to control and are only interested in obtaining a mark, making it a source of anxiety. In the light of this, it is not surprising that testing and assessment are

not highly valued in this context. Clearly, if exams are used as mere snapshots of students' abilities and knowledge and nothing else, with the only evidence of a mark attached to them, one could indeed agree that there are too many exams.

Fortunately, assessment can offer rich information about the students and for the students and help make informed decisions in teaching and learning. For that to happen, assessment needs to connect to the principles of learning and be integrated with learning. The *Common European Framework of Reference for Languages* (henceforth CEFR) has become *the* document of reference. Its Can-Do Statements for each skill and communicative situation can be easily adapted to create suitable learning objectives for each level. The positive way in which they are formulated, showing what learners at each level of ability can do, is an encouraging method to design the different steps language learners climb. By aligning educational objectives to the CEFR, institutions create a more coherent learning path and contribute to the internationalisation of teaching and learning. Once the learning objectives are clear, explaining the connection between them and the classroom practice and activities becomes paramount because it is likely to increase motivation since students understand the reason why different activities are necessary and feel they have more control of their own learning. What is more, connecting the learning objectives with assessment and familiarising students with the assessment criteria and methods makes the whole teaching and learning processes more transparent. This is because, on the one hand, students know what to expect and how they will be assessed; and, on the other hand, because the grade or mark carries more meaning than just a pass or a fail. The feedback helps learners to identify their weaknesses and to identify the aspects that require more intense work on their part. However, the feedback will certainly show what learners have done well and become a source of motivation that gives them energy to continue working. Producing this feedback entails thoughtful reflection on the teachers' part, which leads to a better understanding of their students' performance against the assessment criteria and also facilitates a more accurate plan of the teaching, with more realistic stages and outcomes.

The connection between learning and the assessment that takes place in the classroom and is carried out by teachers is probably easier to understand than that between learning and summative assessment, which usually happens at the end of a course because it is perceived as something that marks the end of a learning stage. This connection is even more difficult to see when the assessment is carried out by external assessment bodies, who do not know the students,

their circumstances or their aims in life. Nevertheless, the alignment of proficiency exams to the CEFR, the importance given to lifelong learning, and the concern of assessment bodies with the effects and impact of a test on individuals, educational processes and on society in general has strengthened the connection between summative assessment and learning.

The alignment of proficiency exams to the CEFR and to the principles of language testing has made the testing methods, instruments and results more transparent and has allowed for the internationalisation of qualifications. Attaching a CEFR level to any grade or mark automatically provides the necessary context to interpret the results and to make decisions in terms of what a person can do and the situations they can deal with based on their language abilities. It has opened a number of opportunities for universities and internationalisation programmes such as Eramus, thus making, for instance, the admission of foreign students more automatic and easier. In the professional sector, selection processes have also been facilitated, and the same could be said when language ability is a factor for immigration.

Nevertheless, these possibilities are based on the premise that the results obtained by candidates meet all the necessary requirements i.e. they are valid, reliable and fair. Exam boards are in constant evidence-based evolution to design exams which are up-to-date with the principles of language learning and the latest trends in assessment because they are aware of the stakes attached to their exams because of the recognition they have. In addition, a number of associations such as Ofqual and ALTE strive to ensure that language tests meet the necessary quality standards to make them fair, reliable, practical, valid and to maximise positive impact on individuals and society as a whole and to minimise any unintended effects.

The relevance of lifelong learning and the fact that one never stops learning a language connected with the principle of impact and washback of exams has given proficiency exams a new role. They should no longer be perceived as the last stage in the language journey but rather as a full stop, another milestone reached. Proficiency exams are thorough analysis tools. They must be so because their results are used to make life-changing decisions (Raban, 2008:x and University of Cambridge Local Examinations Syndicate, 2016). Therefore, it is useful to take advantage of all the data exams gather from candidates to help them to understand what they can do well and the difficulties they have, in other words, to help them to improve and continue learning and know their abilities better. This can be done by producing a score that is associated with a description of

what learners at that level of ability can typically do, so here again we see the connection between a score, Can-Do Statements and learning and it is likely to be an example of positive washback.

1.5. It is complicated

> I cite this as an example of how important it is to research one's beliefs, rather than simply to accept what appear to be truisms. (Cheng, Watanabe & Curtis, 2008:x)

The last section finished with the word *likely* limiting the positive effect of the connection between exams, Can-Do Statements and learning. While one could think that there are no drawbacks to this connection, relevant authors on the subject such as Cheng, Watanabe and Curtis (2008:11) suggest that it is probably safer to add this limitation especially if we take into account that the positive or negative nature of washback can be influenced by many contextual factors, among which the authors cite *test related factors* such as the test methods, content, the skills tested, the purpose of the test, and the decisions that will be made of the basis of the test results. Furthermore, some exams and assessment bodies have a certain *prestige* attached to them within the educational system, which together with the *stakes* of the test can mediate how washback plays its role. In addition, the context in which the learning takes place, be it the *micro-context* i.e. the school where it takes place, or the *macro-context* i.e. the society, the city or the region where the test is used also influences the washback (Cheng, Watanabe & Curtis, 2008:22).

The main actors in the teaching and learning process are *teachers* and *learners*. Teachers' educational backgrounds, their beliefs about assessment, teaching and learning and their personal perceptions of their exams are also identified as intervening in the washback. Finally, the learner, an area that has been under-investigated in the literature, despite the beliefs of leading experts such as Green (2007:314) that the response of the individual learner to the demands of a test and to other features of the learning context have a greater influence on the learning outcomes than the classes or the materials used. Some studies have tried to access the learner through their teachers. However, there is evidence to suggest that the nature and extent of washback to learners does not bear a transparent relationship to washback to the teacher.

The complexity of washback has been evidenced by the work of leading experts in the field. Alderson and Wall (1993), who have led the research in the matter with their seminal work where they wondered "Does washback exist?" and enumerated several "Washback Hypotheses",

concluded that further research on washback was needed. The questions to answer now are what washback looks like, what causes it, and why it happens (Cheng, Watanabe & Curtis, 2008:ix) and the two routes signalled by Alderson and Wall (1993; cited by Cheng, Watanabe & Curtis, 2008:12) are: studying the role of exams as a lever for motivation and change, which has attracted considerable attention; and focusing on motivation and performance.

1.6. What to expect

> Washback is thus grounded in the relationship between preparation for success on a test and preparation for success beyond the test, in the domain to which the test is intended to generalise and to which it may control access. (Green, 2007:1)

The complexity of washback and the context-specific factors that influence its nature account for the need for studies with very clearly defined contexts and aims, where nothing is taken for granted. The driving force behind this Doctoral Dissertation is to understand how professionals in teaching and assessment can help learners to learn better and also to give learners the tools to have a more active and effective role in their own learning. The study is built on the author's recent experience teaching English and preparing students for Cambridge English Qualifications and the knowledge gained as a result of the MA thesis published in 2015 (Peña Jaenes, 2015), which looked at the washback from the perspective of language courses and the assessment of writing skills. The present PhD project focuses on one of the two key research lines identified by Alderson and Wall (1993), that is, language learners, and analyses their performance and motivation. Its ultimate objective is to obtain an evidence-based understanding of students' progression by trying to answer three research questions:

i) Do students enrolled on more exam-oriented (Centro de Estudios Británicos, CEB) courses show a better performance in B2 First mock exam than those enrolled on general English (Centro de Estudios Avanzados en Lenguas Modernas, CEALM) courses?

ii) Do students enrolled on more exam-oriented (CEB) courses improve their language knowledge and abilities?

iii) Do students become more autonomous and independent learners as a result of preparing for Cambridge B2 First exam?

The main objective of the project and the three research questions have guided the project and enable us to draw conclusions based on evidence that despite being surprising at times – which only supports the quote by Cheng, Watanabe and Curtis (2008:x) – equips the author to have a better understanding of washback and students' performance and motivation. It also sheds light on the balance between preparing for success on a test and preparing for success beyond the test, which should be the ultimate goal of any language course. As I personally believe that it happens with most research projects, they open our eyes to new lines of research. The present one encourages the author to continue studying and working on assessment. This learning process is narrated in this project, which is structured as follows:

The chapter that follows the Introduction is a review of the literature on the subject. It starts by exploring the relationship between testing, assessment and evaluation, their evolution and the types and purposes they serve, and the qualities that any test should have. The chapter pays special attention to all four skills, from a more theoretical perspective to more practical terms when analysing the implications for assessment. The second part of the Literature Review zooms in and deals with washback. It describes and comments on the research carried out to date in an attempt to identify possible gaps where the contribution of this project can be considered of use, and studies how the perception and the concept of washback has evolved to the more mature understanding of the term, tapping into the complexity of the effect and the difficulty in identifying what causes and influences it. The third part focuses on accreditation exams in general as an introduction to the exam that is at the heart of this research, Cambridge B2 First. It describes the assessment body behind it, the evolution of the exam and its design, all of them key elements to understand the value of the exam.

Before moving to the fourth chapter of the project, and in light of the literature review, the rationale that inspires the project is defined. The gaps in the research are identified and the personal and academic interest of the author is explained, which leads to the main objective of the project, its specific objectives and the more detailed research questions. They determine the methodology followed and its description makes up the third main chapter of the project.

The fact that the research was carried out while the author was working as a teacher and the objectives of the project were crucial when deciding to opt for classroom research. The literature reviewed and, again, the objective – to gain evidence-based knowledge – demanded the

use of a comprehensive range of data collection methods such as classroom observation, teacher and student questionnaires and different test instruments to obtain data about the students' performance – quantitative data – and also information to contextualise and explain the findings – qualitiative data. In such a way, the author was trying to solve some of the problems found in her previous research project – although new ones would appear, and was trying to build multiple channels for the different type of information to flow. The chapter offers a detailed description of the instruments used and the validation processes used and the other quality control measures implemented to increase the validity and usefulness of the findings. The participants, institutions and the city where the study was conducted are described as they are crucial to contextualise the washback and the findings. Finally, the data collection process and the variables that structure the discussion and results are also analysed.

The fifth main chapter of this project reports the results obtained from the statistical analyses carried out and analyses and discusses them in the light of the qualitative and quantitative data obtained from questionnaires, the observation schedule and relevant studies on the matter. The results description, analysis and discussion try to answer the research question and explain the possible contributions that this research project can make.

The last chapter is devoted the conclusions drawn from the analysis of the results, which answer the three research questions. It also highlights some relevant findings which, despite not being directly related to the research questions, could be considered interesting or relevant. The chapter finishes with the main limitations encountered during the project, which in many cases inspire the author to pursue future research to try to overcome them and continue studying the subject.

For clarity, some linguistic choices and conventions will be explained. The main focus of this research project is the effect that tests have on learners, this effect is known as *washback* or *backwash* although, as we will see, there are other terms that some experts use. While both terms are acknowledged, the term *washback* has been used, mainly because of the influence of Alderson and Wall (1993), who used this term in their seminal work "Does Washback Exist?" and of Anthony Green (2007), who uses the term washback in his research. Alderson himself explained their choice and the relationship between both terms in Cheng, Watanabe and Curtis (2008:xii) as follows:

> If I may permit myself the luxury of a footnote, in reference to the use of two terms to refer to the same phenomenon, namely backwash and washback, I should explain that one of the reasons why the Alderson and Wall article was entitled "Does Washback Exist?" was because it seemed to us that the word washback was commonly used in discussions, in presentations at conferences and in teacher training. When I was studying at the University of Edinburgh, Scotland, for example, Alan Davies, the doyen of British language testing, frequently used the term washback and I do not recall him ever using backwash. Whereas in what literature there was at the time, the word "backwash" seemed much more prevalent. Hence another reason for our question: "Does Washback Exist?" But to clarify the distinction between the terms backwash and washback: there is none. The only difference is that if somebody has studied at the University of Reading, UK, where Arthur Hughes used to teach, they are likely to use the term backwash. If they have studied language testing anywhere else, but especially in Edinburgh or Lancaster in the UK, they will almost certainly use the term washback. (Cheng, Watanabe & Curtis, 2008:xii)

The participants in this project studied or worked at two well-known language schools in Jaén (Spain). However, the aim of this project is not to compare the two institutions, whose directors kindly allowed the author to carry out her research, but to understand the progression of students who are enrolled on more exam-oriented courses and of students who are enrolled on more general English courses. Nevertheless, for practical reasons the former have often been identified as CEB students, because all of the students at the Centro de Estudios Británicos (henceforth CEB) were enrolled on more exam-oriented courses, while the latter have often been identified as CEALM students, because all the participants in this research studying at the Centro de Estudios Avanzados en Lenguas Modernas (henceforth CEALM) were enrolled on more general English courses.

The intended contribution of this PhD Thesis is to obtain an evidence-based understanding of students' progression. To be more precise, it aims to measure the effectiveness in terms of score gain in the mock exams of B2 First exam of courses that are oriented towards preparing students for Cambridge English B2 First exam in comparison with general English courses; the effectiveness in terms of improvement in general language ability of courses that are more exam-oriented as

compared with more general English courses; and to gain evidence-based knowledge of the potential washback that B2 First exam may have on learner autonomy and independence.

Figures and tables are numbered following the order in which they appear in the text. There is an index of figures, tables and appendices to help the reader to retrieve them if necessary. Finally, acronyms are only expressed in full the first time they appear in the text and there is an index of acronyms for the reader's reference.

2. LITERATURE REVIEW

2.1. Testing, assessment and evaluation: an overview

Testing, assessment and evaluation can sometimes be used as synonyms. However, when comparing these three concepts, testing is a more specific concept. *Testing* refers to a formal or informal task or procedure designed and used to elicit certain behaviour from which one can make inferences about a person's ability or knowledge in a given area (Carroll, 1968:46; cited by Cruz Trapero, 2016; and Brown, 1987:219; cited by Bueno González, 1996). As for *assessment*, it covers a much broader cycle of activities (Green, 2014:6), including testing – deciding on the contest of the test, measuring and scoring the performance and deciding on the meaning of the scores, collecting and combining information, and providing feedback (Shohamy, 2017; cited by Muñoz, Véliz-Campos & Véliz, 2019:101). To this definition, Bachman (2004; cited by Bachman & Palmer, 2010) adds that this collection process should use procedures that are systematic and substantially grounded. Clapham (2000) states that assessment has traditionally been associated with more informal methods including checklists, journals, logs, self-assessment and teacher observations (Huerta-Macías, 1995; cited by Díez-Bedmar, 2010:69) and has hence been considered to be closer to classroom-based teacher assessment as opposed to testing, which has been identified with standardised tests. Finally, *evaluation* has a wider scope as it encompasses all the data and information obtained from the learning process either elicited from tests or through other means and involves making value judgements and decisions on the basis of this information (Cruz Trapero, 2016:46). For Bachman and Palmer (2010) evaluation can be seen as one possible use of assessment.

Looking at language testing and assessment in particular, Shohamy (2008; cited by Díez-Bedmar, 2010:69) points out that they are interdisciplinary disciplines which are informed by linguistics, applied linguistics, language acquisition and language teaching as well as the disciplines which are related to testing, measurement and evaluation. For Shohamy (2008) language testing and assessment are the result of two fields of expertise: on the one hand, the language testing field, devoted to the study of the construct that is to be assessed; and, on the other hand, the field of testing and measurement, which focuses on the way to assess such construct.

2.1.1. Evolution

Language assessment, which was already used in the second century in the form of oral exams used to select the most suitable candidates to work as civil servants in China (Alcaraz, 2015; cited by Muñoz, Véliz-Campos & Véliz, 2019:98), has been shaped by a wide range of influences, from practicality to established customs, including political expediency and developments in language teaching, applied linguistics and other allied disciplines. Social changes such as mass migration and tourism have brought new reasons for learning and using languages and have also impacted on assessment (Green, 2014:173). As a result, it is not surprising that the reality when the Cambridge first exam was taken back in 1913 by three candidates (Hawkey & Milanovic, 2013) has changed dramatically and now the number of candidates taking Cambridge exams in a year has reached the figure of 5.5 million (University of Cambridge Local Examinations Syndicate, 2020c). At present, the use of assessment has extended for selection processes or to inform learning and teaching (Brown, 2004; cited by Muñoz, Véliz-Campos & Véliz, 2019:99).

Although as mentioned by Alcaraz (2015) assessment has a long tradition, this account of the evolution of language assessment starts in what Bueno González (1996) called the pre-scientific period, before the early 50s, when test qualities considered vital today such as reliability and aspects widely used today as statistics were ignored and testing was considered intuitive (Madsen, 1983). Even though the situation has changed dramatically, the influence of this period remains in the interest in authentic and contextualised language as well as on the emphasis on communication and interaction. Lado (cited by Green, 2014:180) rejected the techniques of traditional tests on the grounds that knowing rules would not necessarily mean that candidates can apply them in real world communication. Instead, he insisted that test takers should demonstrate the ability to apply what they knew to real samples of language. However, he prioritised efficiency over authenticity since he was in favour of discrete-point testing. This approach also fostered a more humanised attitude from the tester point of view (Bueno González, 1996; cited by Peña Jaenes, 2015:12). During the 50s and 60s one of scholars' main concern was to produce objective and reliable tests, which resulted in the creation of multiple-choice tasks and also to the testing of usage. Looking at this movement now, the main shortcoming was that testing was considered to have an atomistic view and that the desire to maximise reliability compromised the validity of the tests (Bueno González, 1996; cited by Peña Jaenes, 2015:13).

During the 60s and 70s, applied linguistics showed their disagreement with both Bloomfield's structuralist linguistics and behaviourist learning models backed by Lado (Green, 2014:185). In 1985, Traynor and other scholars pointed out that the fact that productive tasks were not included in tests could result in less teaching time devoted to these skills, and assessment institutions started offering speaking and writing tests first as optional extras and then fully incorporated into the tests (Green, 2014:185). Carroll (1961:318; cited by Green, 2014:188) shared Lado's interest in including items testing very specific linguistic aspects and skill but gave more importance to the skills than to the components, and recommended the use of, what he called, integrative tests, which could measure the rate and accuracy with which candidates could put into practice a variety of aspects of language ability under realistic time constraints to produce or understand complete utterances or texts. Lado was also aware of the relevance of context validity, as evidenced by the fact that, when working in the area of language assessment for admission purposes of international students, he suggested that it was necessary to investigate the task types that students may need to perform at university and the levels of performance required to deal with them successfully. However, university students were not Carroll's only concern as he also investigated into how to best select written texts for children of different ages (Carroll et al., 1959; cited by Green, 2014:189): the cloze procedure.

At the end of the 70s, Oller (1979; cited by Green, 2014:198) supported the idea that valid tests should be integrative and pragmatic and showed his position against discrete-point testing while he supported cloze tests. In line with Oller's views regarding integrative assessment and cloze procedures was the Natural Approach, which became popular in the 80s. This trend supported the idea that teachers should present learners with linguistic input at a level of proficiency just above their own production level. However, the concept of a unitary competence was not sustained and research showed that cloze procedures were limited mainly to assessing grammatical and lexical knowledge, just as discrete-point testing. As a consequence, the Natural Approach was left behind, although its integrative approach to assessment had a lasting effect, and scholars started advocating communicative or specific purposes approaches to assessment as socio-linguistics started taking shape leading to a shift in perspective. Scholars such as Cooper (1968), Campbell and Wales (1970:274), Hymes (1972), and Morrow (1979) focused on the context and on the ability of producing and understanding language which is appropriate to the context in which it is produced. The candidate was perceived as an *insider* that communicates a message on the basis of a well-

defined situation (Bestard Monroig & Pérez Martín, 1992; cited by Peña Jaenes, 2015:13). Authenticity became a key issue when selecting test material because the objective was to reproduce real-life conditions in testing (Green, 2014:199). However, scholars such as Weir (1981:29; cited by Peña Jaenes, 2015:13) pointed out that authenticity was compromised in the testing context since the conditions for real-life communication cannot be replicated there. In addition, the focus on authenticity compromised the reliability of the test to some extent. As a result, multiple-choice questions remain present (Powers, 2010; cited by Peña Jaenes, 2015:13) and they are combined with other types of tasks as a way to reach a balance between authenticity and reliability. The term *integration*, which Lado (1961) had used to refer to listening or reading, was now being used by scholars such as Caroll (1980) to refer to the integration of modalities involved when carrying out tasks in real life.

In 1990 Bachman introduced his model of communicative language ability (Bachman, 1990; Bachman & Palmer, 1996 and 2010; cited by Green, 2014:202). This model was built on earlier models suggested by Canale & Swain (1980) and Canale (1983), among others, and was present in all subsequent language testing practice. However, it has also received some criticism, for example, on the grounds that it offered little information about how users processed language in their minds (Weir et al., 2013; cited by Green, 2014:204).

As a result of social and technological changes, the concept of *permanent education* (Council of Europe, 1970) or *lifelong learning* started taking shape, leading to a more flexible approach to teaching and learning. This was also reflected on assessment, as qualifications were required to reflect this greater flexibility and become modular and more learner-centred. Can-Do Statements were introduced and learners became involved in setting objectives of their own learning and became responsible for keeping track of their own progress. By the 1990s, self-assessment, peer assessment and feedback were being used in language classrooms. The publication of Black and Wiliam's *Inside the Black Box* in 1998 fostered learner-centred approaches in language assessment through the idea of *assessment for learning* (Green, 2014:207). This new approach to assessment relied on the value of feedback as a learning tool that learners can use to identify where they need to improve and learn from the experience of assessment. It also introduced the idea of *mediation*, understood as the distance between what learners can do with support and what they are able to do by themselves. Within the assessment-for-learning movement, two different trends can be observed. On the one hand, some scholars felt closer to Bachman's model and socio-cognitive

approach while, on the other hand, some experts were more in favour of a more socio-cultural analysis. The former accept the model of learning based on stages of acquisition and growth in individual functional abilities. They perceive assessment for learning as a way to support students all the way up an imaginary language ladder than takes them from beginner to mastery. This approach sees the usefulness of standardised testing and classroom-based assessment because they are perceived as complementary (Green, 2014:210). The latter, however, see language learning as a social process of assimilation into a specific local set of cultural practices and relationships. The learner takes an active role in a specific community and the teacher acts as mediator between the student and the target culture (Green 2014:10). From the socio-cultural perspective, present frameworks and scales are a reflection of one – dominant – group of values (Leung & Lewkowicz, 2006; cited by Green, 2014:210) which could be subject to change. An example of this can be observed in the description of language ability: in the past, the native speaker was used as a reference and this has now been replaced by adjectives describing the production and reception such as “precise”, “sophisticated”, “completely natural”, or “flexible” (University of Cambridge Local Examinations Syndicate, 2019:26)

While it is true that some of the questions that concerned language testers in 1961 are still in discussion at present (Green, 2014:172), over the past two there has been growing interest in aspects related to values: the who and why of assessment, this is what Davies (2008a) calls principles and involve the use of language exams, their fairness and impact (Green, 2014:172). As a result, works related to the history of assessment have been published by scholars such as Spolsky (1995), Davies, (2008a, 2008b) and Weir et al. (2013), as well as guidelines for ethics and fairness such as the International Language Testing Association’s (henceforth ILTA) Code of Ethics (2000) and Guidelines for Practice (2007) (Green, 2014:215). Tsagari (2006:349) reports that since the late 80s there has been a rapid increase in the number of studies conducted on language.

Over the last years the impact of technology, which has always been relevant in language assessment, has become more evident in delivery and scoring. The majority of large scale international tests and many of the exams done in the classroom are now delivered by computer. In terms of scoring, while it has long been used to mark selected response items, it has started to be used to assess written and oral production thanks to Artificial Intelligence. Another major advancement has been the use of multilevel computer adaptive tests such as Linguaskill, which adds more flexibility to proficiency tests. Also related to technology is the established relationship

between applied linguistics and measurement theory (Bachman, 2004; cited by Green, 2018:218), which involves the use of increasingly sophisticated computer-based analysis tools and statistics to describe test scores, improve test material and evaluate test use. From a different perspective, but also highlighting the relationship between technology and assessment is the use of corpora. On the one hand, corpora have been used to design tests for specific purposes (Baker, 2013; cited by Green, 2014:218) and, on the other hand, to have a better understanding of the language that can be produced by learners at different levels and of different backgrounds and the errors they make. An example of this is the English Profile, a research programme by Hawkins and Filipovic (2012; cited by Green, 2014:218), which explored, among other aspects, how participants with different levels of the *Common European Framework of Reference for Languages* used language.

In terms of research, although awareness of the importance of consequences produced by tests is not new and in fact has existed since the beginning of modern testing (Latham, 1877; cited by Cheng, Sun & Ma, 2015:436), over recent decades, the prevalence of large scale, high-stakes testing worldwide (Ungerleider, 2003; Cheng, 2008; Klinger, DeLuca & Miller, 2008; all cited by Cheng, Sun & Ma, 2015:346) and two main movements in language assessment have fostered the increasing attention paid to washback, which, according to Cheng, Sun and Ma (2015:436), has been the preferred area of research over the last twenty years. On the one hand, as reflected in Hughes (1989) the trends in test design towards performance testing aimed to produce test tasks similar to what language learners can face in real life. On the other hand, there was a shift in the perspective of validity, which was now understood to include the use of tests as instruments of social policy (Green, 2013). However, the main focus of research into washback has been teachers and teaching (Cheng, 2008; cited by Cheng, Sun & Ma, 2015:456) and it was not until very recently that empirical studies into washback on learners and learning began to appear in large numbers trying to gain a better understanding of washback and more precisely of the various factors related to the student and to the context. These studies have very often provided evidence of the complexity of washback, an effect that is often mediated by a range of factors (Cheng, Sun & Ma, 2015:459).

2.1.2. Types of assessment and their purposes

There are different categorisations of assessment, the most common one being that distinguishing between formative and summative assessment. However, in this section assessment will be described from different perspectives and, as a result, different categories of assessment will emerge, although a certain overlap will be found.

First, if the frame of reference is considered, two types of assessment can be distinguished i.e. *norm-referenced* and *criterion-referenced*. In the former, the candidates' results are compared with those of a given group. Tests are designed to maximise the distinction among individuals. Conversely, the results obtained from the latter are to be interpreted with reference to a level of ability, for instance the levels of the *Common European Framework of Reference for Languages* (Council of Europe, 2001 and 2018). As a consequence, tasks in criterion-referenced exams need to elicit performance at these ability levels (Bachman, 1990). So, an example of a task could be listening to a lecture and the objective to assess candidates' ability to understand a lecture.

If we look at the type of data obtained, Green (2014:12) distinguishes between *educational* assessment and *proficiency* assessment. *Educational* assessment focuses on teaching and learning and allows for more flexible approaches such as observation, the use of portfolios, and the administration of informal tests as well as more formal tests administered under more controlled conditions (Green, 2014:13). In contrast, *proficiency* assessment aims to obtain data about what a candidate can do – following the CEFR Can-Do Statements – (Council of Europe, 2001 and 2018) or accomplish through their use of language. To some extent, results in proficiency tests can help to predict how candidates will behave in a future situation of language use (McNamara, 2000), they look forward and, in this sense, they have often been associated with gate-keeping decisions. This is the case of B1 Preliminary, which is accepted by many universities in Spain to certify the language level required to graduate, or of International English Language Testing System (henceforth IELTS), which is used for immigration purposes in many countries. They are often produced by national governments, national and international assessment institutions such as Educational Testing Service (henceforth ETS), Cambridge Assessment English or language schools such as Escuela Oficial de Idiomas or Asociación de Centros de Lenguas en la Enseñanza Superior (henceforth ACLES). In terms of the language domain, it can be general or focused on a specific language domain such as legal English or business English. As Green (2014:16) points out, educational and proficiency

assessment are often perceived as totally different areas, with teachers mainly in charge of the former, and governments and external assessment institutions in charge of the latter. Nevertheless, they share the same aim, i.e. to provide useful evidence of learners' abilities.

Within educational assessment, different agents can take part in the assessment process. Probably the one that first comes to mind is the teacher and teacher-led assessment. It can take the form of tests, observation, portfolios, etc. However, self-assessment and peer assessment have emerged as powerful means of engaging students in the learning process and making them more autonomous learners. Devoting time to talking about the learning objectives – more in the long term – in a way that students can fully understand and explaining how they and their teacher are going to attain them – more in the short term – gives students a sense of direction. In addition, familiarising students with the assessment criteria and methods and giving them the opportunity to take an active role and use them for peer assessment and self-assessment – with checklists written down to level so that students can fully understand them – encourages a reflective attitude, which helps students identify their strengths and weaknesses and understand how they can improve and attain their goals.

If the variable considered is when the assessment takes place, we can distinguish between *prognostic*, *aptitude*, *progress* and *exit* tests (Green, 2014:14). *Prognostic* assessment can be very useful as a placement tool. Very often language schools receive students coming from different backgrounds and with different levels of ability so it is useful to find out how much of the course content learners are already familiar with and to what extent before assigning them to the classes available. *Aptitude* tests can offer guidance in terms of the potential ability for learning a language. *Progress* tests can be valuable as a diagnostic tool for both teachers and learners because they can point at aspects that have been taken in successfully and at the areas that need to be consolidated. They can help teachers to plan the rest of the academic year more accurately and realistically and they can also guide students and make them aware of what content they are comfortable with and the aspects that they need to work harder on. Finally, *exit* tests are frequently used to decide who passes or fails the course on the basis of the extent to which learners have achieved the learning goals but can also be used to offer feedback that can inform independent learning.

As already mentioned, the most frequent categorisation of assessment is the one distinguishing between *formative* assessment, also called assessment *for* learning, and *summative* assessment, also referred to as assessment *of* learning (Black & Wiliam, 1998; cited by Green, 2014:14). Clearly, progress tests can be identified as part of the more formative uses of assessment since the results obtained are likely to have an impact on the teaching and learning process and, hopefully, help teachers and learners make informed decisions about what to do next or how to do it. Although exit tests could be associated with more summative assessment, as they will show what students have learnt, it is important to emphasise the need for including feedback that can make students aware of what they are doing well and what aspects they need to improve so that they can become more independent learners.

From a different perspective, Fulcher (2010; cited by Muñoz, Véliz-Campos & Véliz, 2019:99) distinguishes two main types of assessment: *traditional* and *alternative* assessment. *Traditional* assessment is designed to yield objective and measurable data (Muñoz, Véliz-Campos & Véliz, 2019:102) while the focus of *alternative* assessment is mainly on eliciting meaningful communication and offering feedback to students (Brown, 2004; cited by Muñoz, Véliz-Campos & Véliz, 2019:99). Traditional assessment is associated with high-stakes exams and their most deleterious effects on teaching and learning (Muñoz, Véliz-Campos & Véliz, 2019:99). However, alternative assessment is perceived by authors such as Petre (2017; cited by Muñoz, Véliz-Campos & Véliz, 2019:99) and by McNamara (2000) as more innovative movement and a motivating, tool that allows students to apply the skills developed and the knowledge gained exploiting self-assessment and weaving a constructive relationship between teaching and learning. Although some governments and institutions may opt for one or the other, Muñoz, Véliz-Campos and Véliz (2019:99) support the view already expressed in relation to other types of assessment that traditional and alternative assessment should not be conceived as conflicting forms of assessment but as complementary.

While there is a tendency to see different forms of assessment as incompatible, several scholars – Green (2014) or Muñoz, Véliz-Campos and Véliz (2019) – have emphasised the opportunity of using different types of assessment to complement each other. This is the idea behind Cambridge Assessment English model for Learning Oriented Assessment (LOA). LOA is an innovative approach based on the premise that all levels of assessment – the macro level, which sets educational goals and evaluates outcomes, and the micro level of individual learning – can and

should contribute to both the effectiveness of learning and the reliable evaluation of results. What is new about the concept of LOA is that it combines formative and summative assessment and presents a structured approach to gathering and using evidence from test and from the classroom. This evidence is then used to identify learning needs, evaluate progress and help make decisions to promote continued learning (University of Cambridge Local Examinations Syndicate, 2020a). LOA encompasses *quantitative* measurement, which is the domain of assessment expertise, and *qualitative* individualisation of learners, which is the domain of teaching expertise, establishing a complementary relationship between assessing and teaching. LOA acknowledges that teachers play a key role when it comes to creating an environment which is productive for learning and perceives their work as complementary to the role of more formal forms of assessment (University of Cambridge Local Examinations Syndicate, 2020a).

2.1.3. Test Qualities

For a test to fulfil its purpose and be useful it should meet a set of requirements or test qualities i.e. validity, reliability, practicality, authenticity, impact, interactiveness (Bachman & Palmer, 1996) and quality of service (Association of Language Testers in Europe, 2001). If a suitable balance among test qualities is reached, difficult as it may be, a test can be said to be useful and fair. A fair test is a test in which the test construct is the most important focus and where all irrelevant obstacles that can affect candidate's performance have been removed. The principles described below are crucial for any organisation since accreditation tests are high-stakes and could have life-changing consequences.

Validity has to do with the degree to which evidence and theory support the interpretations of test results entailed by proposed uses of tests (American Educational Research Association et al., 1999; cited by Green, 2014:75). With this definition, Green emphasises that validity is linked to the purpose of the test results in the sense that validity is not properly thought of as a quality of the assessment as such, but rather as a quality of the interpretations that users make of assessment results. Bachman and Palmer (1996) define validity as the extent to which tests scores reflect accurately the candidates' level of language ability. Salkind (2000) refers to the same concept but in very simple terms when claiming that if a test does what it says it does, then the test scores have validity. Fulcher and Davidson (2007; cited by Green, 2014b), among others, consider validity as the most important aspect in test evaluation as it refers to the appropriateness – how well the content

matches the target population, meaningfulness – how transparent the scores are – and usefulness – how easily the test scores can be used – of the test scores. The Council of Europe (2001:177; cited by Cruz Trapero 2016:53) refer to the same idea when stating that an exam has validity to the degree that it can be proved that the construct is what, context concerned, should be assessed, and that the information obtained is an accurate reflection of the proficiency of a given candidate. Validity can be seen from different perspectives and, as a consequence, different types of validities emerge. The first one that is going to be discussed is *construct validity*, which embraces all forms of validity evidence and refers to the "meaningfulness and appropriateness of the interpretations that we make on the basis of test" scores (Bachman & Palmer, 1996). In other words, construct validity is based on the theory on which the test is based i.e. if we are talking about a listening test this theory is our definition of listening and how this theory is transformed into a test (Green, 2014b). Validity, in this case *content validity*, also has to do with the materials used in the test and to what extent they represent the full range of knowledge, skills and abilities that the assessment should cover (Green, 2014:78). *Context validity* (Galaczi & Ffrench, 2011) focuses on the test tasks and how accurately they reflect the essential characteristics of tasks in the target language domain. *Cognitive validity* (Field, 2011) refers to the cognitive processes engaged by candidates in undertaking test tasks and to what extent they are similar to the processes that language users engage with when they carry out tasks in the Target Language Use (henceforth TLU) domain. The way non-expert stakeholders, i.e. candidates, parents and employers, perceive the test is what scholars refer to as the *face validity* of the test (Green, 2014:79). In this sense, students' perception of the test plays, according to Watanabe (cited by Tsagari, 2006:55), a major role in the presence or absence of washback. Validity also has to do with the extent to which the results of a test are compatible with other evidence of the candidates' knowledge, skills or abilities. Depending on the author this type of validity is called *criterion-related* or *criterion-oriented validity*, as it is measured against an external criterion, (Green, 2014:96) following Weir (2005) and Cronbach and Meehl (1955) or *external validity* since it involves linking tests to predictions on external factors (Cruz Trapero, 2016:52). It encompasses *concurrent* and *predictive validity*. *Concurrent* validity has to do with how candidates' scores on one test relate to those on another (Green, 2014b). This other test could be a previous version of the same test, e.g. B2 First before and after the 2015 revision, or two tests produced by different institutions e.g. B2 First and the B2 test exam produced by Escuela Oficial de Idiomas. *Predictive* validity refers to the relationship between the scores on a given test

and the ability of the candidates to perform a certain activity in real life (Green, 2014b) e.g. Linguaskill for Business and candidates' ability to carry out work related tasks in an English-speaking environment. Finally, *consequential validity* (Hawkey, 2011) refers to the use of tests to maximise beneficial consequences and minimise harmful effects (Weir, 2005; cited by Green 2014:96). Some scholars such as Messick (1996) perceive consequential validity and washback as closely intertwined even placing washback within the concept of validity and arguing for the need to embed validity considerations in test design in order to promote positive washback (Cheng & DeLuca, 2011:106). Morrow (1986), Frederiksen and Collins (1989), Weir (1990), and Shohamy et al. (1996), all cited by Tsagari (2006:19) support Messick's view that the effect of a test on teaching and learning is a major aspect of validity. Messick goes further and promotes the examination of the two main threats to test validity, i.e. construct under representation and construct-irrelevant variance, to determine the potential consequences that a test can have on teaching and learning (Tsagari, 2006:350). Nevertheless, other scholars such as Ferman (2004:245; cited by Tsagari, 2006:19) and Davies (1997:335) support the opposite view and argue that validity is not a property of the test as such, but rather the meaning of the test scores and, therefore, do not see a direct connection between washback and validity. Bailey (1996:68; cited by Tsagari, 2007:351) also criticises Messick's position arguing that any test, whether good or bad in terms of validity, can have negative or positive washback depending on whether it promotes or impedes the achievement of the learning goals defined by the students or the teachers. With this statement, Bailey points out that washback might be different for different groups of stakeholders. This view is shared by Burrows' (2004) study, which supported the washback hypothesis by Alderson and Hamp-Lyons (1996:282; cited by Cheng, Sun & Ma, 2015:456) that tests will have washback effects for some teachers and some learners but not for others. Alderson and Wall (1993:116) justify their position on the basis that while validity is a property of a test in relation to its use, washback is likely to be a complex phenomenon which cannot be related directly to a test's validity. In relation to this, Messick (1996:142; cited by Tsagari, 2007:19) points out that the relationship between washback and validity is based on the premise that it can be proved that washback is an effect of the test only and not of other forces operative on the education scene.

Authenticity is closely related to validity, and more precisely to content validity, as it has to do with how relevant the task is to the TLU domain (Bachman & Palmer, 1996). A test is said to be authentic if performance in that test corresponds to language use in specific domains other than the language test itself.

Reliability refers to whether the test results are stable, consistent and free from errors of measurement (Association of Language Testers in Europe, 2001). The Council of Europe (2001:177; cited by Cruz Trapero, 2016:53) defines it as the extent to which the same rank order of candidates is replicated in two separate administrations of the assessment. Henning's (1987; cited by Green, 2014b) definition encompasses all the factors that come into play when talking about reliability: it is a measure of accuracy – is the candidate really B2?, consistency, dependability – will the candidate be B2 tomorrow and if not why? is it because of the rater, the rating scale or the candidate's performance?, or fairness of scores – is the score a fair reflection of the candidate's ability? – resulting from the administration of a particular examination. Reliability contributes to overall validity and, as pointed out by Cruz Trapero (2016:54), the higher the stakes, the more important reliability becomes. As reflected in Henning's definition, reliability concerns not only the tools used to test i.e. the quality of the tasks, the number of tasks, the choice of tasks but also, especially in productive skills, the raters, the rating scale, the training the raters have received, the number of raters involved, and the time allowed to carry out the rating (Green, 2014b).

Equally important nowadays is *practicality*, which has to do with the relationship between the resources that the assessment institution has available and the resources necessary to produce and administer the test in its intended context of use (Association of Language Testers in Europe, 2001 and Bachman & Palmer, 1996).

The fourth quality is the *impact* that accreditation exams have on education and on society in general (Association of Language Testers in Europe, 2001). Institutions and associations such as ALTE try to promote a better understanding of the objectives and procedures of testing and the proper uses of exam results. The beneficial impact of accreditation exams on teaching and learning is another crucial element and that is why research on washback, which has only attracted attention from writers and researchers from the early 90s (Tsagari, 2007:13), and on face validity is so relevant.

Last, we should mention *interactiveness* i.e. the degree and type of involvement of the test taker's individual characteristics in the accomplishment of a task. The individual characteristics that are most relevant for language testing are the candidate's linguistic ability, topical knowledge and affective schemata (Bachman & Palmer, 1996; cited by Peña Jaenes, 2015:16).

Finally, *quality of service* is crucial to round off the whole testing process. It involves the provision of secure examination materials, the confidentiality of exam data and results, and procedures to handle enquiries about results and appeal procedures (Association of Language Testers in Europe, 2001). To meet the requirements of quality of service, examination institutions try to avoid any content and language that could be insensitive and to adapt exams and administrative procedures for candidates with special requirements. Besides, assessment institutions should provide test takes with all the necessary information to be familiar with exam format and assessment criteria.

When the appropriate balance among all test qualities is achieved, there is fairness in the construction of the test, in its impact, in its administration, and in its evaluation. In the following sections, we will see how the different test qualities come into play to assess all four skills.

2.1.4. Assessing speaking

Speaking as a skill can be considered from two different perspectives which have to do with the modes of communication (Council of Europe, 2018). On the one hand, speaking can be seen as oral production and, in this sense, it can be said to be more of an individual activity. On the other hand, speaking can be seen as interaction, so more of a social activity, where each individual is both a speaker and a listener and where participants construct the communicative situation together and share the right to influence the results (Luoma, 2004:20). Such distinction, which can be regarded more as a continuum (Green, 2014:128) than as a clear division, has an impact on the processing of information, which in turn, impacts linguistic aspects. In the next lines, oral language is going to be analysed considering the mental processes involved, its features and the consequences for the assessment of speaking.

2.1.4.1. Mental processes

To describe language processing Cutler (2005) compared productive language processing and receptive processing and observed that they mirror each other. In production, the speaker or writer have a message to convey, and to do it they need to put this message into an appropriate physical form as can be seen in Green's model of language production (2014:128) (Figure 1).

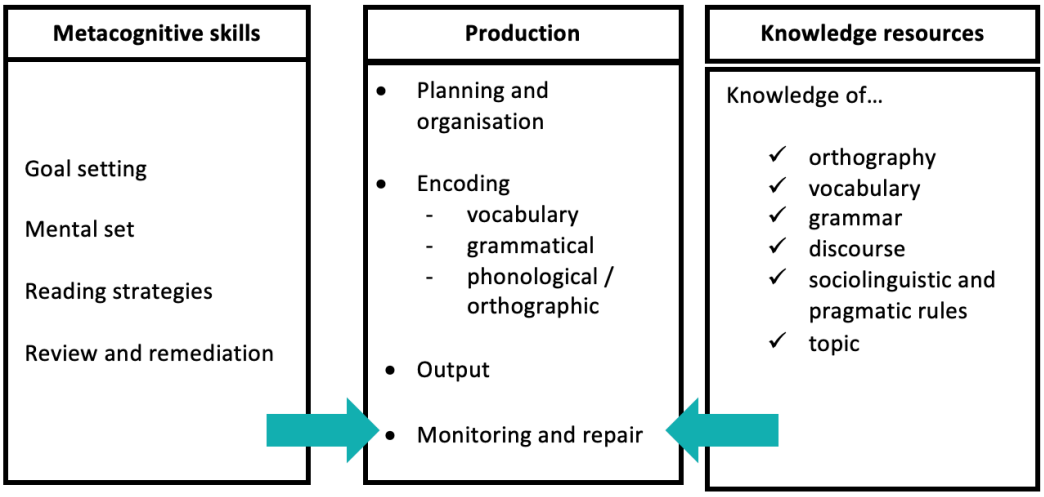


Figure 1. A simplified model of language production based on Levelt (1989), Chenoweth & Hayes (2001), cited by Green (2014:128)

In this model the speaker, like the writer, reader or listener, has a communication objective. This objective is usually represented by the message that they wish to convey. The speaker draws on a mental store of lexis and grammatical structures and on knowledge of discourse, encodes the message and transmits it to the addressee as auditory or visual input (Green, 2014:128). The speaker – and the writer – will decide on how best to express the message in mind in order to achieve their goal. As a result, they will need to use their background information about the topic, about the listener or reader, about the communicative situation and about the types of speech that are suitable for the situation (Green, 2014:128).

While the mental processes are common to the most productive side and the most interactive side of speaking, the time each participant of a conversation has to process the information as a listener and react to it is much shorter and this, as we are going to see, impacts on the linguistic aspects of speaking.

2.1.4.2. Language, language functions and the communicative situation

The language we use depends on the communicative situation where speaking takes place and on the communicative goal in mind. Cummins (1979; cited by Green, 2014:128) distinguishes between Basic Interpersonal Communication Skills (henceforth BICS) and Cognitive Academic Language Proficiency (henceforth CALP) and points out that while most native speakers of a language are proficient in BICS, CALP requires education. Focusing on the functions that the speaker wants to perform, Bygate (1987) divides speech into two main types i.e. *factually oriented talk* and *evaluative talk*. The former has the purpose of conveying information while the latter involves expressing a point of view towards the content. Brown and Yule (1983) categorise language into *interactional* language, which aims to build and maintain social relationships, and *transactional* language, which conveys information. Depending on the function we want to perform and the situation in which it will take place, speakers follow a sequence. Sequences differ from one language and culture to another so language learners need to become familiar with the different sequences used for the different functions and situations.

Language learners' progress is often measured in terms of the the range of grammatical structures and lexis which they can use accurately. There are situations where speakers can prepare what they are going to say in advance, for instance when researchers give a presentation at a conference, they will probably prepare it carefully and even rehearse it several times. In these situations, speakers produce *planned speech* (Ochs, 1979; cited by Luoma, 2004:12), which is characterised by more complex language. Nevertheless, if we move along this continuum to less prepared communicative situations where *unplanned speech* is produced, greater differences between speaking and writing can be observed in terms of the complexity and organisation of language. Luoma (2004:12) points out that speakers speak in *idea units,* i.e. short phrases and clauses connected by simple conjunctions or just spoken next to each other. Unplanned speech is often characterised by shorter idea units, especially when compared with a written text because speakers know that their listeners are trying to understand ideas in real time. Also, Luoma

(2004:17) observes that the vocabulary of spoken interaction is usually quite generic and, when the situation is informal, language is often marked by references to the immediate physical environment. Another feature of oral production and interaction is the need to use strategies to obtain thinking time. This can be done through words and phrases, also known as fillers or hesitation markers. Repetition is also used for the same purposes. These resources are widely used by native speakers and research shows that when language learners use them, the listener usually interprets it as a proof of higher level of ability (Towell et al., 1996; cited by Luoma, 2004:18) and of greater fluency (Hasselgren, 1998; cited by Luoma, 2004:19). However, candidates may sometimes be penalised for using them. This could be the case for instance of repetition, which can sometimes be interpreted as a sign of a limited range of vocabulary. Hence, assessment criteria need to be flexible enough to recognise repetition as a way to obtain thinking time or as form of cohesion. Apart from the differences in grammatical and lexical complexity, a progression in register can be observed from more formal language, typically used in planned speech, to more informal register, as the speech becomes more unplanned.

As we have seen, the situation and the mode of communication influence how we encode our message but they are also crucial when organising it. Organisation is important at the level of the individual sentence or speech unit and at the level of the text or unfolding interaction. It includes deciding what information is already known and what is new. New information is usually given prominence through stress or is put at the end of the sentence. Information already shared is often omitted (Halliday & Hassan, 1976; cited by Green, 2014:129). Luoma (2004:15) adds "two structures which clearly belong to spoken-like language use": *topicalisation*, which gives special information emphasis to the initial element of a clause in informal speech, and *tails*, which are noun phrases that come at the end of a clause.

The aspect that is probably the least impacted by the communicative situation although, as stated by Luoma (2004:10), it is one of the first aspects one notices when one hears someone speak, is pronunciation. On the basis of what listeners hear, they make some tentative and possibly subconscious judgements about the speaker's personality, attitudes, home region and native / non-native speaker status (Luoma, 2004:10). In the field of language learning and teaching, students need to articulate the different sounds of the foreign language and take into account and reproduce as far as possible aspects such as stress, intonation and pitch (Green, 2014:131). In the field of assessment, pronunciation has been a controversial topic for assessment experts.

Finally, one of the key features of the nature of speaking is that speakers can often see their audience or interlocutor's reaction and modify their message as a result of feedback signalling that the communicative goal has not been met. This modification may include pronunciation, grammar, vocabulary, or register, among others.

2.1.4.3. Implications for assessment

The nature of speaking, briefly described above, has implications for assessment. In terms of marking, McNamara (1996; cited by Green, 2014:136) distinguishes between a strong and a weak sense of performance assessment. Strong sense considers whether the task has been fulfilled e.g. buying a certain product. The weak sense analyses the language used to buy the product rather than whether the task was fulfilled or not. Assessment institutions often include a combination of both of McNamara's approaches and assess the global achievement of the candidate as well as other criteria, such as language, organisation and pronunciation. The sound of speech has been a controversial topic for language assessment. As we saw before, people tend to decide whether a person is native or not on the basis of pronunciation, which leads to the idea that this is the standard that should be used to assess learner pronunciation. This poses problems in terms of which native accent to consider as criterion since English, like other languages, is spoken as a first language by people from different countries, who have very different accents. In addition, research into learner language evidences that very few learners are capable of achieving a native-like standard in all respects, so most learners would not fulfil this requirement. As a result, it is widely acknowledged that communicative effectiveness seems to be a better standard for learner pronunciation (Luoma, 2004:11). Considering lexis, when assessing language, the descriptions of vocabulary use at higher levels often expect candidates to use a wide range of lexis with "accuracy" and "ease". However, Luoma (2004:16) points out that while this precision and richness can be relevant in professional contexts, generic, simple and ordinary words are also very frequent in normal spoken discourse, and being able to use them naturally in speech also evidences advanced speaking skills. This aspect of word use should also be rewarded in speaking assessment. In fact, incorporating the use of generic words in the rating scales may produce positive washback as it may make students aware of the fact that generic words are an important feature of natural talk.

If we look at the tasks types, depending on the construct of the test, test designers should design tasks to elicit planned and unplanned speech considering aspects such as planning time and speaker roles and role relationships. Speaking tests usually assess two – or even three – modes of communication (Council of Europe, 2018) because production, interaction and even mediation are elicited by the tasks. The way to approach each mode of communication creates different challenges but also allows for a wider variety of output. He and Young (1998:1; cited by Green, 2014:135) argue that to assess candidates' speaking abilities, examiners have to get them to speak. Although scholars such as Green (2014:138) acknowledge the value of controlled tasks such as gap-fill, ordering and read-aloud task types as indicators of productive abilities and general proficiency, they also point out that these more controlled task types are most useful for specific types of assessment such as placement tests or at early stages of language learning. At present, assessment institutions follow He and Young and rely on the use of performance assessment. Even if this choice poses problems for tests reliability since extended performance is without a doubt less predictable, which makes marking more difficult. In an attempt to reduce variability, tasks can be controlled using recordings or scripts but this may in turn compromise the authenticity of the task. As usual, a balance needs to be reached.

When assessing speaking in a conversation, several approaches can be followed. On the one hand, candidates may interact with the examiner. This is probably the most controlled option because examiners are trained to minimise variation from one exam to another and they often follow a script (Fulcher, 2003; cited by Green, 2014:13). However, the downside is that the conversation produced tends to be very one-sided as the examiner is the one mainly asking questions and the candidates answers them. This type of interaction compromises the authenticity of the task because it does not reflect everyday conversations, where language users have a more balanced relationship and responsibility in directing the interaction (Green, 2014: 139). On the other hand, candidates can work in pairs or groups and interact with one another. This option allows for a wider range of talk as candidates ideally initiate, take turns and can move the conversation forward (Swain, 2001, cited by Green, 2014:140). What is more, this type of tasks is also more practical because more candidates can be assessed in less time. Finally, they bring about positive washback in the sense that they may lead to more interactive classroom activities (Green, 2014:140). The main problem this task type poses is that the candidate's experience will differ depending on the language level and even the personality of their partner (Green, 2014:140).

Probably in an attempt to overcome this problem, assessment criteria used by institutions such as Cambridge Assessment English are designed to individualise scores so that candidates are not impacted by their partner's performance. Nevertheless, this has been criticised by scholars such as He and Young, who express their concern about the validity of interactive speaking tests because scores are awarded to the individual candidate although talk is the result of shared interaction (Green, 2014:132 and Luoma, 2004). A further challenge for assessment designers posed by both task types is that the meaning of any sentence is not universal and, therefore, can vary greatly depending on the situation (Green, 2014:132).

2.1.5. Assessing writing

Writing and speaking share some common features but also differ greatly. Speaking in our mother tongue is a skill we develop from an early age in a natural way. Writing, on the contrary, is a skill one develops more consciously both in the mother tongue and in a foreign language.

2.1.5.1. The writing process

Writing and speaking are very similar in terms of language processing, which, as we saw before, mirrors receptive processing in many ways (Cutler 2005; cited by Green, 2014:127) because the writer has an idea that s/he wants to convey and needs to put it into words to achieve their communicative goal effectively so s/he will have to consider the aspects of the communicative situation. A similar perception, but in this case adapted to the assessment context, is found in the Cambridge English Model of Writing (University of Cambridge Local Examinations Syndicate, 2020d) (Figure 2). From an assessment point of view, everything starts with the message, i.e. the writing task that candidates produce at the end. Once this is clear, the cognitive aspect activates and candidates come up with ideas for that specific task. This is what raters will analyse as content. After that, candidates need to put their ideas into words and this is when the linguistic aspect of the process is activated. However, in order to choose the right words, candidates need to take into account who they are writing to, i.e. the target reader; and why they are writing, i.e. the purpose and the type of language that they need as a result of this. This is the sociolinguistic component of the model and includes aspects such as being aware of the conventions of text type. All these aspects are intertwined in the process of producing a piece of writing.

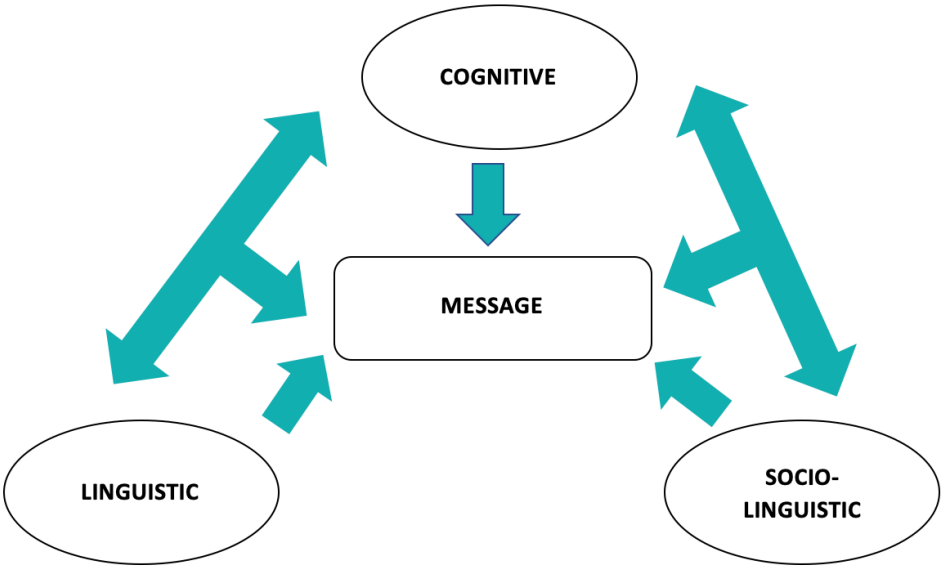


Figure 2. Cambridge English Model of Writing (University of Cambridge Local Examinations Syndicate, 2020d)

In writing there is usually more opportunity for planning and this allows writers to produce more complex language, which includes using a wider range of complex grammatical structures, such as subordinate clauses, but also eliminating redundancy. Organisation is also influenced and writers have longer planning time to organise their sentences, paragraphs and texts (Green, 2014:130). The nature of writing impacts on the lexis too. Since there is usually distance – both hysical and social – between the writer and the reader, the writer needs to make their language more specific as references may not necessarily be shared, and this leads to a higher proportion of lexical to grammatical structures (Manchón et al., 2005:383; cited by Peña Jaenes, 2015:22). Besides, writers need to read their piece of writing from a distance i.e. from the perspective of the target reader (Manchón et al., 2005:383) to try to predict and solve potential communication problems that may arise. They also have to rely on visual elements such as punctuation or typeface to offer some extra paralinguistic information (Green, 2014:131). As Manchón et al. (2005:383) point out, spelling can also be an additional difficulty in the case of English, as it does not match pronunciation.

In addition to planning, Manchón et al. (2005:383; cited by Peña Jaenes, 2015:22) distinguish two more stages in the writing process: formulation and revision. Formulation is the stage that receives the greatest attention on the part of students. In fact, teachers and experts in assessment

often criticise the little importance learners give to planning and revising. Revision has received increasing attention and has been perceived as an opportunity for peer assessment and self-assessment. It has also been identified as a vital stage in process writing to produce a good text because it allows the writer to fine-tune all the above-mentioned aspects.

When analysing the nature of writing, it is clear that planning, formulating and revising are paramount. They interact with one another in a recursive manner (Manchón et al., 2005:387; cited by Peña Jaenes, 2015:22) and the ability to use them adequately is not innate.

2.1.5.2. Implications for assessment

Hughes (2003:75) argues that the best way to test candidates' writing performance is to get them to write, and Hamp-Lyons (1991; cited by University of Cambridge Local Examinations Syndicate, 2020d) seems to agree since she defines writing using three parameters. First, the length: writing is defined as a piece of continuous written text of 100 words or more. However, this length would not be suitable for all candidates because we cannot expect an A2 student to produce such a long text and we would probably expect a C1 candidate to produce a much longer text with ease. Second, Hamp-Lyons explains that a candidate needs enough freedom to create a response to a given stimulus. The keyword here is *enough* because, as we also saw with speaking tasks, we want to assess all the pieces of writing against a common yardstick so there must be specifications that make the candidates' writing production comparable. Grammar and vocabulary exercises (Green, 2014:135) have proved useful to predict oral and written production and this is why there is room for Use of English papers in English language exams. Moreover, these items are more practical because they are less time-consuming to administer and mark, and also more objective (Haertel 1999; cited by Green, 2007:18). Nevertheless, most assessment institutions include a writing paper where candidates need to write a longer piece of writing in reply to input because it is considered more authentic and direct, which can lead to positive washback (Messick, 1996; cited by Peña Jaenes, 2015:24), despite the challenges it poses in terms of time and reliability. The last aspect Hamp-Lyons includes in her definition of writing has already been mentioned and has to do with the criteria used. Assessment criteria must be the same for all the candidates taking the exam and raters have to be trained to apply them reliably.

2.1.6. Assessing reading

The purpose of assessing reading is to determine how well candidates can use their language abilities to understand written input (Green, 2014:97). To develop reading tests, it is vital to have an understanding of the nature of reading (Alderson, 2000:31) and, as we are going to see in the next lines, there is a number of models and theories to explain the nature of reading, each of them with implications for testing and assessment.

When one thinks of assessing receptive skills, probably the main challenge that comes to mind is the fact that they cannot be observed directly, which means that evidence of a language learner's reading or listening ability can only be obtained by asking them to do something else in order to show how well they have understood (Green, 2014:97). One of the most frequent distinctions when it comes to reading is the *process* and the *product* of reading (Alderson, 2000:3). When talking about *process* we refer to the interaction between a reader and a text, which involves looking at print, deciphering what is on the page or the screen, deciding on meaning, thinking about the meaning of the text and how it relates to their previous knowledge, etc. If we want to assess reading on the basis of the end of the process, we will compare the product with the text originally read (Alderson, 2000:4).

2.1.6.1. Types of reading

Urquhart and Weir (1998) categorise types of reading using two dimensions: time and scope. Khalifa and Weir (2009; cited by Green, 2014:99) focus on the time variable and distinguish between *expeditious* and *careful* reading. The former refers to reading that is quick, selective and efficient. This is the type of reading we do when we are flicking through pages in a magazine to find something that interests us or when we are trying to find a good restaurant and we are reading different reviews. In contrast to fast reading, people often engage in careful reading. This type of reading is linear because the reader starts at the beginning and follows the structure of the ideas and it is not selective because the reader tries to take in all or the majority of the information in the text (Green, 2014:100). This type of reading is necessary when you are following instructions or when you are studying something to fully understand it.

Khalifa and Weir (2009) further distinguish different types of expeditious reading: *scanning*, *search reading* and *skimming*. When scanning, the reader is trying to find specific words or data. This is what we do when we are trying to find the time when the film starts at our favourite cinema. Search reading is used when we are trying to locate an idea about a certain topic, for instance when a PhD student is trying to find the paragraph discussing washback in a research paper. Finally, skimming is used when we have a very long text and we want to obtain the gist of it, i.e. an effective general understanding of the complete text. This is what we do when we are reading a newspaper article to have an overall idea of the situation but we are not really interested in the details. If we go back to Urquhart and Weir's (1998) dimensions and we consider the scope of reading, scanning and search reading have a more local focus while skimming has a more global focus (Green, 2014:101).

Skilled readers decide on the best type of reading depending on the text type and their objective. For language learners, this is an ability that is developed consciously and that requires practice. In the field of teaching, it is commonly believed that there is a hierarchy of types of reading. At lower levels, much of the learners' attention is focused on decoding. This process should become automatic before we can expect learners to read quickly. Therefore, fast reading can be introduced to classroom and assessment activities at an intermediate level and should receive a more intense focus from B2 level upwards. Moreover, it should be taken into account that some fast reading abilities are more challenging than others, so it can be considered more appropriate to introduce scanning at B1 level and then move on to skimming and search reading at B2. At advanced levels, students can be expected to have awareness of text structure and they are fully prepared to choose between different types of reading or combining them as appropriate (University of Cambridge Local Examinations Syndicate, 2020e). Depending on the type of reading that the reader considers most appropriate in a given situation, they will engage in different mental processes. This ability to adapt to the text type and the task objective is one of the problems candidates face in reading tests. They often complain about the little time they are given to read a text and answer the questions. However, assessment designers make decisions regarding text length and time allotted to test candidates' ability to make the right choice in terms of the reading subskill they need to use.

2.1.6.2. Mental processes

When considering the mental processes, a recent and widespread perception is that reading is divided into two areas: *decoding* and *comprehension*. Decoding has to do with recognising words and comprehension encompasses parsing sentences, understanding sentences in discourse, building a discourse structure, and integrating this understanding of the text with one's previous knowledge (Alderson, 2000:12). This view brings together reading and listening, as comprehension is said to be centrally-determined and does not depend on how the input is presented, i.e. orally or in writing (Larsen & Feder, 1940:251; cited by Alderson, 2000:12).

Green's (2014:101) understanding of reading is based on the work of Weir (2005a) and Field (2008) and combines metacognitive skills and language knowledge to contribute to the reception process (Figure 3 below). At lower levels, readers' mental capacity is mainly focused on the local decoding of words and the grammatical parsing of each clause or sentence (Green, 2014:100). This is what Gray (1960; cited by Alderson, 2000:7) calls *reading the lines*. As they climb the language ladder, usually at B1, language learners learn to understand implied meaning or to *read between the lines* (Gray, 1960; cited by Alderson, 2000:7), and to understand information across sentences or paragraphs – usually at B2 – and across a whole text – usually at B1 and B2. Finally, learners are ready to understand and combine information across different texts and make a critical evaluation of the text. This last stage is what Gray (1960, cited by Alderson, 2000:7) calls *reading beyond the lines*. While it is believed that the progression in the ability language learners have to process information outlined above is an accurate representation of how reading comprehension improves, relevant scholars such as Alderson (2000:8) point out that there is little evidence supporting it. In fact, Alderson (1990c; cited by Alderson, 2000:11) states that the reading process involves "the simultaneous and variable use of different, and overlapping skills".

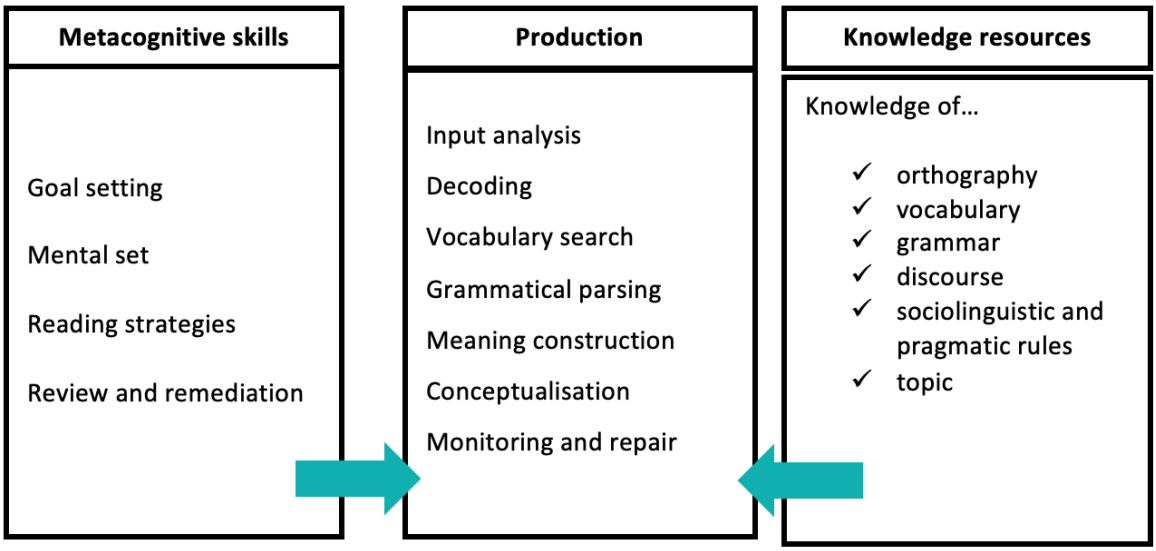


Figure 3. An outline model of reading language processing based on Green (2014:101), Weir (2005a) and Field (2008)

Over the last twenty-five years, the idea that readers engage in bottom-up and top-down approaches has gained momentum. When readers follow a bottom-up approach, they start with the printed word, recognising it and decoding its meaning. However, when they use a top-down approach they activate their schemata to use the information in the text and interpret it. Alderson, however, rejects this way of understanding reading as he claims that the most accurate model is the one that acknowledges that reading components interact with each other. He goes on to quote Grabe (1991:384) suggesting that processing is thought to be parallel rather than serial.

2.1.6.3. Implications for assessment

The views towards reading discussed above have implications for the assessment of reading. If we look at the product and process approach, it seems obvious that understanding how the reading process takes place poses challenges. Research has relied on eye movement, reading aloud, think-aloud protocols or verbal retrospection interviews as means of externalising this process. If we opt for analysing the product, scholars have tried to compare it with the original input. However, this poses problems in terms of what understanding readers reach. This is because different readers may reach different understandings of what a text *means*, since meaning is created in the interaction between a reader and a text (Alderson, 2000:6).

When looking at reading as a series of strategies, Alderson (2000:15) explains that there are two options available. On the one hand, test designers may try to isolate the different components of reading, select the ones considered suitable for a given level, and test whether readers have successfully engaged with them. On the other hand, test designers may try to simulate as much as possible the conditions considered suitable and assess whether the reader has successfully completed the task (Alderson, 2000:16).

Finally, Alderson (2000:20) points out that it would be difficult to imagine a reading test that would require candidates to adopt either a bottom-up or a top-down approach as it is most likely that there will be an interaction between the text and the reader's background knowledge.

2.1.7. Assessing listening

Listening comprehension has received increasing attention in the recent past as a result of the theoretical developments in language learning and teaching and the emergence of English as a Foreign Language (henceforth EFL) courses. Before that, listening was considered, together with reading, as a secondary skill when compared with productive skills (Nunan, 2002:238) mainly because it was thought to be a passive activity (Jung, 2003; Vandergrift, 2004; cited by Dhanapala, 2019:1) and a skill that could only be acquired by exposure to the language and not taught (Dhanapala, 2019:1). Thanks to the work of scholars such as Vandergrift (1999), it is clear that listening is far from being a passive activity since when we listen we activate a number of processes, which pose heavy cognitive demands (Rubin, 1995; cited by Dhanapala, 2019:1).

The purpose of assessing listening is to determine how well candidates can use their language abilities to understand spoken input (Green, 2014:97). This will often be recordings in the target language reflecting what learners may need to deal with in real life, as part of a language learning programme (Green, 2014:97) or in the workplace because, while listening is just as important as the other skills in our daily life, it can be a crucial skill in contexts such as the workplace (Brownell, 2002) or the academic context (Suchithra et al., 2014:476).

One of the main obstacles faced by listening research is that listening is a cognitive process and therefore difficult to observe (Takei, 2002; cited by Suchithra et al., 2014:476). Traditionally, research into listening focused on the product of listening and the performance of listeners. However, scholars such as Buck (2001) or Rost (2002) pointed out the limitations of studying

listening from the product perspective, claiming that assumptions about the listening process cannot be made relying only on the features of the product produced during the listening because the characteristics of this product can be altered by a different level of ability in productive skills. As a result, there is a growing interest in using verbal protocols as a means of investigating the nature of listening and in using the products of listening as an aid (Smith & King, 2013; cited by Suchithra et al., 2014:476).

2.1.7.1. Types of listening

Just as there are different types of reading, one can also identify different types of listening (Green, 2014:103). Probably the types that usually come to mind, are listening for gist, listening selectively for specific pieces of information, or listening in depth, paying attention to every detail (Green, 2014:103). Other types of listening may include listening to understand opinion, to follow an argument or to infer something that is not directly stated. Clearly, like with any communicative situation, the purpose of listening will determine the type or types of listening that are the most suitable. The purposes of listening vary according to whether learners are involved or not in listening as a component of social interaction (Dhanapala, 2019:5). Brown and Yule (1983) distinguish between *interactional*, which takes place in social activities, and *transactional* listening, which mainly focuses on communicating information and, as a result, makes accuracy and coherence particularly relevant when transmitting the message.

Nunan (2004:239) establishes a different classification of listening activities, depending on whether they are *reciprocal* or *nonreciprocal.* In reciprocal listening, the listener is also an interlocutor, which adds to the cognitive load as we saw in the Speaking section. In nonreciprocal listening activities, the listener has no opportunity to interact and, hence, has no opportunity to check their comprehension. Successful listeners, like successful readers, are good at choosing the type of listening that suits the input, the communicative situation and their objectives (Green, 2014:103) and, apart from understanding the specific demands posed by their purpose of listening, also need to be good at other abilities that are vital when it comes to listening. Most of them are common to reading and listening, e.g. understanding key words and being able to understand paraphrasing, being able to concentrate and focus on the question and process a large amount of information, understanding linking words and referencing, coping with inference and interpretation, as well as understanding tone, attitude and paralinguistic features.

2.1.7.2. Mental processes

According to scholars such as Klatzky (1980) and Brown (2008), both cited by Dhanapala, (2019:3), the listening process goes through five stages: sensory memory, attention, short-term memory, rehearsal, and long-term memory. Some scholars claim that beyond the point of word recognition, listening and reading require essentially the same processes (Gough et al., 1996:2; cited by Green, 2014:103), as they both start from input and involve building up mental conceptualisations of a message. The model of listening proposed by Klatzky (1980) suggests that first, listeners register the information and then they engage in pattern recognition processes and register the resulting information. This stage is where the greatest differences between listening and reading can be observed. The decoding process is more challenging in listening as word boundaries may be unclear and phonological effects such as weak forms, reduction, speaker variation and word transitions can make it difficult to discriminate sounds and recognise words (Field, 2008; cited by Green, 2014:104). In addition, grammatical patterns often differ from the ones found in writing. Other challenges that listeners need to overcome have to do with interpreting intonation and stress or coping with speed. Aspects such as hesitation, fillers and organisation should not be forgotten either (Hughes, 2010; cited by Green, 2014:105). Once input has been registered, listeners need to retain it. This process takes place in the last two stages i.e. rehearsal and information preservation, of Klatzky's (1980) listening process, where listeners need to interpret the information using the immediate and the larger sociocultural context of the utterance (Vandergrift, 1999; cited by Dhanapala, 2019:2). In their research using verbal protocols, Suchithra et al. (2014:480) offered further detail of the different processes candidates engage in when listening and also shed light on the activities they do in preparation for listening. In the pre-listening stage, participants in the research reported that they looked at rubrics and questions before starting listening, they also highlighted specific words to prepare for listening and answering questions, and identified the type of listening that best matches the task. While listening students evaluated the information presented, the speaker and the delivery of the information, they also reported using their background knowledge to make sense of the text and made linguistic inferences to deal with unknown vocabulary and engaged in speculation. Participants monitored their comprehension during the listening and looked at previous questions to make sense of problematic aspects. Interestingly, participants took notes to aid understanding and to identify problematic vocabulary.

The cognitive and metacognitive processes described by Suchithra et al. (2014:480) are not very different from the ones readers would use, which supports Gough et al. (1996:2)'s position. In fact, just like in the case of reading, listening has also been analysed from the perspective of bottom-up and top-down dimensions since the 1980s. In the bottom-up model, listening is a process of decoding the sounds that one hears in a linear fashion, from phonemes to full texts. Information builds up until meaning is derived at the end of the process. Conversely, the top-down approach suggests that listener uses background knowledge to make sense of what is heard. Nunan (2002:239) suggests that successful listening comprehension relies on the integration of both bottom-up and top-down approaches.

2.1.7.3. Implications for assessment

Field (2008; cited by Green, 2014:103) claims that most listening tests only include a narrow range of types of listening, mainly listening for specific information or listening for detail and neglect other types of listening such as interactive listening, i.e. the type of listening we do when we are having a conversation with someone.

Another aspect that needs to be considered when designing a test – be it a reading or a listening test – is whether the texts will be taken from the real world or created for the test. Although ideally to maximise the authenticity of the test, input should be obtained from the situation in the real world that the assessment is intended to reflect; in practice, these texts even if carefully chosen are unlikely to follow the test specifications and, hence, pose problems in terms of practicality and reliability of the test. Authenticity also has to do with the purpose for reading and listening. In an exam, the purpose should reflect the aim that a reader or a listener would have to read such text or listen to such recording in the real world. Although the text may be authentic, the candidates' engagement may be different from the one they would have in real life, and that would compromise the authenticity of the task. As a result, the objective is to design exam tasks that activate the same mental processes as the activities in the target language use domains. In this sense, Khalifa and Weir (2009) studied how closely cognitive processes used in real life situations are reflected in the texts included in the Cambridge English Qualifications reading papers. Several scholars such as Alderson (2000), Buck (2001), Davidson and Lynch (2002), all cited by Green

(2014:108), support the view of creating tasks that are not necessarily taken from the real world but which activate the same mental processes as the "real" activities and claim that if it is not possible to find recordings which are suitable for use as input to listening tests, an alternative is to use recordings that are as similar as possible to real world language use.

2.2. An updated approach to washback

2.2.1. Research into washback and evolution of the concept

As a result of the evolution of language testing and social changes, there has been an increasing interest in washback over the last four decades, starting first with test washback on teaching and learning in the field of general education and then expanding to the field of language education (Ha, 2019:3). A review of the literature on the subject shows that scholars' initial concern was to state whether washback really existed and find evidence for it because authors often talked about the existence of washback although no one had accounted for it (see Pearson, 1988; Hughes, 1989; Fredericksen & Collins, 1989). This is what Gosa (2004:29; cited by Tsagari, 2006:14) calls *The Myth Phase* and covers the period before 1990.

The next phase was *The Metaphor Phase* and it started in 1993 when Alderson and Wall questioned whether washback was just a metaphor that led researchers to explore the relationship between teaching and testing and between learning and tests (Alderson & Wall, 1993:121). At this stage, the concept of washback and what it entailed was unclear – in fact, it remains unclear – and Alderson and Wall (1993) already pointed at the complexity of the phenomenon due to the variety of factors that come into play and its intricate nature (Saglam, 2018:158), suggesting an agenda for future research (Ha, 2019:3).

This complexity was widely reported in a number of empirical studies which were published in the so-called *The Reality Phase*. Scholars such as Messick (1996:242) tried to explain this complexity by pointing at the relationship between washback and validity. Others such as Bailey (1997) and Bachman and Palmer (1996:30) tried to explain it from a different perspective distinguishing two different levels (micro and macro) at which the effect of a test can take place. A similar division, in this case between *participants in washback processes* and *processes* and *products of washback*, can be found in Bailey's work (1997). Since then, more recent research

projects (e.g. Collins, Reins & Stobart, 2010; Polesel, Rice & Dulfer, 2014 or Cheng, Sun & Ma, 2015) have tried to shed light on the complexity of washback.

In addition to the concern about washback complexity, research has focused on understanding the effect of specific tests which are considered to be high-stakes. The exams under study range from school leaving tests, university English exit examinations, to well-known tests like IELTS, and the Test of English as a Foreign Language (henceforth TOEFL) (Wei, 2017; cited by Sultana, 2018:151).

In terms of the context of the studies, at first scholars were usually interested in how assessment innovation affected teachers and how tests could shape teaching; later, the focus has widened and studies have been conducted in areas were high-stakes tests have long been present in the education system.

As for the subjects under study, only recently have the effects of washback on students and their learning gained momentum (Green, 2007a, 2007b; Shih, 2007; Tsagari, 2007; Pan & Newfield, 2013; Cheng & Sun, 2015; Sun, 2016; Kim, 2017; all cited by Sultana, 2018:153; Gosa, 2005; Xie & Andrews, 2013). In fact, when one compares the number of washback studies focused on teachers, learner washback research is lacking (Cheng, 2008; cited by Pan, 2014; Shih, 2007; Cheng, Sun & Ma, 2015; Damankesh & Babaii, 2015; cited by Ha, 2019:4; Booth & Davis Lee, 2019:19) despite being equally important in the washback process (Tsagari, 2007:314). Research has evidenced that students are the ultimate stakeholders in any assessment and play an important role in their own learning process as they determine for themselves how to prepare for a test. Moreover, their views, attitudes and feelings towards the test are crucial in the presence or absence of washback (Tsagari, 2007:314). When trying to explain the mechanism of washback, scholars such as Gosa (2004), Green (2007a), Mickan & Motteram (2010) and Xie & Andrews (2013) argue that washback to the learner does not flow in a straightforward manner either directly from the test or from washback to the teacher. Among the aspects explored, questions about what washback looks like, students' attitude and motivation towards the test, parental influence in test preparation, and test preparation have been the most common (Green, 2006 and 2007a; Shih, 2009; Cheng & Sun, 2015; Allen, 2016; all cited by Sultana, 2018:153).

Despite the number of studies conducted, which make Green (2013; cited by Sevilla Morales & Chaves Fernández, 2020:207) feel optimistic that a good deal of progress has been made, it is

clear that more research is necessary as questions such as how a test influences teaching and learning, the mechanisms of washback (Cheng et al., 2004; Xie & Andrews, 2013; all cited by Dong, 2020) and what is understood by student learning remain unclear (Tsagari, 2006:56). Green (2013; cited by Sevilla Morales & Chaves Fernández, 2020:207) argues that washback needs to be studied and understood within *specific contexts of test use* and that more research is needed to understand the roles of different educational actors in the generation of test washback and once again emphasises students as "perhaps the most important participants of all". Scholars such as Wall (2002:502) also point at questions related to test design and learning outcomes as interesting subjects for future research. In addition, in recent years testing experts concerned with ethics in language testing have argued for the need to be accountable to test-takers and to investigate how tests impact upon them (Shohamy, 1997; Hamp-Lyons, 2001; cited by Tsagari, 2007:226).

2.2.2. Defining washback

When referring to the effects of tests, the term *washback* is used in applied linguistics, language education and language testing to refer to this effect on teaching and learning (Hughes, 1989; cited by Saglam, 2018:156). The term *backwash* is more commonly used in the general education field to refer to the same concept. Some scholars narrow this concept down by specifying when the effect can be felt more strongly. This is the case of Buck (1988:1; cited by Riswandi & Wahyudi, 2018), who defines washback as "a natural tendency for both students and teachers to tailor the classroom activities to the demands of the test, especially when the test is particularly important for test takers". Alderson and Wall, whose seminal work led the path for many other scholars in the field, defined washback in 1993:17 (cited by Saglam, 2018:156) as the influence of tests which "lead teachers and learners to do things they not necessarily otherwise do". This definition and term have been used by relevant authors such as Bailey (1996), Messick (1996), Hamp-Lyons (1997), Cheng et al. (2004), and Tsagari (2007). However, other scholars have opted for referring to the exam influence with different names: *measurement-driven instruction* (1987), *washback validity* (Morrow, 1986 and Weir, 1990), *systemic validity* (Fredericksen & Collins, 1989), *curricular alignment* (Smith, 1991), and *impact* (Wall, 1997). The latter has been more controversial though, because for some language testers impact is understood to have a much broader scope encompassing any of the effects that tests may have on individuals, policies or practices, within the

classroom, the school, the educational system, or society as a whole (Tsagari, 2007:4). Following this definition, washback is only one of the dimensions of impact. However, other language testers distinguish washback and impact as micro and macro effects of testing within society (Taylor, 2005; cited by Ha, 2019:4).

2.2.3. Dimensions and complexity of washback

The acknowledgement of the complexity of washback dates back to 1993. Since then, scholars such as Watanabe (1997) or Cheng (2001); all cited by Estaji (2013:222), have echoed Alderson and Wall, suggesting that washback is a dynamic phenomenon which involves a range of intervening factors such as tests, test-related teaching, learning, and the views of stakeholders. These views have been supported by a large number of empirical studies evidencing this complexity (Wenyuan, 2017:62), which can be explained by the fact that washback is "an interactive multi-directional process involving a constant interplay of different degrees of complexity among the different washback components" (Estaji, 2013:222). In addition, testing and assessment impact not only the educational context but also society as a whole, which entails that, apart from the direct participants, it is affected by indirect ones that give washback greater significance but also add to its complexity. Bachman and Palmer (1996:39) refer to this reality and identify a micro level to refer to the effect felt by the individuals – test-takers and teachers – and a macro level involving the educational system or society. However, as we have seen before, authors such as McNamara (2000; cited in Cheng, Sun & Ma, 2015 and in Muñoz, Véliz-Campos & Véliz, 2019:103) argue that impact refers to the macro level effect of tests, for example of national testing policies, while washback refers the micro level, i.e. the classroom.

In the following lines washback will be analysed from several dimensions and different models to explain the mechanisms and scope of washback will be explored.

Watanabe (2004) distinguishes five dimensions of washback: specificity, intensity, lengths, intentionality, and value. *Specificity* refers to whether the washback is general or specific. Washback is said to be general when if refers to an effect that may be produced by any test (Watanabe, 2004:20). Alderson and Wall (1993) also used the term *general washback* when a test impacts the content taught by teachers. On the other hand, *specific washback* applies to the effect

produced by only one specific aspect of a test or one specific test type (Watanabe, 2004:20). For example, specific washback can be observed if a study reports that multiple-choice questions do not encourage learners to learn productive language skills.

Intensity, also referred to as *extent* by Bachman and Palmer (1996), describes the strength of washback. It was first coined by Cheng (1997:43; cited by Ha, 2019:4) and then further developed by Watanabe (2004) and Green (2007a) to refer to the "degree of washback effect in an area or a number of areas of teaching and learning affected by an examination". Intensity depends on several factors that mediate the process of washback being generated, e.g. prestige of the test, participants' perceptions of test stakes, test quality and test difficulty, which, in turn, tend to vary from person to person. Several studies evidence this idea reporting that washback is not to be attributed only to characteristics and qualities of the test, but to a variety of factors that might influence test performance (Muñoz, Véliz-Campos & Véliz, 2019:112). For example, the research conducted by Muñoz, Véliz-Campos and Véliz (2019:112) identified the teacher and more precisely their relationship with their students as a potential factor that influenced the perceived washback. Karabulut (2007; cited by Toksöz & Kılıçkaya, 2017:185) established a connection between the stakes of an exam and the intensity of the washback it produces. In this sense, if a test is significant for the test taker, it displays strong washback. When this is the case, the exam will determine everything that happens in the classroom, and will lead all the teachers in the same way towards the exam. However, if the test is not fundamental at all, it presents weak washback and, hence, will affect only a part of the teaching and learning, or only some teachers and students, but not others (Hakim & Tasikmalaya, 2018:59).

Length is used to describe how long the washback of a test lasts (Watanabe, 2004) and the effect can be said to be short term or long term. An example of a study reporting an instance of short-term washback can be found in Shohamy et al. (1996), who found that the effect of the national test of Arabic as a Second Language existed only before it was first administered.

If *intentionality* is considered, washback can also be intended or unintended (Messick, 1996; Andrews, 2004; Tsagari, 2006). Heyneman and Ransom (1992) refer to intended washback when they claim that tests can be a powerful, low-cost means of influencing the quality of what teachers teach and what learners learn at school. Assessment institutions strive to produce fair, reliable and practical tests which meet high quality standards and, hence, produce positive washback.

Nevertheless, as Wall and Alderson (1993) point out, it would be naïve to assume that the effect of a test having a set of qualities is enough in itself to produce change. There is a number of factors such as teacher factors (Watanabe, 1996:331) or lack of communication between test providers and test users (Tsagari, 2006:48) which may outweigh the influence of an examination. This may lead to effects that were not initially considered, therefore unintended, by the assessment institution.

Some scholars connect intentionality with *value* (see below), this is the case of Buck (1988), Bachman and Palmer (1996), Davies et al. (1999), and also Hughes (2003). The term *value* is equivalent to *direction*, which is used by other authors e.g. Alderson and Wall (1993), Green (2007a), Tsagari (2007), and Bailey and Masuhara (2013). This dimension has been widely studied mainly in areas were high-stakes tests have long been present in the educational system e.g. Collins, Reiss and Stobart (2010), Polesel, Rice and Dulfer (2014) and Cheng, Sun and Ma (2015), and evidence has been found of the influence of high-stakes exams on teaching (Muñoz, Véliz-Campos & Véliz, 2019:105-106). Commonly, washback is considered as a neutral term (Alderson & Wall, 1993) and may refer both to positive – beneficial – or negative – harmful – effects. According to Bailey and Masuhara (2013:304; cited by Ha, 2019:4), the value of washback is not absolute as it depends on our view of the desirable outcomes of language learning. A test may be thought to exert negative washback when its content or format is based on a limited definition of language ability and, as a consequence, has a detrimental effect on the breadth of, or the variety to be found within a curriculum, preventing students from learning real-life skills, for instance. Tests also produce negative washback when they bring anxiety both to teachers and students and affect their performance (Ahmad & Rao, 2012; cited by Hakim & Tasikmalaya, 2018:62). Some educators identify standardised testing with negative washback although some testers have pointed out that test design does not affect the nature of washback – or is not the only factor affecting it – and argue that it is the misuse or overuse of test results that produce negative washback (Messick, 1998; Shohamy, 2001; both in Xie & Andrews, 2012:50). Positive washback can be investigated taking into account four different variables i.e. the purpose of language learning, authenticity of testing, students' autonomy and self-assessment, and feedback of test results (Bailey, 1999; cited by Hakim & Tasikmalaya, 2018:59). It occurs when a test brings about good teaching practice (Taylor, 2005:154), for instance, encouraging teachers to cover their subjects more thoroughly.

Washback can also be considered positive when a test motivates students to work harder and thus enhances learning (Hakim & Tasikmalaya, 2018:62).

The effect of a test can be felt by the participants involved, on the processes, and on the products according to Hughes (1993; cited by Pan 2014) and his trichotomy of backwash model. Hughes calls everyone "whose perceptions and attitudes toward their work may be affected by a test participants", for instance, teachers, students, administrators, materials writers and publishers (Hughes, 1993:2). *Process* is understood as any action undertaken by participants to carry out teaching and learning tasks e.g. materials development, syllabus design, modifications in teaching methods or content, as well as the learning or development of test-taking strategies (Hughes, 1993:2). For example, processes for students include using the target language skills, studying, learning, memorising, worrying, or cheating. The process for the teachers are what they teach, how they teach, the intensity of teaching, and additional tutorials, whereas the process for programmes are changing curricula, scheduling test preparation classes, using new materials, and cancelling classes (Hakim & Tasikmalaya, 2018:63). Finally, the *product* refers to the learning outcomes and the quality of learning (Hughes, 1993:2). It includes changed teaching, leading to increased interaction and studying and better learning, new materials, and new course syllabi (Hakim & Tasikmalaya, 2018:63). Hughes's trichotomy of the washback model provides a theoretical framework that sheds light on the relationship between participants, processes and products and on the washback mechanisms. In his view, the nature of a test may first affect the perceptions and attitudes of the participants, while in turn, these perceptions may affect how participants react when performing their work – their process. This, then, affects the learning outcomes, i.e. the product (Dong, 2020).

Bailey's (1996) *Basic Model of Washback* is based on Hughes's, and on the fifteen hypotheses from Alderson and Wall (1993). She distinguishes between *washback to the programme* i.e. what and how teachers teach and the rate / sequence and degree / depth of teaching, and *washback to the learner* i.e. what and how learners learn and the rate / sequence and degree / depth of learning, to explain the way in which washback works in the teaching and learning context (Tsagari, 2007:24; Estaji, 2013:220). She includes five of Alderson & Wall's Hypotheses (2,5,6,8,10) under the washback to the learner heading and give ten examples of the actions that learners might carry out when studying for an important exam. These go from doing activities similar in format to those that can be found in the exam, to putting into practice test-taking strategies, to

attening test-preparation courses and to studying for the test instead of going to language classes (Bailey, 1996:264-265; cited by Tsagari 2006:24). Although Bailey's model (Figure 4) is a step forward in the explanation of the mechanism of washback since she explains in more detail what is understood by the products and by the participants – among which she includes researchers – and adds on the potential interactions of the participants, Pan (2014) argues that her model still fails to explain the possible differences in how the test impacts individuals.

1. A test will influence teaching
2. A test will influence learning
3. A test will influence what teachers teach
4. A test will influence how teachers teach
5. A test will influence what learners learn
6. A test will influence how learners learn
7. A test will influence the rate and sequence of teaching
8. A test will influence the rate and sequence of learning
9. A test will influence the degree and depth of teaching
10. A test will influence the degree and depth of learning
11. A test will influence attitudes to content, method, etc. of teaching/learning
12. Tests that have important consequences will have washback
13. Tests that do not have important consequences will have no washback
14. Tests will have washback on all learners and teachers
15. Tests will have washback effects for some teachers and some learners, but not for others.

(Alderson & Wall Hypotheses, 1993:120-121; cited by Tsagari, 2007:8)

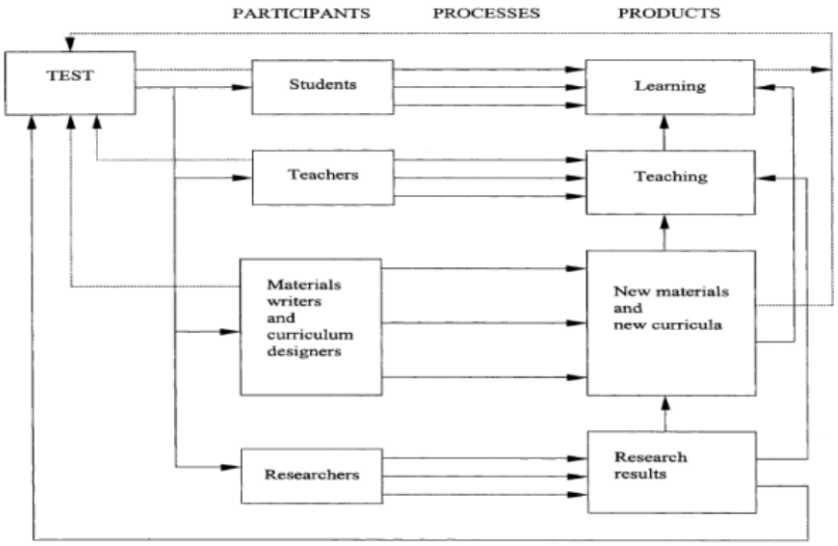


Figure 4. Bailey's Model of washback (1996)

In terms of the areas where washback can appear, Spratt (2005) in Wall (n.d.) identifies five. First, the curriculum, washback can be observed in decisions regarding what to teach – content – and also on the time devoted to teaching this content. Second, washback can also be present in the materials used e.g. choice of course books, use of past papers, teacher-made materials, etc. Third, the methodology can also be affected as teachers may alter their choice of methods and decide to include test-taking strategies. From the learning point of view, we may try to find evidence of washback by analysing test results to try to determine if learning has improved due to a test. Finally, the attitudes and feelings of both learners and teachers can also be influenced by a test. As we have already seen, teachers may feel pressured by exam results and a test may encourage some students to work harder while it may increase the anxiety of others.

Prodromou's (1995) also focused on the areas where washback can be observed and identified that washback could be felt mainly in the methodology and the classroom atmosphere. To explain this, Prodromou used the terms *overt* and *covert* washback. Overt washback is more evident and impacts the choice of materials and the types of activities done in class, mainly past papers and task types that are similar to the ones students can find in the exam. Covert washback may be less obvious but can be felt on lesson sequencing, classroom arrangement, and on specific aspects of teaching methodology. Some professionals may modify their lesson planning as a result of the test by removing some activities and prioritising others. Teachers may decide not to use

warm-up activities, which are used to engage students, activate their schemata and prepare them for the main activity. By doing so their lesson sequencing becomes more similar to the procedure used in exams in which input is typically presented without any introduction and no interaction is possible (Prodromou, 1995; cited by Peña Jaenes, 2015:36).

Covert washback can also be observed when dealing with errors. While there is a tendency among some assessment institutions such as Cambridge Assessment English to follow a positive approach towards marking by giving candidates credit for what they can do (University of Cambridge Local Examinations Syndicate, 2018) and introducing more information on their certificates to provide candidates with feedback regarding their performance, Prodromou (1995; cited by Peña Jaenes, 2015:36) mentions that proficiency tests usually penalise errors without offering useful feedback and that some teachers may do the same. In doing so, they fail to use mistakes and errors as opportunities for improvement and learning.

Finally, covert washback can also be observed in seating arrangement. Seating arrangement has an effect on the social and academic aspect of the classroom. On the one hand, the way chairs and desks are arranged influences the atmosphere in the classroom and students' relationships with each other (Gest & Rodkin, 2011; van den Berg et al., 2012; McKeown et al., 2015; all cited by Gremmen et al., 2016). On the other hand, classroom space influences learning and the teachers' and students' attitudes towards school (Denton, 1992) and students' engagement. A seating arrangement in rows, as opposed to groups, may favour academic behaviours such as raising one's hand to ask for help and and comply with requests, which may facilitate classroom management. Similarly, rows also support students' concentration when they are carrying out individual activities (Gremmen et al., 2016). However, Prodromou perceives this seating arrangement as a way to give students the idea that they are under exam conditions. Conversely, seating arrangements in small groups can make interaction between students easier since they are closer and that can foster collaborative work and be of use in some stages of the lesson such as brainstorming (Gremmen et al., 2016).

The nature of washback can be affected by several factors, which according to Spratt (2005) are: the exam, resources, the school, teachers' beliefs, teachers' attitudes, teachers' education and training.

Spratt identified several aspects that are directly linked to the exam, for instance, its *proximity* – as we have already seen, the closer the exam is the stronger the washback can be felt. Also, the *stakes* of the exam for the students and the *status of the language* it tests can be said to increase the washback because if the stakes are high and speaking the language is considered important, students may feel more motivated to work hard to pass the test and certify their level. Apart from these aspects, Spratt mentioned test characteristics such as the purpose of the test or its format as having an impact. In addition to those, contextual factors such as when the exam was first introduced and the teachers' degree of familiarity with the exam should also be considered.

As for the academic context, the characteristics of the schools and the resources and materials used also play their role influencing washback. The language school and its atmosphere was identified by Watanabe (2004:164) as mediating the influence of washback in the sense that a positive school atmosphere focused on helping students improve their skills may transfer to exam classes. In addition, the time devoted to exam classes and the number of students in them, the amount of pressure that school administrators or even parents put on teachers to achieve results, and cultural factors can account for big differences in the nature of washback. Similarly, the course book can determine the contents being taught, how frequently the different skills are practised or even classroom interaction. Moreover, the availability of exam support material with information about test specifications can be useful for teachers and can also influence their attitude towards the exam. All of them can also mediate the effect of the exam on the classes, the teachers and the learners.

Finally, teachers: their beliefs about the exam, i.e. its reliability and fairness, its usefulness and stakes can determine how much they are influenced by it. Without a doubt, teachers have their own personal methodology and their beliefs about teaching and exam preparation. If the exam contravenes their teaching philosophy or their practices, or if they do not see a connection between test results and their instructional approaches (Mehrens, 1998; cited by Peña Jaenes, 2015:32), they are more likely to have negative attitudes against it and hence the washback may be weaker. These beliefs and attitudes are based on their education and training, how much general methodological training they have received and whether they have received any training in specific exams, which has explained the exam rationale and philosophy and has made them familiar with test specifications. In fact, Watanabe (2004:141) reports that some teachers placed undue blame on the effect of the exam on their classes on the basis of misconceptions about it.

In the light of the above, it is not surprising that washback to the teacher has received a great deal of attention. However, Cheng, Watanabe and Curtis (2004:x) explain that it is not only the teacher but the people in the classroom the ones who bring about washback and point out that test developers can only exert a limited influence on how people prepare for their test.

2.2.4. Washback to the teacher

The way in which the literature reviewed describes the *influence* of teachers differs greatly. On one end of the continuum, Swain (1985:43) reports that teachers teach to the test if they are familiar with the content and format of the test. Prodromou (1995:19) explains this position on the basis of the pressure that professionals may feel to cover the examination syllabus. This pressure can come from the school, as we saw before, from the parents, or from the students, who may feel reluctant to learn content which does not seem to be in the form of examinations practice. This deterministic view of teaching is aligned with the idea that "what is assessed becomes what is valued, which becomes what is taught" expressed by McEwen (1995:42; cited by Cheng, Watanabe & Curtis, 2004) and with the view expressed by Popham (1987), that tests and examinations can and should drive teaching. As for where the effect can be felt, scholars seem to agree that tests will have more impact on the contents taught and on the materials used than on the teacher's methodology (Wall & Alderson, 1993:68; cited by Watanabe, 2004:130). In fact, test preparation courses are commonly perceived as more limited in terms of contents and activities (Peña Jaenes, 2015:81) as a result of the washback of proficiency tests.

This limitation in the effect of exams brings us to the other end of the continuum, where several scholars believe that teachers and the *teacher factor* have a greater importance in washback than the exam itself and accounts for a variability in teaching that is greater than the effect of the exam. This is expressed for instance by Watanabe (1996:331) and by Alderson and Hamp-Lyons (1996:290), who observed that the differences between the teachers taking part in their study were at least as great as the difference between test and general classes. Alderson (2004) elaborates this idea more and explains that he became aware of the teacher factor in washback when he realised how differently two teachers prepared their students for the same test (Alderson, 2004, in Cheng, Watanabe & Curtis, 2004:x) and how similarly two teachers could teach for different tests and realised that it is at least as much the teacher who produces washback, be it positive or negative as it is the test (Alderson, 2004, in Cheng, Watanabe & Curtis, 2004:x). Green

(2007:19) also supports this view explaining that teacher variables contributed at least as much to the variation in practices as did the test / no test distinction.

The reasons behind the variation in how washback is felt have already been covered in the previous section and include test related factors but also the professionals' characteristics and background. What seems clear is that while no one can deny the importance of exams and exam preparation, the main concern shared by all professionals is how to best support students in their learning.

2.2.5. Washback to the learner

Washback to the learner has received little attention in washback research despite the fact that learners are the individuals who are probably most impacted by exams and exam results. A possible explanation of this gap in research can be the role learners have traditionally had in learning, which could be said to be more passive as opposed to the more active participation expected nowadays. At present, learner autonomy is a crucial element that professionals try to reinforce by making students aware of the learning objectives and making them familiar with the assessment criteria and methods. This sense of *agency* has extended to washback research and a number of aspects which are relevant to analyse how learners react to exams and their effect have been identified.

Green (2007:82) considers, for instance, that learners' age was a relevant factor influencing perceptions about the importance and level of challenge on the test. Authors such as Gosa (2004), Xie and Andrews (2012), Zhan and Andrews (2014), and Zhan and Wan (2016) also perceived the power of personal beliefs i.e. perceptions of test importance and difficulty, educational experience, i.e. the degree of familiarity with the test, and the context, e.g. resources available to meet test demands and their willingness to accept these demands, as mediating factors.

Expanding on personal beliefs and attitudes, although authors such as Shohamy (1993), Cheng (1998) and Tsagari (2007) reported anxiety and negative attitudes held towards the test, tests are also perceived as motivators for learning. In fact, Tsagari (2007:265) explains that many participants in her study reacted positively to the challenge even when the results were below expectations and used poor exam results as learning tools to identify strengths and weaknesses.

Another area where washback can be observed is exam preparation and learning strategies. Allen (2016:5) suggests that learners can modify their learning strategies to adapt to the format of

the exam and on the basis of their initial performance on the test and its perceived difficulty. Green (2007), Mickan and Motteram (2008), and Zhan and Andrews (2014) point out that students tend to focus on test-related tasks and materials to prepare for the test although for those students who prepare for the test independently without attending lessons there is greater variability in the preparation strategies adopted. This could be thought to be connected to the role of teachers and the influence they exert on the strategies used by their students. Nevertheless, washback affects teachers and learners differently, in the sense that they have different ideas of the best way to achieve learning goals (Peña Jaenes, 2015:81). Allen (2016:15) sees here an opportunity for further research because test takers often need to rely on external support to prepare for the test and do not have the autonomy to guide their learning, which could be an obstacle to achieve positive washback.

Apart from learning strategies and students' perceptions, washback can be observed in score gain. Relevant scholars perceive this line as a way forward in analysing the influence of exams on students' progression and learning. However, few studies have been conducted to date and they have often offered limited information since, in many cases, it is difficult to isolate the effect of the exam from all the other factors that mediate washback.

2.3. Accreditation exams

2.3.1. Relevance of accreditation exams

Examinations, in general, are deeply embedded in our culture and have an effect on life-changing decisions such as immigration, education, and career prospects of millions of people around the world (Raban, 2008:x and University of Cambridge Local Examinations Syndicate, 2016). The presence of examinations is such that Raban (2008) states that although it is possible to be educated without being examined, "the reality during the 150 years since the (University of Cambridge Local Examinations) Syndicate's foundation is that public examinations and the experience of them have become an almost universal phenomenon". As we saw at the beginning of the section, the use of language assessment for recruitment purposes dates back to the second century (Alcaraz, 2015; cited by Muñoz, Véliz-Campos & Véliz, 2019:98). Nevertheless, it is clear that if we look at English language testing, the stakes associated with passing a test have dramatically increased in the last decades (University of Cambridge Local Examinations Syndicate,

2016). This is due to the importance of English, which has been identified as a *lingua franca* in the globalised world (Chávez-Zambano, Saltos Vivas & Saltos Dueñas, 2017:761) and a contributing factor to economic success (Graddol, 2006; cited by Khalifa, 2014:7). Being able to speak a foreign language, especially English, has become a necessity rather than a privilege (Biava & Segura, 2010; cited by Chávez-Zambano, Saltos Vivas & Saltos Dueñas, 2017:761). As a result, including the English language in the education of any country is perceived as necessary and urgent worldwide (Jaimechango, 2009; cited by Chávez-Zambano, Saltos Vivas & Saltos Dueñas, 2017:763).

In the light of this, it is not surprising that the number of candidates of proficiency exams has increased dramatically. In the case of Cambridge exams, the first exam was taken by three candidates in 1913 (Hawkey & Milanovic, 2013; cited by Peña Jaenes, 2015:9) and since then the number has reached 5.5 million candidates a year in more than 130 countries (University of Cambridge Local Examinations Syndicate, 2020c). Similar figures report IELTS with more than three million tests taken in the past year (International English Language Testing System, 2020), a significant increase if compared with the 43,000 candidates taking the test in 1995. Also impressive is the figure reported by ETS with 35 million TOEFL candidates since the exam was first launched (Educational Testing System, 2020).

Assessment providers are aware of the great responsibility they have to develop tests that are "fair, accurate and valid" (University of Cambridge Local Examinations Syndicate, 2016:1). In addition, tests need to be fit for purpose, that is, they should be appropriate for the different needs candidates may have. To achieve this level of excellence, the appropriate balance of the test qualities already discussed needs to be reached and maximised (University of Cambridge Local Examinations Syndicate, 2016:1).

2.3.2. B2 First

2.3.2.1. Cambridge Assessment English

Cambridge Examinations have a long history, which dates back to 1858 when the University of Cambridge Local Examinations Syndicate (henceforth UCLES) was formed to set school leaving examinations for non-members of the university (University of Cambridge Local Examinations Syndicate, 2017). Some years later, the Certificate of Proficiency in English (henceforth, CPE) was first taken. This was only the starting point but it already showed the relevance that speaking

foreign languages and being able to certify that ability had. In 1939, the Lower Certificate in English was introduced as a complement to CPE and it proved particularly popular (University of Cambridge Local Examinations Syndicate, 2017).

Since its foundation, Cambridge Assessment English, previously known as Cambridge English Language Assessment and before that as University of Cambridge ESOL Examinations, has evolved to meet the changing social and educational needs. Today's society is dramatically different from that of 1913, when CPE was taken by only three candidates (University of Cambridge Local Examinations Syndicate, 2017) with millions of candidates taking Cambridge examinations worldwide for immigration, academic or professional reasons. Cambridge Assessment English is aware of this reality and offers assessments across the full spectrum of language ability. It provides exams for schools, general and higher education and business (University of Cambridge Local Examinations Syndicate, 20119a:2). In addition, Cambridge Assessment English is fully committed to quality and research, which are paramount to maintain the recognition achieved worldwide. The assessment institution is in charge of a series of studies in Language Testing and follows a number of Principles of Good Practice (University of Cambridge Local Examinations Syndicate, 2018) so that its exams produce a positive impact on teaching and learning. Its commitment with research has produced a new multilevel test, which is adaptive and which can be taken remotely. The exam also relies on Artificial Intelligence to support examiners to mark the writing component (University of Cambridge Local Examinations Syndicate, 2020).

Moreover, Cambridge Assessment English has aligned its exams, with modern developments in language teaching, taking into consideration the latest trends in Linguistics and Applied Linguistics, and has actively involved teachers in the design, setting and marking of Cambridge exams (Tsagari, 2006:6). An example of this is the fact that Cambridge exams are periodically revised to reflect the abovementioned trends in teaching and testing. For instance, B2 First – previously known as First Certificate in English (henceforth FCE) – has been revised three times in the last twenty years. Nevertheless, Cambridge examinations are still thought to reinforce traditional ways of teaching, encouraging teacher-centred classes and individualistic work (Gabrielatos, 1993; Prodromou, 1993; Kenny 1995; all cited by Tsagari, 2006:7). These authors also point at the negative washback produced by the test as it leads teachers and students to focus on test-taking strategies (Tsagari, 2006:7). From a different perspective, Tsagari (2006:9) also reports

that B2 First encourages the development of all four skills and motivates students to work harder during the exam preparation period.

2.3.2.2. B2 First: revisions and current structure

The B2 First exam was first offered in 1939 and it is a qualification at upper-intermediate level, which certifies a B2 level of the CEFR (Council of Europe, 2001). Since 1939, B2 First has undergone several revisions, the last three have taken place in two decades and have attempted to provide *positive educational impact* (Tsagari, 2006:7): one in 1996 (Tsagari, 2006:7), the second one in 2008, and the most recent one in 2015 (University of Cambridge Local Examinations Syndicate, 2013b:3). The last one involved several minor changes but also some key modifications that had to do with the structure of the test, which now has four papers instead of five, its format – candidates can choose to take B2 First as either a paper-based or a computer-based exam (University of Cambridge Local Examinations Syndicate, 2019a:3), and its version since there is a B2 First for Schools, aimed at under age students, and the standard version for adults.

The exam covers all four language skills – reading, writing, listening and speaking – and it requires candidates to have understanding of the structure of the language (University of Cambridge Local Examinations Syndicate, 2019a:3). From the perspective of the modes of communication (Council of Europe, 2018), it tests production and reception and it could be argued that it also assesses interaction in the speaking tests as candidates are expected to interact with one of the examiners as well as with their partner in the speaking test, and in the writing paper because candidates are given clear instructions of the communicative situation so it could be claimed that they have a clear target reader in mind when producing the text. Mediation is also included in the exam construct because candidates are expected to select relevant information for a target reader in the writing test and also build on different contributions to a discussion, stimulate reasoning with questions and collaborate to construct meaning in the speaking test. There are four papers: Reading and Use of English, Writing, Listening, and Speaking, and the weighting of each of the four skills and Use of English is equal. Therefore, the overall performance is calculated by averaging the scores achieved on all the parts (University of Cambridge Local Examinations Syndicate, 2019a:3). In the next lines, the different papers will be described.

2.3.2.2.1. Reading and Use of English

In this exam paper, candidates are expected to understand texts from publications such as fiction and non-fiction books, journals, newspapers and magazines. They have one hour and fifteen minutes to do the 52 questions included in the seven parts of this paper. Each part has a different focus, ranging from more lexical contents such as collocation, fixed phrases and word formation to more textual aspects such as cohesion and coherence. Parts 2 and 3 contain texts accompanying grammar and vocabulary tasks, and for each correct answer the candidate receives one mark. Part 4 consists of separate items with a grammatical and lexical focus, each correct answer receiving up to two marks. Parts 1, 5, 6 and 7 contain a variety of texts and reading comprehension tasks. In terms of marking, each correct answer of Part 5 and 6 receives two marks whereas each correct answer of Part 1 and 7 is awarded one mark (University of Cambridge Local Examinations Syndicate, 2019a:7).

2.3.2.2.2. Writing

Candidates are expected to produce two different pieces of writing: a compulsory task in Part 1, and one from a choice of three – B2 First – or four – B2 First for Schools – in Part 2. It lasts one hour and 20 minutes. Each question on the paper carries equal marks. The task types that may appear in the test are different depending on the version of the test. B2 First for Schools may include an article, an email or a letter, an essay, a story, or a review (University of Cambridge Local Examinations Syndicate, 2019b:27), whereas B2 First may include an article, an email or a letter, an essay, a report, or a review (University of Cambridge Local Examinations Syndicate, 2019a:27).

In Part 1 candidates are asked to write an essay where they should give their opinion and justify it. The task has an opening rubric, which sets the scene, an essay question and two prompts plus an additional prompt that candidates have to produce. The total number of words in the rubric is 120 words. Regarding the topic, the subject of the essay is of general interest and, hence, does not require specialised knowledge. The length of the candidates' essay should range between 140 and 190 words.

In Part 2 candidates can choose from a number of possible text types based on a contextualized writing task. In all the options, candidates have a clear context, topic, purpose and target reader for their piece of writing. The rubric has a maximum of 70 words. In this part of the

test candidates might have to write an article, an email or a letter, a review and either a report if they are adults, or a story if they sit the B2 First for Schools version. Besides, B2 First for Schools Writing Paper Part 2 includes an additional question based on a prescribed book and which students can choose if they have read the set novel. The length of the text is the same as in Part 1 (University of Cambridge Local Examinations Syndicate, 2019a:27).

The writing tasks are assessed considering a scale which has four criteria: Content, Communicative Achievement, Organisation, and Language. A maximum of five marks can be given for each criterion (University of Cambridge Local Examinations Syndicate, 2013b:34).

2.3.2.2.3. Listening

In this part of the test candidates are expected to understand a variety of spoken material ranging from lectures to radiobroadcasts. The duration of this paper is 40 minutes and there are four different parts. Part 1 is a multiple-choice task in which students have to identify feelings, attitudes, topics, opinion, and purposes. The focus is both on gist and detail. Part 2 is a sentence completion activity in which candidates need to understand detail and identify specific information and stated opinion. Part 3 is a multiple-matching task in which candidates listen to five short related monologues. Finally, Part 4 is a multiple-choice task focused on identifying opinion, attitude, detail, gist, main ideas, as well as specific information (University of Cambridge Local Examinations Syndicate, 2019a:51). It must be pointed out that Cambridge listening tasks are recorded for exam purposes. The recordings are played twice and different native accents are used.

2.3.2.2.4. Speaking

This part of the test, which lasts 14 minutes, is taken by two candidates although sometimes there are groups of three. Test-takers are tested on their ability to produce longer monologues and to take part in different types of interaction: with the examiner and with each other. There are two Cambridge examiners, one of them is the interlocutor, who is in charge of conducting the test and giving the Global Achievement mark, while the other one is the person in charge of giving a mark for the rest of the criteria – Grammar and Vocabulary, Discourse Management, Pronunciation, and Interactive Communication. The Speaking test contains four parts: Part 1 entails a conversation between the interlocutor and each candidate and is focused on general interactional and social

language. Part 2 is an individual long turn for each candidate followed by a response from the second candidate. Test-takers are given a pair of photographs, which they have to compare, and a question they have to answer. This question is asked orally but is also written above the two pictures. Each candidate has one minute for their long turn and 30 seconds for their response about their partner's photographs. The objective is to see whether candidates are able to organise a larger unit of discourse, compare, describe, and express opinions. Part 3 involves a two-way conversation between the candidates to discuss a question and some prompts, which are written. The focus is on sustaining interaction, exchanging ideas, expressing and justifying opinions, agreeing and/or disagreeing, suggesting, speculating, and negotiating towards a decision. They have two minutes to discuss the stimuli and one minute to reach an agreement. Nevertheless, they are not penalised if they cannot agree. Finally, Part 4 is a discussion on topics related to Part 3. Candidates are expected to express and justify opinions, agree and/or disagree, and speculate. The duration is four minutes (University of Cambridge Local Examinations Syndicate, 2019a:71).

As already mentioned, there are five criteria: grammar and vocabulary, discourse management, pronunciation, interactive communication, and global achievement. Test-takers are given a mark out of five for each criterion and the pass mark is considered to be three.

3. RATIONALE AND AIMS

For the last two decades proficiency tests have become more and more relevant for immigration, study and work purposes all over the world (Cheng, 2010[1]; Rahimi, Esfandiari & Amini, 2016; University of Cambridge Local Examinations Syndicate, 2017). In spite of their social role, the views towards them are diverse. While some scholars see them as opportunities to encourage good teaching, others such as Bachman and Palmer (2010; cited by Green, 2013) claim that the skills needed to succeed in a test can never fully equate the skills required for success in a target language domain. Crooks (1988; cited by Green, 2013) goes further and points out that the most deleterious effects come from high-stakes tests that control access to opportunities and so are seen as very important for test takers' life chances.

Despite their controversial role, it is a fact that the number of candidates has not stopped growing; from only three candidates in 1913 to 5.5 million candidates a year, in the case of Cambridge Assessment English (University of Cambridge Local Examinations Syndicate, 2020), and from 43,000 candidates in 1995 to three million in the past year, in the case of IELTS (International English Language Testing System, 2020). ETS also reported the impressive figure of 35 million TOEFL candidates since the exam was first launched (Educational Testing System, 2020). Given the context described above, where there is an increasing need to certify the level of English and a vast number of candidates sitting accreditation exams, a wide range and a vast number of test preparation courses have mushroomed. Students of all ages and profile are propelled towards them in the hope to succeed in high-stakes tests (Robb & Ercanbrack, 1999:2) in spite of the relatively little evidence to prove whether special preparation can have a markedly positive effect on test scores (Robb & Ercanbrack, 1999:2).

Research into testing and assessment carried out to date has paid special attention to testing principles (Davies, 2008; cited by Green, 2014:172) to guarantee that major proficiency tests are valid, reliable, practical and fair (Rahimi, Esfandiari & Amini, 2016:8). Since the seminal work by Alderson and Wall (1993), there has been increasing interest in the effect that tests may have on educational practice (Spratt, 2005; cited by Zhan & Wan, 2016; Xie & Andrews, 2012:51; Green, 2013; Rahimi, Esfandiari & Amini, 2016:7). Nevertheless, the focus has been mostly on teachers and

[1] Published online in 2010 but in a printed format in 1997.

classroom practices. The few studies focused on washback and learners (Elder & O' Loughlin, 2003; Read & Hayes, 2003; Gosa, 2004; Hawkey, 2006; Green, 2007; Mickan & Motteram, 2008; Xie & Andrews, 2012) have offered inconclusive findings due to the unpredictability and complexity of washback (Ha, 2019:13). Thus, there seems to be some agreement concerning the need of gaining further first-hand knowledge (Mickan & Motteram, 2009:4) about how tests may affect students in terms of learning (Wall, 2000:502; Alderson & Banerjee, 2001; cited by Tsagari, 2007:43; Andrews, Fullilove, & Wong, 2002:208; Beikmahdavi, 2016:135), of score gains (Wall, 2000:502; Green, 2013; Green, 2014:17), of their attitudes towards learning and exam success (Gosa, 2004; Green, 2007; Mickan & Motteram, 2009; Xie & Andrews, 2012; cited by Green, 2013), of preparation practices (Mickan & Motteram, 2009:5), of test-taking experiences (Templer, 2004 and Green, 2007; cited by Mickan & Motteram, 2009:4) and of experiences in preparation programmes (Mickan & Motteram, 2008:40; cited by Mickan & Motteram, 2009:5).

In relation to score gains and learning, Taylor (2005) and Tsagari (2007) highlight that future research needs to collect empirical data – Cheng (2010:14) also referred to the lack of empirical data in washback studies, which can clearly show whether students have learned better due to their preparation for a particular test. This is because, as Hughes (1993:5) argued, the *ultimate washback objective* of an English language test will be "the English skills that candidates develop" (Green, 2013:48). It seems logical to think that test preparation courses foster the development of English skills and hence the success in the test. However, Green's investigation in 2007 about IELTS test preparation practices suggested that, contrary to teachers' beliefs, there was no substantial benefit in focusing on the test in preference to studying broader English for academic purposes programme (Green, 2013:44). Robb and Ercanbrack (1999) reached a similar conclusion regarding the little benefits that the Test of English for International Communication (henceforth TOEIC) preparatory materials offer to students in Japan. Tsagari shares their skepticism because, of the washback studies reviewed, only one had documented any demonstrable gains in student learning that could be tied to the use of the test (Saif, 2006; cited by Tsagari, 2007) and the remaining studies had either dubious approaches, did not find any considerable gains, or found negative results.

In the light of the above, it seems paramount to rely on the thorough information available about washback and use it to look into "the crucial stakeholders in large scale, high-stakes examinations" (Zhan & Wan, 2016:372), that is, the students. As for the context of the study,

Tsagari (2007:43) points out that research needs to be conducted not only in contexts where a new exam has been introduced – or when a specific examination has been modified and improved upon (Alderson & Wall, 1993; cited by Rahimi, Esfandiari & Amini, 2016:7) – but where exams have been in operation for an extensive period of time. Ha (2019:13) expands on this and explains that given the context-dependent nature of washback, it is necessary for future research to be conducted in new research contexts to fully understand how it operates and to generate positive effects of testing on learning if we want to use testing to drive education.

With that idea in mind, this PhD Thesis builds on knowledge gained as a result of the MA thesis published in 2015 (Peña Jaenes, 2015) and focuses on language learners and a prestigious high-stakes exam.

The main objective of this PhD Thesis is to obtain an evidence-based understanding of students' progression. More precisely, it aims to measure the effectiveness in terms of score gain in the mock exams of B2 First exam of courses that are oriented towards preparing students for Cambridge Assessment English B2 First exam in comparison with general English courses; the effectiveness in terms of improvement in general language ability of courses that are more exam-oriented as compared with more general English courses; and to gain evidence-based knowledge of the potential washback that Cambridge Assessment English may have on learner autonomy and independence. In light of this, this present study has been designed on the basis of these three research questions:

i) Do students enrolled on more exam-oriented (CEB) courses show a better performance in Cambridge B2 First mock exam than those enrolled on general English (CEALM) courses?

In order to answer this question, the performance at the beginning and at the end of the project shown in two sets of B2 First mock exams by the experimental group and the control group is compared to identify potential score gains.

ii) Do students enrolled on more exam-oriented (CEB) courses improve their language knowledge and abilities?

To try to obtain evidence of improvement, the project measures the performance of students in the two types of courses in two sets of independent grammar and vocabulary tests.

iii) Do students become more autonomous and independent learners as a result of preparing for Cambridge B2 First exam?

The students' answers to two sets of questionnaires are analysed to determine this last aspect.

In the light of the above the stages of this PhD project are:

1. The study of the literature available, the design of the project and the statistical validation of the instruments used.

2. The comparison of two groups of students: the experimental group and the control group. The former includes a total of 89 students, enrolled on courses oriented towards the preparation of Cambridge Assessment English B2 First exam at Centro de Estudios Británicos. The latter includes a total of 41 learners who study English in general at Centro de Estudios Avanzados en Lenguas Modernas.

3. The study of the students' attitudes and expectations regarding the course, English and B2 First exam at the beginning of the project.

4. The analysis of the learners' progression in two sets of tests administered at the beginning of the project and at the end. The first set of tests includes two independent grammar and a vocabulary tests designed and validated for this study and the second set of tests includes two B2 First mock exams.

5. The discussion of similarities and differences in methodology, teaching and learning strategies and learners' progression between the experimental and control groups.

6. The contextualisation of the results obtained and their comparison with the literature available.

Hopefully, the results obtained will shed some light on how washback affects students and contribute to filling in the gap identified by relevant scholars on the subject such as Tsagari (2006:56), who point out that the findings on washback studies focusing on students learning are disparate or too mixed to provide a definitive conclusion. The ultimate goals are to offer some valuable data that can be of use for stakeholders to make informed decisions related to content and course design, and to help learners of English to make the most of their time and learn in the best possible way.

4. METHODOLOGY

4.1. Research methodology

4.1.1. Classroom research

As a result of their social relevance, accreditation exams have come under the spotlight of EFL research because of the influence they have in the classroom and the participants in the teaching and learning process. Nevertheless, understanding how this impact takes place is problematic because:

> The classroom has long been considered one of the most difficult places to do research. The connection between input received by students and the output they produce is often characterised as a "black box", reflecting the idea that what actually goes on in this process is dark and unknown. (Madrid, n.d.:5)

Classroom research, which usually involves practical research focused on the classroom, tries to explain what happens inside the classroom, the direct and indirect influence of internal and external variables that have to do with the learner, the teacher and the English Language Teaching curriculum. This kind of research has proved valuable to improve teaching and learning and it has guided the present study, which tries to shed light on how the B2 First exam affects students in terms of attitudes, learning outcomes and autonomy. In order to do so, the teaching and learning context has been analysed and we have tried to include almost all the presage and context variables included in Madrid's (1995:60-62, 1998a, 1998b) framework for L2 teaching analysis and research (Figure 5).

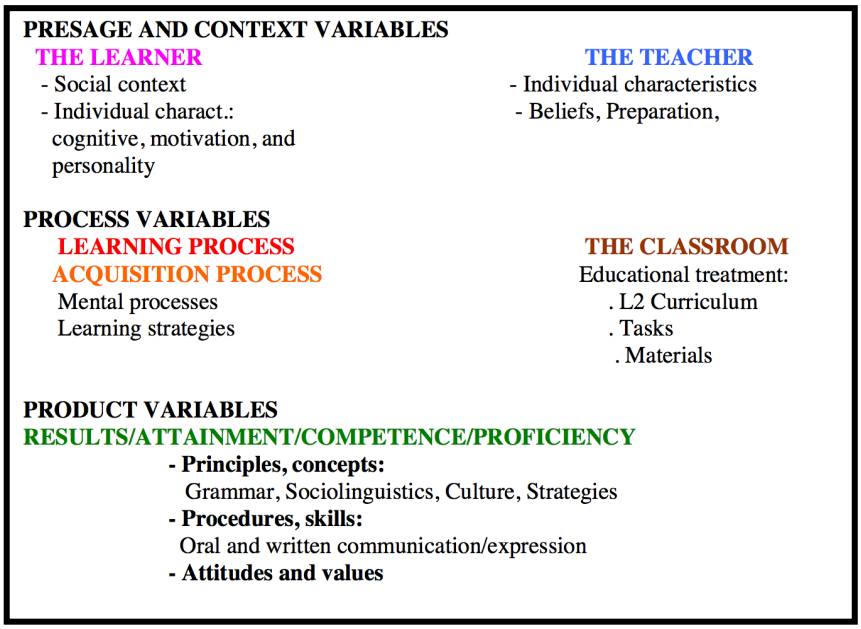


Figure 5. A framework for L2 teaching analysis and research (Madrid, 1995:60-62, 1998a, 1998b)

First, the learners, who are the cornerstone of this PhD Thesis, their social context and individual characteristics have been studied. Second, the teachers' characteristics, beliefs and preparation have also been taken into account. This is because according to Madrid (n.d.:6) both groups clearly influence what happens in the classroom. Without a doubt obtaining data about what happens there has also been vital for this study. Similarly, it has been necessary to look at the learning and acquisition processes and, in this sense, gaining greater knowledge about learning strategies has been one of the objectives of this study. Finally, this project aims to analyse the results obtained and the competence acquired in terms of concepts, skills and attitudes.

4.1.2. Approaches to research

Once the kind of research used in this PhD Thesis has been explained and justified, it is time to look at its characteristics and the approaches that are considered most suitable for it.

> The approach that we adopt needs to be suited to the kind of research we want to carry out, and for the variables we want to control. On some occasions, an observational process will be enough because the data we want to collect cannot be quantified, but on other occasions we may need to illustrate our findings with

> figures and statistics. So, the approach we adopt depends on the nature of the research we aim to do. Very often an eclectic position may be necessary. (Madrid, n.d.:11)

As Gosa (2004:62) points out, there is some debate regarding the difference and relationships between qualitative and quantitative research methods. In fact, it seems a problematic issue since "there have been many attempts to define qualitative research in the social sciences, and to determine whether or not it can or should be differentiated from something called quantitative research" (Mason, 2002:2). Authors such as Riley (1996:9) define qualitative research as the kind of research where one collects opinions, observations and wordy statements, rather than numbers. Nevertheless, Gosa (2004:60) claims that "this view seems limited because quantitative social data often is also made up of words which through the process of analysis are turned into numbers". As a consequence, she follows Mason (2002) and offers a series of labels that have been associated with qualitative and quantitative research.

> - Qualitative: subjective, hermeneutic, interpretive, naturalistic-inquiry, inductive, emic, holistic, data-driven, uncontrolled, unstructured, test-hypotheses, process-oriented, context-dependent and descriptive.
> - Quantitative: objective, nomothetic, positivist, experimental, deductive, etic, structural, theory-driven, controlled, structured, generate-hypotheses, product oriented, context-independent, prescriptive. (Gosa, 2004:60)

In the light of the above, the research methods in the present study can be seen as part of a continuum (Seliger & Shohamy, 1989:31; van Lier, 1989:34; Allwright & Bailey, 1991:43; all cited by Gosa, 2004:63), some of them are closer to the qualitative end while others are more quantitative in nature. The reason for this eclectic approach, deemed important by relevant authors such as Allwright and Bailey (1991:67) and Green (2007:73), is that, on the one hand, it is necessary to record *soft data* (Bueno González, 2015:82), e.g. opinions and attitudes towards English and, on the other hand, it is deemed convenient to obtain *hard data* (Bueno González, 2015:82) about learning processes and score gains, activities, etc. and to analyse it with mathematical and statistical procedures to make the research valid and reliable.

4.1.3. Researching into washback

Obviously, the subject of this project has determined the kind of research and the approaches used. The review of the literature on the matter has shown the limitations of previous studies and those imposed by the circumstances in which they take place. Furthermore, it has pointed in the direction this research should head towards.

> It is clear that the future direction of washback and impact studies to investigate the consequences of language testing needs to be multiphase, multimethod and longitudinal in nature […]. In addition, researchers should pay attention to the seasonality of the phenomenon, that is, the timing of research observations may influence what we discover about washback. (Bailey, 1999; Cheng, 2005; Watanabe, 1996; all cited by Tsagari, 2006:359)
>
> Unpacking the complexity of washback on students' learning outcomes by investigating learning including test preparation, as a mediated activity situated in context where different perspectives, perceptions and practices interact, would be particularly useful. For this purpose, the scope of washback research needs to be expanded to include more groups of participants such as parents and principals and by linking student data and data from other sources. (Cheng, Sun & Ma, 2015:462)

As already mentioned, this study has been designed to combine quantitative and qualitative research methods in a comparative study between experimental – groups of students enrolled on more exam-oriented courses – and control groups – groups of students enrolled on general English courses. It covered six months, which corresponded to the second half of one 120-hour programme and to one complete 60-hour programme. Information and data were collected at the beginning and at the end of this time span to compare outcomes and draw conclusions. Thus, it offers a longitudinal analysis of the relationship between teaching, learning and a well-known high-stakes exam paying special attention to students' perspectives. Such an approach has been suggested by authors such as Tsagari (2007:55) and Wall (2000:506; cited by Gosa, 2004:41). In addition, the study has gathered a wide range of information, from opinions and attitudes, obtained from questionnaires, to objective data. The latter has been collected from direct observation, measures

to assess students' progression in the exam, and alternative measures of the criterion abilities. It was considered relevant to include this type of data because early research into washback was criticised for lack of empirical data (Green, 2007:xi).

4.2. Research tools

When designing the present study, careful thought has been given to the research tools and to the participants, to what they could offer and to their limitations. The objective has been to obtain reliable and accurate data to have as many pieces of the jigsaw puzzle as possible. Therefore, a range of instruments and participants were necessary. Nevertheless, it is paramount to bear in mind the limitations that studies have; in an ideal world, questionnaires, interviews, observation, mock exams, lexico-grammatical exams and study diaries would have been used over an extended period of time. However, the range of research tools was reduced to make the study feasible.

The fact that this Thesis has been carried out in the framework of classroom research make Hamp-Lyons's (1997:299; cited by Gosa, 2004:40-41) words especially relevant. The expert claims that it is not enough to evaluate tests from our own perspective, or to evaluate them by including [only] teachers' perspectives. Thus, she highlights the need for studies about students' views and their account of the effect on their lives of test preparation, test taking and the scores. Similar views have been offered by Struyven, Dochy and Janssens (2005:331; cited by Zapata, 2016:96), who pointed out that "reality as experienced by the often forgotten student is an intervening variable [and] cannot be neglected if full understanding of student learning is the purpose of our educational research and practice", and by Bailey (1996), Alderson and Wall (1993), by Shohamy, Donitsa-Schmidt and Ferman (1996) and by Gosa (2004), all cited by Tsagari (2006:303), who refer to the *major role* played by students. In this sense, this study has counted on information offered by learners and teachers through questionnaires. Several well-known experts such as Wesdorp (1982), Alderson and Wall (1993), Li (1990), Shohamy (1993), Shohamy, Donitsa-Schmidt and Ferman (1996) and Watanabe (1996) have also used this tool. However, while questionnaires offer a personal account of testing, learning and teaching, classroom observation offers insight into what they actually do and this is probably one of the reasons why scholars such as Shohamy (1993), Alderson and Wall (1993), Alderson and Hamp-Lyons (1996), and Watanabe (1996) include it in their studies.

While these tools provide a great deal of details, some more objective data are necessary. In line with this, Tsagari (2007:59 and 2006:360) mentions that it is desirable to conduct studies which look at washback of a specific test from different perspectives in order to investigate the influence it exerts on classroom teaching and learning in depth. The present Thesis analyses the effect that B2 First may have on test takers. For that purpose, it also relies on tests scores – mock B2 First and lexico-grammatical tests – just like relevant scholars such as Wesdorp (1982), Hughes (1988), Alderson and Wall (1993) have done. The information obtained from them will be analysed bearing in mind questionnaires and observation, which may explain why individual variation in score gain may occur (Elder & O'Loughlin, 2003; Read & Hayes, 2003; Green, 2005; Humphreys et al., 2012; all cited by Allen, 2016:3).

4.2.1. Questionnaires

The value of questionnaires and their limitations are clear. On the one hand, questionnaires grant the opportunity of obtaining first-hand information simultaneously, which increases its accuracy (Madrid & Bueno, 2005). This is vital for the study as the population is composed of students and teachers from two different institutions attending lessons on several weekdays. What is more, participants can take their time to reflect and offer honest information. On the other hand, answers may be biased due to misunderstandings or because participants feel that what they say may turn against them. In addition, questionnaires may not be returned and this may compromise the reliability and generalizability of the study.

In order to make the most of the pros and minimise the cons several measures were taken. First, participants received the same questionnaire on paper. Using a digital format of the questionnaires was also considered because it would have made the data processing easier and faster. However, some students are under age and the researcher did not have access to their personal email addresses. Furthermore, students may not have returned the questionnaires if they had to fill them in at home in their spare time. Moreover, if questions had come up, it would have been impossible to answer them. For all these reasons, a hard copy of the student questionnaires – Entry and End-of-Course – was handed out in class for participants to fill in. Thus, the student questionnaires had to be short enough to be filled in in no more than ten minutes and questions should require concise answers. The piloting of questionnaires enabled the researcher to make them as relevant as possible and identify unclear questions or answers. It also offered guidelines in

terms of time, which was limited to 10 minutes. The questionnaires were piloted with a reduced sample of students, who did not take part in the main study, in November 2017. The researcher and a designated person were in charge of handing the questionnaires out, explaining the procedure and collecting them. Strict guidelines were followed to ensure that all participants received the same input in order to minimise potential biased responses. Regarding the teacher questionnaire, it was also piloted with a number of teachers who did not take part in the main study and its length suggested that it should be filled in outside the lesson time as piloting showed that teachers needed between 20 and 40 minutes to complete it.

As for the timing, questionnaires for students were applied in January / February 2018 and April / May 2018. This timing was deemed the most convenient because CEALM courses tend to start in January and finish in May and CEB students start the second part of the course in January and finish it in May / June. Two models of student questionnaires were used following Green's study (2006b:115) and other similar studies on the matter. Although they have the same format, the Entry Questionnaire (EQ) – see Appendix 1 – focuses on the participants' profile, their expectations about the course and the learning process, whereas the End-of-Course Questionnaire (EoCQ) – see Appendix 2 – has a retrospective function and enquires about the course, the learning process, test-taking strategies and the students' perception towards the test. Before receiving the EQ all learners had some teaching hours to be familiar with the level and the test and the end-of-course questionnaire was applied as close to the examination date as possible to reflect the intensity of the potential washback. The objective of keeping the application of questionnaires as distant as possible in time was to be able to record learners' progression, to witness the continuum from covert to overt washback (Prodromou, 1995) and obtain evidence of the seasonality of the phenomenon (Tsagari, 2006:359) since washback becomes more evident as the date of the test approaches (Cheng & Deluca, 2011; Zhan & Wan, 2013; all cited by Cheng, Sun & Ma, 2015:459), particularly when the test taking dates are externally determined and the stakes are high as it is the case of this study. Overt washback takes the form of purposeful preparation to maximise performance and classroom observation and questionnaires have been used to try and record this phenomenon.

In order to reduce bias, participants were informed about the fact that the information obtained from questionnaires and other instruments would be anonymised so that no results can be matched to individual participants once this research is published. Hopefully this may have

minimised students' anxiety over the consequences of their responses. Besides, the person in charge of applying the questionnaires explained how to fill them in, briefly justified the aim of the research, and clarified doubts following strict guidelines, as we have stated above.

As already mentioned, teachers were also given a questionnaire – see Appendix 3 – in May because of their potentially crucial role in washback. The objective of this questionnaire for teachers is to obtain greater insight into the activities carried out in class. Furthermore, research shows that teachers' perceptions of their students are accurate (Peña Jaenes, 2015) so they may be able to explain strategies used by their students, score variations, etc. Nevertheless, it will be thought-provoking to observe that students' practices and views differ from what their teachers believe them to be, as Gosa (2004:42) highlights. Besides, it will be interesting to see whether teachers and students' views towards the exam are in agreement or not. In this sense, Tsagari (2007:48) offers a list of studies which observed examples of both positions.

Finally, questionnaires responses were processed and analysed statistically. To do so, questionnaires were assigned a code that, while protecting the anonymity of the population, made it easier to process the answers. In the next lines, the meaning of the code will be explained. First, the type of questionnaire – entry or end-of-course – is specified, this is relevant as the two questionnaires elicit different types of information. Second, the code includes the acronym of the institution, which allows the researcher to make comparisons and explain data. For instance, it may offer some data to compare with initial findings of the study by Hamp-Lyons and Brown (2007), which revealed "few differences between the views of students who are preparing for TOEFL and those who are not" (Rahimi, Esfandiari & Amini, 2016:9). Third, a number identifies the teacher delivering the course as aspects such as the *teacher factor* (Alderson, 2004:x; Green, 2007:30; all cited by Peña Jaenes, 2015:80) may explain potential differences between groups taught by different teachers. In fact, Zhan and Wan (2016:373) referred to Gosa (2004:227), who also observed that the teacher's tight control over students' preparation seemed to mediate washback on learning hence supporting the idea that teachers play an important role as far as the presence of washback is concerned. Tsagari (2006:301) also identified teachers as an important factor in the presence or absence of washback. Fourth, the code includes the level of the group – B2.1 for Year 1 students and B2.2 for students in Year 2. This distinction is relevant because the courses under

study have a duration that ranges from 60 to 120[2] teaching hours and it is believed that students on average need two 120-hour modules to prepare for the level. Graddol (2006:96) suggested that 300-400 hours of study are necessary to raise one IELTS level, for example from the equivalent of B1 Preliminary – B1 following the CEFR – to the equivalent of B2 First – B2 following the CEFR. Green (2007:86) found out that to pass from one IELTS band to another the number of instruction hours could reach 500. He went on to suggest that higher levels of proficiency especially require a considerable amount of time and intensive preparation (Read & Hayes, 2003; Green, 2005; all cited by Allen, 2016:3). Peña Jaenes (2015:61) points out that most of the professionals who took part in the research believe that adult learners need more than 120 teaching hours to pass from a B1 to a B2 level of the CEFR. The last figure of the code is a number that identifies each student – in the case of teachers' questionnaires this number is 0. All the research tools applied to the students are identified with this number in order to relate relevant aspects such as score variation and strategies used or self-learning time.

As already mentioned, three models have been used: the entry questionnaire and the end-of-course questionnaire for students, and a questionnaire for teachers. Students' questionnaires have a very similar design and both of them are written in Spanish – the students' mother tongue. The reason for choosing this language is that, although the sample is composed of B2 students, there may be questions that could cause problems. Moreover, writing the questions in English would not affect the research. As a result, it was decided to write the questionnaires in Spanish to minimise the risk of obtaining inaccurate answers due to linguistic problems (Peña Jaenes, 2015). As for the teachers' questionnaire, it is in English because using that language is unlikely to cause any problems.

2 CEALM courses have a duration of 60 instructional hours and last for one semester while CEB courses have a duration of around 120 teaching hours and last for two semesters. To be able to make a comparison between CEALM and CEB, the module in the first and the second semester taught at CEALM have been categorised as Year 1.

4.2.1.1. Student questionnaires

4.2.1.1.1. The Entry Questionnaire

It has two different sections and a short introduction which explains why students are given the questionnaire and how the information they provide will be used and by whom. The questionnaire is structured as all the items are closed-ended questions. This is because the author wanted to reduce the number of unrelated answers. Besides, it also minimises problems related to the understanding of the question.

Section 1 focuses on the participants' profile – age, educational experience and current situation, language proficiency, reasons for studying English, beliefs, and other contextual circumstances, which according to Tsagari (2006:360) is worthwhile to look at for washback and impact studies. Information such as time available for personal work (Allen, 2016:15) is valuable because it may influence the potential washback. Other relevant variables are learners' interest (Zhan & Andrews, 2014; cited by Allen, 2016:15); and the importance given to speaking English (Tsagari, 2007:48) because, as Cheng (1997 and 2005; cited by Allen, 2016:15) points out, washback is impeded through lack of perceived importance. Participants were asked about their perception of B2 accreditation exams in general and B2 First in particular following Tsagari (2007:55) and Xie and Andrews (2012:62), who refer to the role that test value plays when understanding the mechanism of washback. Assessment expectations – together with course and outcomes expectations – have been considered when investigating the washback on learners as a way to research into the consequential validity of the IELTS test (Brown, 1998; Elder & O'Loughlin, 2003; Mickan & Motteram, 2009). In fact, Xie and Andrews (2012:51) cite Gosa's study in 2004 and claim that students' expectations of assessment is the single most important factor in explaining students' perspectives on teaching and learning activities. They go on to mention Wigfield and Eccles (2000:68), who claim that expectancy-value theory posits that individual's choice, persistence and performance is determined by the their expectations about how well they will do on the activity and the value they give to it. This connects with other questions in the first section about what students feel they want to do and can do. They argue that if they feel they want to pass the test and are able to do it they are more likely to engage in the task and perform well on it. Students' answers to these questions are thought to be influenced by their personal objectives and perceptions of the task demands (Xie & Andrews, 2012:53) since, according to Booth (2012; cited

by Cheng, Sun & Ma, 2015:459) learners' aims and actions contribute greatly to the effect of the TOEIC on learning. In fact, Tsagari (2006:305) concluded that the acquisition of the certificate was a far more powerful motivation for language learning in the exam preparation classroom than a desire to learn the language for communication or integration with L2 speakers. A similar view was observed by Bailey (1996:269; cited by Luxia, 2005:154), who states that a student's immediate goal is often to achieve a given test score or to exceed the previous one. But presumably the students' long-term goal is one they share with their language teacher, i.e. to enhance their language proficiency. This is particularly relevant for our study because a high percentage of the participants are teenagers, who may be forced by their parents to attend the lessons and do not have attaining the B2 level as one of their priorities. Besides, some of them are attending B2.1 modules and do not see this as an attainable goal in the near future.

Past learning and test taking experience also play a relevant role because they may enable learners to use some of the strategies already acquired for the new test (Allen, 2016:4) and may have a major influence on their choice of the types of test preparation activities (Stoneman, 2006; cited by Chen, Sun & Ma, 2015: 447). This is also interesting for our study because many of the students preparing for B2 First have already sat the B1 Preliminary exam. In fact, washback studies have revealed that these variables mediate the washback to learners and learning (Gosa, 2004; Xie & Andrews, 2012; Zhan & Andrews, 2014; Zhan & Wan, 2016; all cited by Allen 2016). There will also be questions about the stakes of the exam and about perceptions of test design because, as Xie and Andrews (2012:61) indicate, both factors influence preparation simultaneously although perceptions of test design seemed to be more powerful. Answers to these questions were analysed together with students' report of their strategies and personal work and preparation – part of the end-of-course questionnaire. This is because of the close relationship that may be observed between students' perceptions of the test design and content and how students prepared for the test (Xie & Andrews, 2012). This is what Watanabe (2004; cited by Zhan & Wan, 2016:371-372) calls *specific washback.*

Section 2 enquires about what they expect to find in the language programme they are enrolled on and about whether they are aware of the course objectives and assessment. According to Green (2006b:123-125) learners arrived on the courses of different types with comparable expectations and they understood that test preparation includes a traditional focus on lexico-grammar including elements specifically targeted at the test. Thus, these data are relevant to

create a profile of the students and to compare and contrast if differences arise between the groups taking general English lessons and the groups whose objective is to sit B2 First in the near future. Besides, comparisons with other studies in the literature will surely be of use. Another aspect that is also checked is whether learners are more exam-oriented than their teachers (Lumley & Stoneman, 2000; cited by Tsagari, 2007:50). Wall and Horak (2006; cited by Tsagari, 2007:49) found that, according to teachers, students only wanted to do exercises for the TOEFL. This contrasts with Tsagari's study (2006: 273) because it found that the practice of revising and doing past papers in the period before the exam was counterproductive for some of the diarists as this made them feel stressed and resulted in lowering their self-esteem by creating an exam-like atmosphere in the classroom, making them believe they were going to fail.

4.2.1.1.2. The End-of-Course Questionnaire

It also has two different sections and a short introduction which states the reasons why they are given the questionnaire and how the information they provide will be used and by whom. The questionnaire is semi-structured or mixed as eight questions out of 39 are open-ended. This is because the author wanted to minimise problems due to misunderstandings and unrelated answers. However, in some cases it was interesting to have open questions to have more information that had not been expected or so that students could explain their previous answers.

Section 1 focuses on learners' perspectives on their language course, their outcomes (Green, 2006a) and perceptions about improvement. In particular, questions such as the type of activities carried out in the lessons may help us understand test score gains and give evidence for washback, as Mickan and Motteram (2008:39, cited by Mickan & Motteram, 2009:20) concluded, in the sense that a teacher preparing students for IELTS faces a choice "along a continuum between developing language skills, as in general or academic language classes, and training for taking the test". They (Mickan & Motteram, 2008:8; cited by Allen, 2016:3) found that most frequent activities were test practice, skills focused activities and explanations of the format and content of the IELTS components and test-taking strategies. The same detrimental effect on the breadth of, or the variety to be found within the curriculum has been observed in preparation courses for TOEFL (Wall & Horak, 2011) as well as other regional English language exams (Gosa, 2004; Stoneman, 2006; Shih, 2007; Xie, 2013; Zhan & Andrews, 2014; Zhan & Wan, 2016; all cited by Allen, 2016:3) and in other studies on washback on learning such as in Michaelides (2014). Xie (2013; cited by Zhan &

Wan, 2016:374) also has a negative view towards examination-oriented preparation as she claims that while it can boost test scores, it does not help the students in the long run because they are not beneficial for the development of the language skills that students will need in real life. Nevertheless, this author also acknowledges that explaining test content and demands is not necessarily negative, quite the opposite, because knowledge about the test specifications and format could give students a sense of direction and control and they could feel more confident towards test-taking (Xie & Andrews, 2012:63) and, as a result, plan for assessment to perform better and make the most of their language skills and their abilities to self-assess (Muñoz & Álvarez, 2010:38). "Self-assessment can be related to beneficial washback because it helps learners to develop internal criteria for progress and success and thus develop their autonomy". Learner autonomy and responsibility are directly related to Alderson and Wall's (1993:120) 10th washback hypothesis that "a test will influence the degree and depth of learning" (Muñoz & Álvarez, 2010:38). What is more, it could improve the perception that students have about the course as they could see that what is done in class really serves a purpose and helps them improve. Failing to explain the relationship between classroom-based activities and assessment – of any kind – leads students to perceive a disconnection as they believe that the course content taught and practised is very different from what is included in the test (Zapata, 2016:101). For this reason, several questions to determine students' ability to assess their skills have been included and the information they elicit has been compared with the data obtained from similar questions in the entry questionnaire in order to record whether there has been a progression in this aspect. In addition, students were inquired about the connection between assessment and the activities carried out in class and also about their degree of satisfaction with the course they attend.

This section of the questionnaire also enquired about potential differences in preparation for the first mock B2 First and the second and the reasons for them. This is relevant because Allen (2016:12) points out that the experience of the initial test and its results is the primary driving force behind learners' strategies employed for the second test. These findings are in line with previous evidence obtained by Tsagari (2007) and Zhan and Andrews (2014) of how tests can raise learners' awareness and motivate learning. What is more, these findings show how washback is influenced by learners' experience of the test together with previous learning experience. The last part of this section asked students questions about the perceived difficulty of B2 First – it was included in the entry questionnaire as well and has been used to check whether the perceived difficulty of the

exam has evolved as they have obtained more information about the exam – and their intention of taking the official exam. These questions have been analysed together with exam results because it is interesting to see if despite scoring low on the screening tests, students are willing to take the official exam (Tsagari, 2006:276) and if they feel that the exam made them work harder to achieve good scores (Alderson & Wall, 1993; cited by Tsagari, 2006:304 and Cheng, 1998; cited by Tsagari, 2007:50).

Section 2 is aimed at B2 accreditation exams candidates and offers data about key activities for success, their personal work and their perceptions of score gain (Elder & O'Loughlin, 2003; Green, 2007; O'Loughlin & Arkoudis 2009; Humphreys et al., 2012; all cited by Allen, 2016:3). The latter together with their perspectives of the course and outcomes – included in Section 1 – is part of the research into washback as a way to reflect on consequential validity. Participants are asked about the most effective strategies for exam success and the answers to this question are compared with those obtained by scholars such as Mickan and Motteram (2009), Xie and Andrews (2012), and Zhan and Andrews (2014), who have explored the relationship between test design, exam use or beliefs about them and test takers' strategies (Zhan & Wan, 2016:265). This question together with the item asking about how students prepare for the test may offer some insight into whether the test provides a stimulus which orients students towards the study of productive skills as it was observed by Allen (2016), and whether they change their studying habits depending on the type of test they are going to take, e.g. multiple-choice tests, open-question tests. What is more, it may show whether students finally come to terms with exam practice because they see the value in it and if they are happy to make sacrifices to pass the test (Tsagari, 2006:277).

Furthermore, questions regarding test score gains and the inconsistency of test scores found in some studies such as the one carried out by Mickan and Motteram (2009:20) is paramount and requires intensive analysis. Not only because they may identify flaws in teaching and learning but also because they may have an impact on learners' motivation and feelings. In this sense, Murray, Riazi and Cross (2012:579) suggest that positive attitudes produce beneficial effects while real or anticipated negative experiences can lead to attitudes that erode confidence and potentially impact negatively on performance. Yet test takers' reactions have not generally been seen as central to test validation (Elder, Iwashita & McNamara, 2002; cited by Murray, Riazi & Cross, 2012:579). Logically, positive attitudes towards the test can promote a sense of empowerment and foster effective learning opportunities (Murray, Riazi & Cross, 2012:580). For this reason, it was deemed

necessary to ask students about score gain and the reasons behind it and how they felt about negative results – Section 2 of the questionnaire.

Section 2 also includes questions about *folk-knowledge* (Bailey 1999; cited by Allen, 2016:13). This is because it is suggested that peers work as advice givers and play a role along with official test information that can have an impact on test takers preparation and the materials they choose and, as a result, mediate the potential for washback (Allen, 2016:13). Similar findings were obtained by Mickan and Motteram (2009; cited by Allen, 2016:14) in non-instructed English as a Second Language (henceforth ESL) contexts. In fact, they found that "test takers were dependent on the assistance and feedback of others, without which they felt incapable of preparing for the test". Cheng, Sun and Ma (2015:446) also recognised the importance that peers and teachers have when it comes to shaping students' perceptions of tests. That is why a question about the influence that teachers exert on students seems appropriate and it is connected to the usefulness that, according to students, teachers' advice, recommendations and examination practices may have. This is relevant because Xie (2013; cited by Zhan & Wan, 2016:374) points out that teachers play a vital role when it comes to predicting the types and degrees of washback that the Computer-based English Listening and Speaking Test (CELST) have on the participants.

The questionnaire is also designed to be a tool for reflection and aims to raise awareness among participants about how to improve. Hopefully, questions enquiring about the reasons for poor performance may help identify weaknesses and offer hints for improvement. In this regard, Vikiru (2011:20) found that subjects' comments on test success and standards suggest a lack of personal agency in the face of assessment. In the same line, Mickan and Motteram (2009:20) claimed that "the apparent intractability of success for some students is a fundamental issue to be addressed through intensive analysis of individuals' test taking". Besides, it may be a way to motivate students if the answer to the question about whether the English they are learning can be applied to real-life contexts is positive. This is connected to the question which enquires about whether they will continue studying English once the exam is over. Such an issue would be a useful addition to the field as it would offer greater insight into the role that high-stakes tests play in motivating students to learn and whether they can help sustain students' motivation for learning after the exam (Bailey, 1996:269; cited by Luxia, 2005:154) or if, on the contrary, the long periods of stress and anxiety and other negative feelings, especially in the case of low performers, have long-term consequences and impact on their later performance in the language (Harlen & Crick,

2003; cited by Tsagari, 2006:304). Finally, the last item of the questionnaire aims to summarise students' feelings about the test design and preparation because it asked students whether they believe the exam reflects all aspects of their study, surely something interesting because Cheng (1998; cited by Tsagari, 2007:50) found that most of the students provided a negative answer.

4.2.1.2. The Teacher Questionnaire

Schön (1983, 1987, 1990; cited by Madrid n.d.:37) points out that teachers have two distinct kinds of professional knowledge. On the one hand, what Wallace (1991:12; cited by Madrid, n.d.:37) called *received knowledge* obtained from training, courses, etc. and, on the other hand, the *experiential knowledge* (Wallace, 1991:13; cited by Madrid, n.d.:38) based on the teacher's experience in the classroom. This teachers' questionnaire tries to elicit this valuable information that can help contextualise the other data gathered in this study.

The teacher questionnaire – see Appendix 3 – has five sections and a short introduction which states the reasons why they are given the questionnaire and how the information they provide will be used and by whom. The questionnaire is semi-structured, as 11 questions out of 43 are open-ended. This is because the author wanted to minimise problems due to misunderstandings and unrelated answers. However, in some cases it was interesting to have open questions to have more information that had not been expected.

Section 1 tries to gather some information about teachers' experience and education background. It includes questions about Cambridge training, about their current teaching practice and their students' age. Section 2 enquires about the teachers' perception of the language courses and accreditation exams. They are asked about the potential differences between general English courses and test preparation course and there is also a question about the number of instruction hours necessary to pass from a B1 to a B2 level. This question was also included in a previous study by Peña Jaenes (2015) so it will be interesting to see if the opinions changed after the implementation of the revised version of the exam in 2015. There is also a question about motivation and self-assessment as both aspects have been regarded as vital by several scholars.

Section 3 elicits information about the lessons and the activities carried out in class. Some of the questions can also be found in the end-of-course questionnaire for students, for instance, those referring to the objectives, the marking criteria and methods, the connection between the

classroom activities and the exam and feedback. Furthermore, there is a question asking about motivation and about how often materials are used and contents included in the lessons. It also enquires about grammar and vocabulary because, as we saw before, students may expect to find a traditional approach towards grammar and vocabulary so asking teachers about the importance given to these aspects in class will be useful, even more if we take into account that a lexico-grammatical test will be administered as part of this study. The answers obtained have been compared with the students' answers and also with findings in other studies such as in Tsagari (2006) and in Hamp-Lyons and Brown (2007), where it was revealed that teachers and their students had different perceptions and attitudes towards the TOEFL.

Section 4 focuses on their learners and their learning process. There are questions about self-directed learning and about their students' performance and their progression. Section 5 has only been answered by teachers delivering B2 First courses. It asks about factors that could be said to be relevant for success in B2 First; about whether they think that a candidate who passes the B2 First has a B2 level; and about the effect Cambridge exams have on their lessons and on their students. The questionnaire includes questions about the problems students encounter according to teachers and asks them if the problems experienced in the different Cambridge papers are the same as the ones experienced when they work on the skills in general.

4.2.2. Tests

Classroom tests are very useful research instruments for obtaining data on students' competence and performance. They are used to check the progression, if any, of experimental and control groups and to assess if a given intervention programme has had any effect or not (Bueno González, 2015:79). In this study two tests are used. On the one hand, two sample tests of Cambridge B2 First and the results obtained by the students have been used to check how effective preparation courses are in improving students' performance in the exam. On the other hand, a grammar test and a vocabulary test, which are used to check if students in the control and the experimental group progress in a similar way in terms of general English skills. Further information about both formats can be found below.

4.2.2.1. B2 First mock tests

In order to check one of the three key research hypotheses, it was necessary to analyse students' performance in B2 First exams and study their potential progression. For that purpose, Test 6 and Test 7 of Book 2 (Cambridge University Press and University of Cambridge Local Examinations Syndicate, 2016) for adults have been used. The reason for choosing the adults' version was mainly practical as it was the most recent book of sample papers at the time and the researcher wanted to make sure that no student had already done any of the exams.

The first mock test was applied between January and February 2018 and the other was applied three months later, in April and May 2018. The reasons for choosing these dates were that, on the one hand, the students need some instruction time to get familiar with the level and, on the other hand, the researcher wanted to keep both exams as distant in time as possible to maximise students' progression. In addition, the last mock test ought to take place as close to the official exam as possible to be able to capture the potential washback. As the overall duration of the test is 3 hours 30 minutes, the different papers were applied in different weeks in order to combine it with normal lessons and adapt it as much as possible to the pace of the lessons in the institutions. The different papers were applied during the same weeks and under exam conditions so that the situation is as similar as possible in both institutions. The researcher was present most days and careful instructions were given to the teachers who were to invigilate – see Appendix 4.

Before the first mock test – in January 2017 – a pre-test preparation activity was carried out with those students enrolled on the general English course. First, a document with the format of the exam and recommendations was handed out to all the students – see Appendix 5. The information it contains is based on the B2 First Handbook for Teachers (University of Cambridge Local Examinations Syndicate, 2019), on different websites (University of Cambridge Local Examinations Syndicate, 2017 and Cambridge English De, 2012), and the researcher's teaching experience. It is divided into two sections: Section A offers a description of the exam papers with the different parts and the marks awarded; in Section B there is a list of bullet points with tips for each exam paper. Later, a shortened, demonstration version of the B2 First was administered to all the groups. For that purpose, the website Examenglish (Exam English Ltd., 2014) was used. The reason for choosing an online website is that the format is more interactive and students can share their doubts and find it more enjoyable than sitting in front of a piece of paper. For the Reading and

Use of English and Listening papers one item of each part was answered in class. As for the Writing component, students were provided with a sample test so that they could see what was expected of them. Finally, they watched a video of a B2 First Speaking paper (Cambridge English, 2014). This one-hour preparation was considered important as a way to partially offset the *practice effect* that according to Allen (2016:2) has hampered washback studies into learning whereby students generally score higher on second and successive administrations of a test merely due to greater familiarity with the test itself (Robb & Ercanbrack, 1999:8). This preparation was especially useful for the purpose of this study because it includes Year 1 and Year 2 students as well as general English students and their degree of familiarity with the exam differs.

4.2.2.2. Vocabulary and Grammar tests

While the data from the questionnaires outlined the context of the study, and the results in the mock B2 First exams enabled an analysis of students' progression in the test, it was necessary to use an alternative measure tool and compare the results obtained with Cambridge scores for benchmarking purposes.

The idea was to design a test that could be easily integrated into the language courses under study. This is because the institutions which kindly agreed to take part in the study could be reluctant to involve their students in lengthy exams with apparently no use for them and the same could be true for students, who may not be willing to do one more test, which in addition does not resemble the one they intend to take at the end of the academic year.

Applying a grammar and vocabulary test was deemed a good alternative measure because testing these linguistic aspects by using a multiple-choice format is quite practical as it is objective, easy to mark and not very time consuming. What is more, the CEFR (Council of Europe, 2001) includes the lexico-grammatical competence for all the levels of proficiency and so do Cambridge English Qualifications (University of Cambridge Local Examinations Syndicate, 2013a; 2013b). Consequently, it could be seen as another method to record students' progression.

To reduce bias, it was considered that choosing a test which had nothing to do with any of the textbooks used in the courses would be the fairest and most neutral alternative. The *Vocabulary Levels Test* (VLT) (Xue & Nation, 1983; Nation, 1990) has been described as the best available measure of vocabulary size (Schmitt, Schmitt & Clapham, 2001) and has been widely used

both amongst teachers and researchers to offer an estimate of vocabulary knowledge of learners (Read, 1988; Cobb, 1997; Schmitt & Meara, 1997; Laufer & Paribakht, 1998; Read, 2000; Shiotsu & Weir, 2007; all cited by Green, 2007:147). The original versions of the test are divided into five frequency bands: 2,000-word level, 3,000-word level, 5,000-word level, 10,000-word level and academic vocabulary – although in the version used – see Appendices 6 and 7 – the word level is not specified because, on the one hand, it was believed that this information would not bring any benefits to the students and, on the other hand, it could lead them to think that the 10,000-word level is above their level and hence, they may not try as hard to answer correctly. Each level contains 10 clusters with six words in a column on the left and the corresponding meaning of three of these words in a column on the right. Learners have to match each meaning in the right-hand column with one single word from the left-hand column (Kremmel & Schmitt, 2017). The VLT has two parallel test versions, which makes it especially suitable for this study, which requires the application of one vocabulary and one grammar test at the beginning of the course and one of each at the end to compare potential improvement. While it is true that Kremmel & Schmitt (2017) point out that the two versions of the test are not equivalent enough to measure the learning gains of individual learners and that the same version should be used twice, Green (2007:149) claimed that there were arguments for using two versions rather than a single repeated version. First, participants accepted doing the test if it contributed to learning so feedback was considered necessary. As a consequence, if the same test were repeated, the second exam would only test students' memory. Second, the test was used for group comparison and not for high-stakes decisions about individuals. The same reasons can be used to justify the use of the VLT in this study. This is because, on the one hand, the idea is that all the data collection tools are useful both for the research and for the instruction and learning process and, on the other hand, the study and the comparison group took the same tests at the same time so comparisons could be made. Furthermore, Green (2007:148) found that the VLTs did appear to reflect learning gains made over the course of instruction, and to reflect the greater gains made by learners over the longer of the two courses. In addition, the version of the VLT used was trialled and its results analysed statistically and changes were made to make the instrument suitable for the purposes of this study. Finally, it must be noted that there is new version of the VLT (New English Vocabulary Test, henceforth NVLT) and although it was analysed and considered for its use in the present study, the fact that it only had one version did not make it suitable for this.

After selecting the test, another factor came into play: time constraints. Since the participants are all at B2 level because they have done a placement test or an achievement test to join the programmes, it was decided that the 2,000-word level and the academic part would not be administered. The former, because most of the words have been categorised as below B2 level by the tool English Vocabulary Profile (Cambridge University Press, n.d.) and hence it can be understood that the students are likely to know them, and the latter, because most of CEB students are secondary school students and hence have not had contact with academic English so they would be at a disadvantage. What is more, this would reduce the time necessary to complete the test and would make it more practical. A trial of the original versions of the VLT was conducted in a B2 group and statistical analyses were applied and the content of the original versions of the VLT test was revised to obtain a Cronbach's Alpha value of 0.86 in the Entry Vocabulary Test (EVT) and a Cronbach's Alpha value of 0.65 in the End-of-Course Vocabulary Test (EoCVT). As for discrimination, the distractors of those items with a discrimination index below 0.3 were modified. Regarding facility values, distractors were modified if the value of the item was below 0.2 or above 0.8. In terms of duration, the trialling suggested that the vocabulary tests could be completed in 20 minutes.

As for the Grammar test – see Appendices 8 and 9, it was inspired by the CEFR and was designed using resources obtained from Cambridge University Press website (Cambridge University Press, 2017), *Advanced Grammar in Use* (Hewings, 1999) and *Intermediate Grammar in Use* (Murphy, 2004) and its content is based on the grammar points included in four B2 textbooks: *English File Upper Intermediate* 3rd Edition (Latham-Koenig & Oxenden, 2014), *Gold First* (Bell and Thomas, 2014), *Complete First Certificate* (Brook-Hart & Owen, 2011) and *Objective First* (Capel & Sharp, 2014). The idea was to identify the grammar points that can be classified as typically B2 level. As the four B2 textbooks included the same content points, short-listing the grammar to include was straight forward. Two equivalent versions of the test were created with 25 four-option-multiple-choice questions. As with the vocabulary tests, the grammar tests were piloted with a group of B2 students and statistical analyses were applied. As a result, their content was revised. The Cronbach's Alpha value of the Entry Grammar Test (EGT) and the End-of-Course Grammar Test (EoCGT) is 0.65. The items with a facility value below 0.2 and above 0.8 or with a discrimination index below 0.3 had their distractors modified. In terms of duration, the trialling suggested that the grammar tests could be completed in 15 minutes.

The last question to answer was whether the Grammar and the Vocabulary tests would be applied on the same day. At first, it was thought that doing both of them one after the other would be the best idea because teachers would not need to devote time in two different lessons. Nevertheless, after working out the amount of time that is necessary to do both tests, it was considered that spending more than 30 minutes of one-hour lessons would be too much. For this reason, the tests were applied on two different days in the first three weeks of the course so that the teachers can use the exams as diagnostic information and find this instrument useful for their lessons and their students.

4.2.3. Observation

Traditionally, observation has been considered as a useful instrument to study what happens inside the classroom in a systematic way (Madrid, n.d.:26). This is because the researcher can observe a series of behavioural patterns in the classroom, which help understand the complexity of the teaching and learning process (Bueno González, 2015:76). In fact, washback studies have relied on classroom observation to provide detailed information about what teachers and students actually do in the classroom and to complement other methods (Cheng, 2010:43) because, according to Green (2007:137) and Bailey (1996, 1999), among others, information obtained by asking participants about their behaviour does not provide enough evidence to demonstrate that washback occurs and, as a result, empirical evidence of what occurs in classrooms is also required to contextualise, corroborate, or correct data (Alderson & Wall, 1993; Wall, 1996; Turner, 2001; Watanabe, 2004; cited by Green, 2006a:334). Nevertheless, classroom observation may have its advantages and disadvantages, like any other research instrument. The upside is that it allows the study of a phenomenon at close range with many of the contextual variables present, but the downside is that the closeness may introduce biases which may affect the researcher's objectivity. What is more, the presence of the observer in the classroom may alter the behaviour of the subjects observed (Seliger & Shohamy, 1989:162; cited by Bueno González, 2015:76).

Green (2006a:334) suggests two options for observation: the first one involves watching a small number of participants intensively over a sustained period of time (Alderson & Hamp-Lyons, 1996; Read & Hayes, 2003) whereas the second option increases the number of participants being

watched during lesson time and has a broader perspective (Wall, 2005; Hawkey, 2006). For this study, the second option has been chosen for practical reasons. A representative sample – two of the three groups at CEALM and four of the eight groups at CEB – of the groups that took part in the study have been observed – teachers were informed in advance – during two non-consecutive weeks. However, the researcher was not present in the room because she was teaching at the same time and audio recording was necessary. Although at first this was seen as a disadvantage, it turned out to be positive since the students did not feel they were being observed – so one of the abovementioned drawbacks was minimised – and the lessons were more natural.

When opting for classroom observation there are two possibilities: spontaneous or structured observation (Bueno González, 2015:76). In this case, the second one has been selected because the researcher has concentrated on pre-fixed issues derived from the content of the questionnaires. For observation to offer accurate and objective data, it is necessary to use a well-designed instrument. Spada and Fröhlich (1995) created the Communicative Orientation to Language Teaching (henceforth COLT) observation schedule. Although this observation instrument has been widely used in washback studies (Watanabe, 1996; Cheng, 1997 and 2005; Burrows, 1998; Read & Hayes, 2003), Read and Hayes (2003) and Green (2007:138) realised that the COLT schedule needed some adaptation, the reasons being that first, the COLT schedule does not capture references to test-taking strategies; second, the description of materials provided for by the COLT schedule is not sufficiently sensitive; third, it does not include a section for homework activities. For this reason, the authors chose to supplement the COLT with the draft IELTS Impact Study (henceforth IIS) observation schedule. This instrument was commissioned by University of Cambridge Local Examinations Syndicate from the University of Lancaster in 1995 at the inception of the IIS. It includes lists of text types and activities that were likely to occur in preparation classes and is specially designed for IELTS (Green, 2007:138-139). Nevertheless, further adaptation became necessary for Green's study (2007) and the changes implemented were the result of piloting. These adaptations are in line with the Spada and Fröhlich's (1995) view as they recommend that the schedule be adapted to the specific purpose of defined research objectives as they are tools that should serve rather than direct research. In the design of the observation schedule (see Appendix 10) used in the present study, Flanders' Interaction Analysis Categories (henceforth FIAC) (Flanders, 1970), Foreign Language Interaction (henceforth FLINT) (Moskowitz, 1971), COLT and Green's adaptation of COLT and IIS observation schedule have been used together with guidelines offered

by Mickan & Motteram (2009:17). Besides, the author has had access to teachers' lesson plans and used this information, together with the questionnaires and her own experience to design a tool that is in line with the lessons in both institutions.

The schedule includes some data that identify the group observed – sheet number, institutions, level, group and date. It also includes a table and some instructions to complete it. The table is divided into five main categories: time, activities and episodes, content, test, materials and homework. The section entitled activities and episodes is subdivided into teacher actions, student actions and approach. Examples of activities can be warm up, homework correction, grammar or vocabulary explanation, drilling activities, setting homework and house-keeping – attendance, exam dates, discipline issues, etc. By episodes the author understands giving instructions, asking for clarification, explaining doubts, observing, correcting activities or giving immediate feedback. Finally, the approach refers to the interaction in the class such as group activity, individual activity, working in pairs, and also to the approach of the task, i.e. if it is a process task to work on writing or a product task.

The second main category is content and refers to the skills or contents on which the activity is focused: speaking and oral interaction, writing, translation, reading, listening, grammar or vocabulary, culture, and pronunciation and phonetics. These subcategories are in line with the contents mentioned in the students' questionnaires. In some of them more detail can be added. For instance, in the case of speaking and oral interaction, the observer can add if it is a debate or if it is student-teacher interaction or student-student interaction. As for writing, some text types, mainly those tested on B2 First, have been given as examples although other text types can be added. Regarding phonetics or pronunciation, it is stated that this should be the main focus of the task and should not be part of feedback.

The third category is entitled test and includes Cambridge-related information or information about any other examination institution. It also offers the opportunity of providing more details about the type of information given such as format, strategies, feedback, practice, and performance. If the teacher and the students are talking about the format, they mention aspects such as the parts of the exam or the marking system; if they talk about strategies, this has to do with tips and recommendations to make the most of one's abilities; feedback is also included here because in some cases teachers may create an activity using feedback to highlight mistakes or

positive aspects and this may not happen exactly after the students' performance but some days later as can be the case with speaking or writing exams. Exam practice refers to exam tasks done in class.

The fourth section has to do with the materials used and it is subdivided into textbook – student book, work book or CD-ROM, authentic material taken from any source, exam preparation material including books or websites, and extra support material such as photocopies. The observer needs to explain the type of material and keep a copy of it if possible. In the case of the textbook, it is necessary to write the page and the activity. The last section is homework and it is subdivided into textbook homework, exam preparation or other. In the case of exam preparation, it could be for the official exam or for the achievement test at the end of the course and it should be specified. Finally, there is a section for comments.

The observation schedule has also been trialled in a B2 group.

4.2.4. Quality control

When one carries out a study, one realises that it is essential to design the research tools well because otherwise all the effort will be wasted. Equally important is to obtain results that are reliable, valid and that can be generalised because without these qualities any study is useless. The next section focuses on the measures taken to guarantee the quality of the present study.

4.2.4.1. Triangulation

Triangulation (Denzin, 1970:472; cited by Madrid, n.d.:72) is necessary if an accurate picture of a particular phenomenon is to be obtained. It can take several forms and most of them have been used in this project as explained below.

First, data has been triangulated because they have been collected at the beginning and at the end of the study in order to be able to compare results and analyse the seasonality of the washback. Although the fact that the researcher does not want to disturb the pace of the lessons makes it more difficult to coordinate the sampling process, data has been gathered in similar circumstances in both institutions. Second, there is methodological triangulation because a range of methods, e.g. observation, questionnaires, tests, has been used to collect data and they have offered information from different perspectives that has complemented each other. This type of

triangulation has been followed in recent research as mentioned by Green (2007:xi). Finally, the researcher has analysed and studied the information with an open mind and considering different possible interpretations that could explain the results, thus ensuring theoretical triangulation. Due to the nature of this project, a PhD Thesis, investigator triangulation was not feasible but the researcher has counted on the help and guidance of her Thesis tutor.

Bearing in mind the limitations posed by any study carried out in different institutions which accept to take part in it without any benefit obtained, the triangulation procedures try to guarantee the reliability and validity of its findings.

4.2.4.2. Reliability

Just like the results of an accreditation exam, the results of any study and hence the research procedures must be reliable, in other words, they need to be consistent and replicable both over time and across the variety of people who might use them (Madrid & Bueno, 2005:660). In this project, internal reliability has been guaranteed since the data collection procedures have been followed consistently as the application of questionnaires has been closely monitored by the author to ensure that no additional information is provided to the participants and most of the questions are structured to reduce the number of unrelated answers. As for observation, the lessons have been recorded so it is possible to listen to them more than once and clarify doubts. Besides, the observation schedule includes instructions to avoid misunderstandings. Finally, the tests – mock B2 First and the lexico-grammatical tests – have been administered under strict exam conditions.

4.2.4.3. Validity

According to Bueno (2015:86) validity refers to the extent to which the experiment investigates what was intended to measure. In this sense, the description of the research tools above shows that they have been designed to answer the research questions of this study not only when it comes to formal aspects – face or external validity – but also in terms of construct validity – see the last paragraph of this section for more details. As for internal validity, Madrid and Bueno (2005:660) claim that, "a study is said to have internal validity if the outcomes of the experiment can be directly and unambiguously attributed to the treatment applied to the experimental group, rather than to uncontrolled factors". This PhD Thesis has been carried out in two institutions, which kindly

agreed to take part in it, so the researcher has strived to have the minimum impact on the lessons. Nevertheless, the planning has been thorough to minimise unrelated variables and hence guarantee its internal validity as we can see below.

First, the instruments were piloted in a B2 group that could be described as equivalent because the students were attending lessons in one of the institutions taking place in the main study the in the second semester of 2017 and the results obtained were analysed statistically in order to ensure that they are adequate for the research.

Second, the timing of the data collection process was planned to try to make it as similar as possible in both institutions despite the different duration of the courses. Besides, it has coincided with the time when mock exams take place in the exam preparation programmes so neither the students nor the professionals have perceived they are doing something additional that may be a waste of time.

Third, there are groups which could be labelled as B2.1 and B2.2 depending on the number of instruction hours they have received before the study started. In addition, the programmes at CEB have a duration of around 120 hours whereas the programmes at CEALM have a duration of 60 hours. For this reason, all the CEALM groups have been considered to be Year 1 because in any case these students have had fewer than 120 teaching hours and the CEB groups have been categorised as either Year 1 if they are in their first year of preparation or Year 2 if they are in their second. Finally, the study has been conducted in the second half of the 120-teaching-hour course so that the time span between the application of first set of instruments and the second one in both groups is similar.

Fourth, there are several authors such as Robb and Ercanbrack (1999:8) and Allen (2016:2) who explain the improvement in exam performance on the basis of *practice effect,* that is, they claim that students perform better in the exam only because they are familiar with its format. In fact, this variable has been the reason for criticism of previous washback studies so, in order to eliminate it, all the students have received a handout with a summary of the exam format and tips and recommendations to succeed in the exam. In addition to that, they had practice with exam tasks.

Finally, the objective of this Thesis is to prove that students who prepare for B2 First improve their performance in the exam, make a progression in general language skills and become more independent learners. What the researcher understands by improvement in the exam can be objectively observed in the potential score gains. As for the general language skills, the tests that will assess this progression have been described in Section 4.2.2.2 of this study and can be found in Appendices 6, 7, 8 and 9. Regarding independent learning strategies, they have been described in the teachers and students' questionnaires, which are available in Appendices 1, 2, 3.

4.2.4.4. Generalizability

Madrid and Bueno (2005:660) define generalizability or external validity as "the extent to which the findings of a study can be generalised, or applied, to other (external) situations". The concepts of population and sample become relevant when talking about whether the results of a study can be generalised to other circumstances or not. In fact, Green (2006a:339) and Tsagari (2007:43) point out that some studies which included few classes or students had their generalizability compromised. This study includes data collected over four months from two different institutions in Jaén (Spain) and involved a total of 132 students and 8 teachers.

4.2.4.5. Data analysis

Given the sample available, qualitative and quantitative approaches along a continuum have been used. In the quantitative end, statistical techniques have been applied to check that the findings obtained can be generalized.

4.3. Study

The experimental part of this PhD Project was conducted between January 2018 and May 2018 in Jaén (Spain). In 2019, Jaén had 112,999 inhabitants (Instituto de Estadística y Cartografía de Andalucía, 2020). The average age was 42.6 (Instituto de Estadística y Cartografía de Andalucía, 2020) and the percentage of inhabitants younger than 20 was 20.5% (Instituto de Estadística y Cartografía de Andalucía, 2020) and that of people older than 65 was 17.9% (Instituto de Estadística y Cartografía de Andalucía, 2020). In terms of population growth, over the last decade its population decreased by 3.1% (Instituto de Estadística y Cartografía de Andalucía, 2020). Regarding

the economy of the province, agriculture is the main economic activity with 55.29% of new contracts, followed by the service sector with 33.40% and both the industrial sector and construction activity with the very similar percentages – 5.64% and 5.67% respectively – (Servicio Público de Empleo Estatal, 2019:64). Unemployment reached 19.12% in 2019 in the city (Instituto de Estadística y Cartografía de Andalucía, 2020).

4.3.1. Institutions

4.3.1.1. Centro de Estudios Avanzados en Lenguas Modernas (CEALM)

This language school, which is part of the University of Jaén, was officially founded in 2011 to offer language training and certification services to the staff and students of the University of. For this purpose, the institution follows the Communicative Approach, which is focused on adults since most of the students are university lecturers and students. As for the type of lessons, it offers test preparation and general language courses. In the first semester of 2018, when the study was conducted, 180 students were doing one of the 15 general English courses available – 3 of them were the B2 courses included in this study, there was another B2 course offered in Linares, a town near Jaén where the university has a campus – or the only course with a more exam focus offered (Universidad de Jaén, n.d.). Apart from being a language school, CEALM is also a venue for different English accreditation exams, including B2 First and Aptis as well as Certacles. From January 2018 to July 2018, 10 students sat B2 First exam, 382 sat Aptis and 20 went for Certacles (Universidad de Jaén, n.d.). The school employed a total of 6 English teachers who follow the institution guidelines in lessons. As for teacher training, they attend seminars and workshops organised by assessment institutions such as Cambridge Assessment English or TOEFL, courses in foreign universities such as Lancaster University and in-house seminars (Centro de Estudios Avanzados en Lenguas Modernas, n.d.).

4.3.1.2. Centro de Estudios Británicos (CEB)

The school was founded in 1990 in Jaén and delivers only English courses. Its rationale is based on the Communicative Approach and its main focus is to make learning enjoyable and to prepare students for Cambridge English Qualifications. In 2018 there were seven teachers delivering general English lessons as well as more exam-focused courses. Regarding teacher training, the school offers

an induction course for new teachers as well as monthly meetings and regular seminars for both new and more experienced teachers. Training is complemented with external workshops. All professionals are expected follow the school guidelines, which range from using English as the only medium of instruction from A2 level to setting homework on a daily basis. In terms of assessment practices, teachers must apply CEB criteria and students sit a speaking, a listening and a written test every term and regular Cambridge mock tests. Besides, students taking the B2 First exam can attend specific seminars to prepare each exam paper. Finally, students' profile is diverse because the youngest learners are 3 years old and there are also adults and teenagers. In the first semester of 2018, when the study was conducted, 643 students were enrolled on one of the courses taught by one of the 7 teachers who worked there at the time. In that academic year, 136 students sat an official Cambridge Assessment English qualification, that is, 21.15% of the total of students.

4.3.2. Students

The participants in this project studied at CEB and CEALM. They were divided into eight groups at CEB – three groups with Year 1 students and five groups with Year 2 students – and into three groups at CEALM – two B2.1 and one B2.2; however, for the purposes of this study all of them were categorised as Year 1 students because courses at CEALM run every semester and have a duration of 60 teaching hours.

In terms of population description and regarding age (Figure 6) and educational background (Figure 7), most students (58%) had not started their university studies as they were younger than 18. Those who had completed university studies, either an undergraduate or a postgraduate degree, reached only 19% of the total and this percentage went down to 14% when asked about participants who currently had a job. In the light of the above, our population are mainly full-time Secondary students. When asked about when they started learning English, only 7% started in Secondary education or later, which is in line with general trend in the Spanish Education System of having English as the first foreign language.

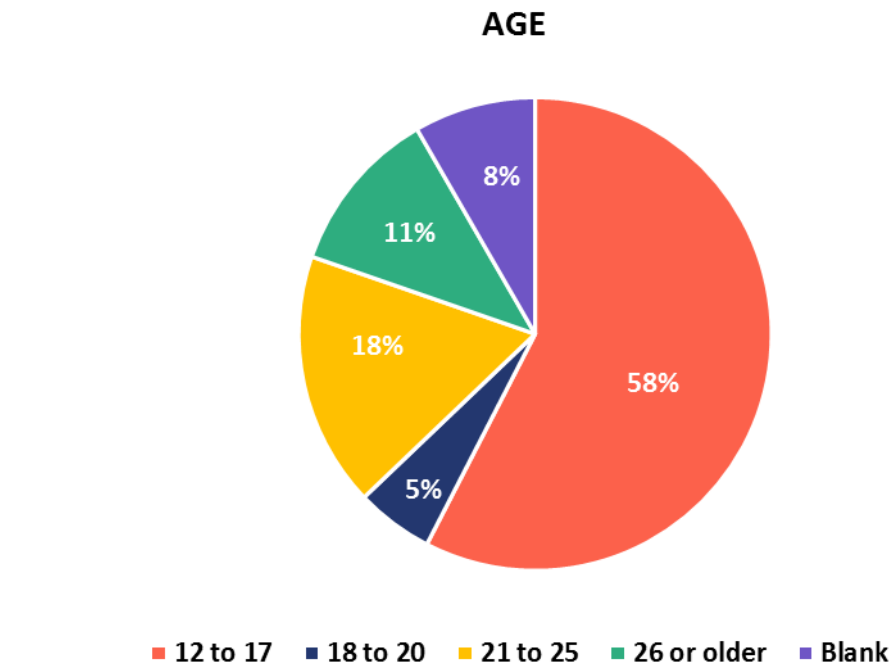


Figure 6. Students' profile: age

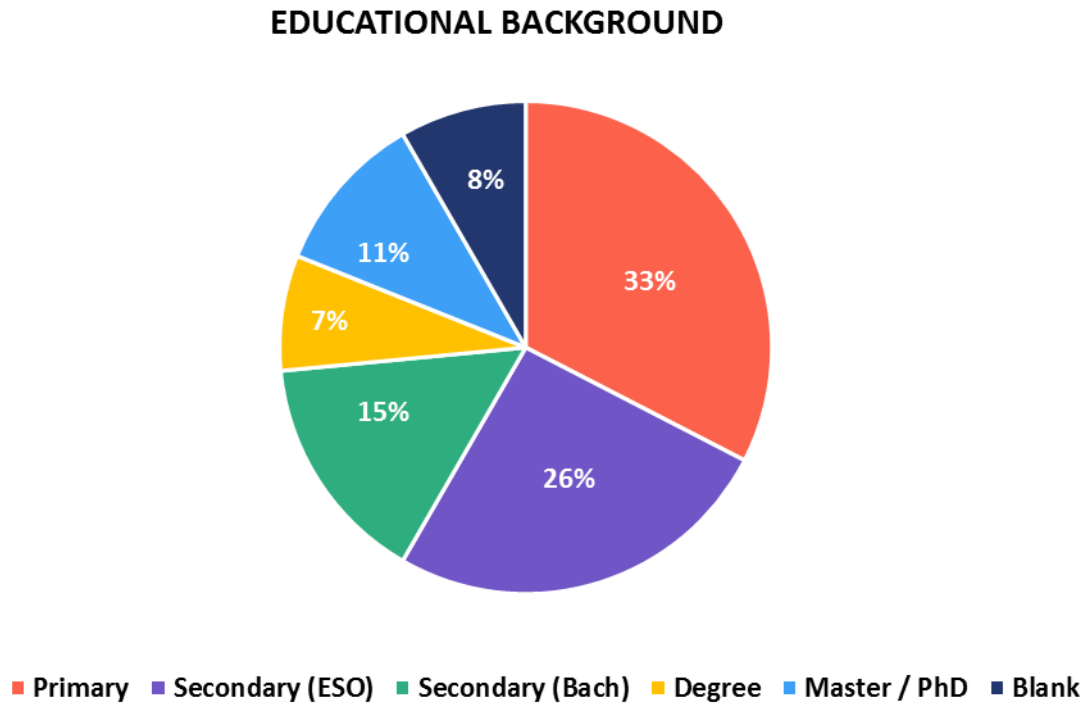


Figure 7. Students' profile: educational background

In terms of their experience with accreditation exams, the majority of participants (70%) had sat an official accreditation exam and 53% had done it in the previous two years, the most popular level being B1 (Figure 8).

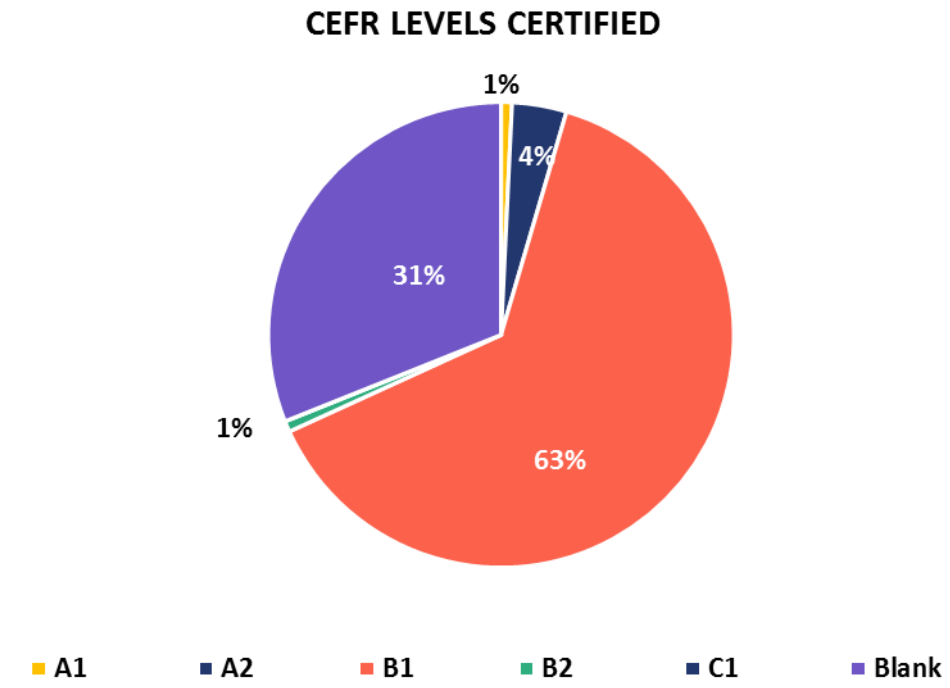


Figure 8. Students' profile: CEFR levels certified

When asked about their main reason for learning English (Figure 9), participants point at going abroad (23%), something that could be explained by the economic situation of Jaén, which was briefly described above. However, if education and employability are considered as one big group, 78% see English as a gateway to make progress either in Spain or abroad. To some extent, Wenyuan (2017:65) also noted that employment-related reasons were the most powerful factor when deciding to sit an official accreditation exam about Chinese candidates, followed by the need to obtain a certificate and to prove their level of English.

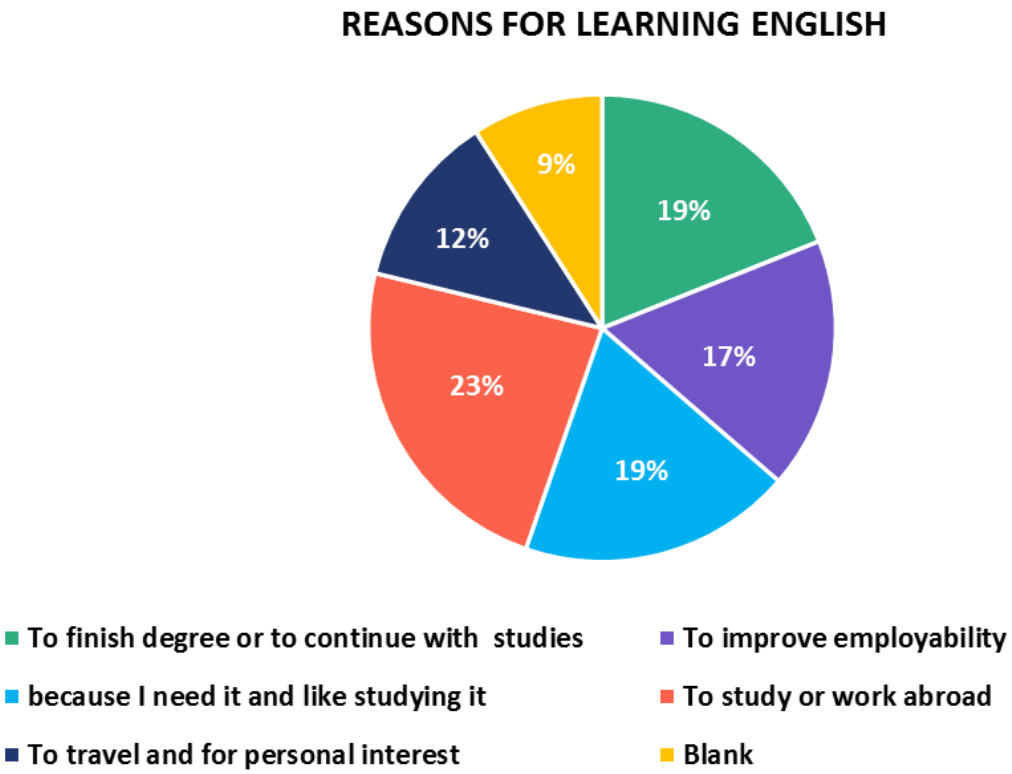


Figure 9. Students' profile: reasons for learning English

In fact, when asked about how important having a good command of English is 91% believe it is either important or very important and the percentage is very similar (89%) when asked about the course they are studying.

In terms of participants' perception of accreditation exams and their value, the majority (67%) aimed to pass a B2 exam – most of them B2 First with 66% of the answers – in the year when the study was conducted (2018). Interesting enough was the answer to how important passing an accreditation exam is because even more participants (84%) described it as important or very important. When choosing an official exam, prestige and recognition is the key factor with 55% of participants going for this option. Regarding the content and complexity of B2 accreditation exams, 62% of those who answered the questionnaire are quite confident they can pass the exam B2 First after 18 months preparing for it and 87% identify the four skills and interaction as the most essential aspects to pass a B2 exam.

4.3.3. Teachers

A total of eight teachers took part in the study. Six of them worked at CEB and two at CEALM. Of those working at CEB, four of them taught B2 groups. They were all qualified English teachers who had a degree in English or Translation and Interpreting and/or who had done postgraduate studies, e.g. Teaching English as a Foreign Language courses or a MA in Education or its previous equivalent in Spain – Certificado de Aptitud Pedagógica (CAP). All teachers except one had received Cambridge training. All of them had attended the seminars organised by Cambridge Assessment English in Jaén every year. In addition to those, CEB teachers received in-house training either at the beginning of or throughout the school year. As for the content of the training, it focused on marking and assessment (85% of the teachers selected this option), exam format and test-taking strategies (100%), exam practice activities (85%), students' motivation (28%) and resources available (71%).

Regarding their experience, all teachers were teaching different CEFR levels, mainly up to C1. When asked about the type of course they were delivering, up to 57% of the teachers delivered extensive test preparation courses – with a duration of at least 60 teaching hours, 14% of the teachers taught intensive general English courses – with a duration of fewer than 60 teaching hours, 28% of the teachers delivered intensive test preparation courses and 57% taught extensive general English courses.

In terms of the students' profile, all CEB teachers (71% of the total) taught teenagers and 60% of them also taught adults. All of CEALM professionals taught adult groups. It is also interesting to note that 86% of the teachers were preparing students for a B2 accreditation exam: 71% B2 First, 14% Aptis and 28% Trinity.

When asked about the difference between test preparation and general English courses, most teachers 57% agreed that test preparation courses give a greater focus to exam-based activities and to test-taking strategies, which is in line with the findings reported by Green (2007:76), as the teachers taking part in his study reported that, in this case, IELTS influenced their choice of activities. When Green (2007:86) asked teachers about what activities they thought could boost test scores, teachers agreed that the activities that are based on test content such as memorisation of phrases, question analysis and direct test practice were of value. However, they could not agree about how activities which are less closely related to test content impacted on score gain. Back to the present study, two teachers (29%) mentioned the narrowing of the

curriculum – “The focus is on doing the exam, on showing what the examiners want to see rather than on learning English” (Teacher Questionnaire) – and the type of activities – “less fun” (Teacher Questionnair)e – as negative aspects of test preparation courses. Regarding the time students need to move from B1 to B2 57% of the teachers agreed that learners may need between 120 and 180 teaching hours and 29% thought that they need at least 180 teaching hours.

4.3.4. Data collection

In this section, the data collection process and calendar (Table 1) will be described. It must be noted that the administration of the instruments was done simultaneously in both institutions.

Table 1. Data collection calendar

Instrument	Date
Entry Questionnaire	January 2018 - February 2018
End-of-Course Questionnaire	April 2018 - May 2018
Teacher Questionnaire	May 2018
Entry Cambridge Test	January 2018 - February 2018
End-of-Course Cambridge Test	April 2018 - May 2018
Entry Grammar Test	January 2018
End-of-Course Grammat Test	April 2018
Entry Vocabulary Test	January 2018
End-of-Course Vocabulary Test	April 2018
Classroom observation	February 2018 - April 2018

4.3.4.1. Questionnaires

Questionnaires were handed out on paper and were applied during class time. Students filled in an entry questionnaire in January and February 2018 and an End-of-Course questionnaire in April and May 2018. Although at first the initial plan was that the end-of-course questionnaires were filled in at the end of the course, i.e. May / June 2018, the fact that students started dropping out in early May led the author to prioritise having as many participants as possible. Teachers only filled in a questionnaire and this was done in May 2018.

4.3.4.2. Exams

4.3.4.2.1. Cambridge Tests

The Entry B2 First mock exam, referred to as Entry Cambridge Tests (henceforth ECT) in this Thesis, was administered in January / February 2018 during lesson time on different days and weeks in order to cause as little disruption as possible in the lessons. The end-of-course B2 First mock exam, referred to as End-of-Course Cambridge Test (henceforth EoCCT) in this Thesis, was administered in the same conditions in April / May 2018.

4.3.4.2.2. Grammar Test

The Entry Grammar Test was administered in January 2018 and the End-of-Course Grammar Test was applied in April 2018. Both of them were applied in a printed format and during lesson time.

4.3.4.2.3. Vocabulary Test

As with the Entry Grammar Test, the Entry Vocabulary Test (EVT) was applied in January 2018 and the End-of-Course Vocabulary Test (EoCVT) was administered in April 2018. Both of them were applied in a printed format and during lesson time.

Test results were processed by the author with the support of Ms. Valentina Cueva López, who works as a teacher of Statistics at the University of Jaén and who collaborates with the department of English Studies of the University of Jaén carrying out statistical analyses.

4.3.4.3. Observation

Including observation has been recommended as a means of contextualising, corroborating, or correcting data from surveys (Alderson & Wall, 1993; Wall, 1996; Turner, 2001; Watanabe 2004; all cited by Green, 2006:334). Lessons from a sample of groups and levels – Year 1 and Year 2 – were observed in both institutions from February to April 2018 in two non-consecutive weeks. On the one hand, the author attended the lessons delivered by one of the two CEALM teachers, including a B2.1 group and a B2.2 group – although both categorised as Year 1 – so eight hours were observed in week 1 and eight hours in week 2. In this case, the observation sheet was filled in during the class. On the other hand, the author recorded the lessons taking place at CEB and delivered by four different teachers so 12 hours were observed in week 1 and 12 hours in week 2. In this case, the observation sheet was filled in after the lesson. For that purpose, the lessons were recorded with students and parents' consent using the author's iPhone 4s recorder.

4.4. Variables

4.4.1. Time 1

One of the variables that are considered in this study is the teaching hours received by the students. Students who have received fewer than 120 teaching hours are categorised as *Year 1* and those who have attended more than 120 hours of lessons are categorised as *Year 2*. As mentioned above, all CEALM students are categorised as Year 1 because courses at CEALM have a duration of 60 teaching hours. CEB students are categorised as Year 1 or Year 2 depending on whether they are in their first or second year of preparation because CEB courses have a duration of 120 hours approximately. This variable is relevant because students who have attended lessons for longer may be more familiar with the contents and skills expected at B2 than those who have just started studying English at B2.

4.4.2. Time 2

There is a second variable related to time, in this case it refers to when the data collection took place i.e. at the beginning of the project or at the end. Including this variable is vital as the study is based on the analysis of potential progression and learning. Also, according to Prodromou (1995;

cited by Xie & Andrews, 2012:55) washback behaves as a continuum from covert to overt influence – considered to be the most intense form of washback accessible to observation and measurement – as the date of the test approaches, particularly when the test taking dates are externally determined and the stakes are high. This description applied to the situation of the participants in the experimental group so, from a methodological point of view, it was considered appropriate to apply the instruments at the beginning of the project when the exam dates are still considered to be distant and at the end of the project when the exam is just a few weeks away. The tests and questionnaires administered at the beginning of the study are categorised as *Entry* and the tests and questionnaires administered at the end of the project are categorised as *End-of-Course*.

4.4.3. Course type

Another key variable in this study is the type of course that students are attending. On the one hand, students may be enrolled on a course focused on teaching General English. The course book used is *English File Upper Intermediate* (Oxford University Press, 2014) and is not designed to prepare students for any exam in particular. In terms of assessment for the course, it does not follow the style of any proficiency exam specifically and it is designed by the teachers. Finally, students attending this course may or may not aim to sit a proficiency exam at the end of it. All the general English courses are organised by CEALM. On the other hand, students may be enrolled on a course which is more exam oriented. The course books used are *Complete First* (Brook-Hart & Owen, 2011) for Year 1 students and *Objective First* (Capel & Sharp, 2014) for Year 2 and both are designed to develop students' skills at B2 level using similar tasks to the ones that can be found in B2 First exam. In terms of assessment, students sit an exam which does not follow the style of any proficiency exam in particular and which is designed by the teachers and they also sit a B2 First mock exam. All exam oriented courses are organised by CEB and students usually aim to sit the B2 exam when they are ready for it.

4.4.4. Test type

To measure students' performance two different types of test have been used. On the one hand, two sets of a grammar test and a vocabulary test were administered at the beginning and at the end of the project and are categorised as the *independent* tests because their design is not based on any proficiency test and students should be equally familiar with their format. On the other

hand, two sets of B2 First mock exams were administered at the beginning and at the end of the project. The objective of introducing this variable is to compare score gain in both types of tests to have a better understanding of it and how students make progress.

4.4.5. Components

Apart from studying learners' performance in the test as a whole, performance in the different B2 First test components is analysed. The components are: Speaking, Writing, Listening, Reading and Use of English. It must be noted that although Reading and Use of English are one paper in the real B2 First exam, this project administered the two components and analysed them separately for practical reasons. The objective of introducing this variable was to have an additional tool to compare performance and obtain richer results.

4.4.6. Learners' autonomy and independence

Lifelong learning has become a reality in all aspects of life but is paramount when it comes to communicating in a foreign language. Developing learners' autonomy so that they are aware of their strengths and weaknesses and are able to identify the main difficulties they have is vital to continue learning. Being familiar with assessment criteria and developing a critical approach to one's performance should be part of any language course and this is why introducing this variable was considered useful.

5. RESULTS: PRESENTATION AND DISCUSSION

This section aims to describe the results obtained from the statistical analyses carried out and to analyse and discuss them in the light of the qualitative and quantitative data obtained from questionnaires, the observation schedule, and relevant studies on the matter. A t-test for independent samples was used to compare the experimental and the control group and also Year 1 and Year 2 students. A t-test for dependent samples was used to analyse the progression of students from the beginning to the end of the course. The programme SPSS version 24 was used. The significance value is 5%.

First, the performance shown by CEB and CEALM students in B2 First mock exam will be described, analysed and discussed separately to have a better understanding of each group's progression and abilities. After that, attention will be paid to the skills profile of students in each institution to understand the areas where students show a better performance, to identify the aspects on which students need to work harder and to spot potential differences derived from the course aims. Later, the comparison between the experimental group – CEB Year 1 students – and the control group – CEALM students – will be analysed to look into the potential washback of Cambridge B2 First exam on those students enrolled on courses with a greater exam focus. The discussion of these data will try to answer the first research question: i) Do students enrolled on more exam-oriented (CEB) courses show a better performance in B2 First mock exam than those enrolled on general English courses (CEALM)?

Then, the results of CEB and CEALM students in the independent tests will be discussed and compared with those obtained on the Use of English component. Although grammar and vocabulary are assessed in the different components of Cambridge B2 First exam, they are more explicitly assessed in the Use of English component. For this reason, it was considered useful to compare the results obtained on the independent tests and the Use of English test to have a better understanding of how exam practice may influence the results obtained. Finally, the results obtained by the experimental and the control group will be described and analysed to understand how exam preparation may impact the learners' grammatical and lexical ability. The analysis of these data will help answer the second research question: ii) Do students enrolled on more exam-oriented (CEB) courses improve their language knowledge and abilities?

Finally, the third research question will be considered: iii) Do students become more autonomous and independent learners as a result of preparing for Cambridge B2 First exam? In order to answer this question, students' answers to the two sets of questionnaires applied will be discussed paying special attention to students' ability to identify ways to improve their level of English and their strengths and weaknesses, to the activities done in class and to prepare for the B2 exam, and to the washback of B2 First on motivation.

The data will be contextualised with information obtained from the students and teachers' questionnaires and with findings from relevant studies to identify potential similarities or differences.

5.1. Do students enrolled on more exam-oriented (CEB) courses show a better performance in Cambridge B2 First mock exam than those enrolled on general English courses (CEALM)?

5.1.1. Students enrolled on more exam-oriented (CEB) courses

In order to have a better understanding of how the B2 First exam affects students' learning, an Entry B2 First mock exam (Entry Cambridge Test, ECT) and an End-of-course B2 First mock exam (End-of-Course Cambridge Test, EoCCT) were applied. Table 2 presents the results obtained. Interestingly enough, CEB students as a whole improve their performance from the ECT (54.00%) to the EoCCT (56.70%) (p value 0.007).

5.1.1.1. Speaking

If performance in the Entry Speaking Test (EST), which is part of Entry B2 First mock exam, is compared against performance in the End-of-Course Speaking Test (EoCST), which is part of the End-of-course B2 First mock exam, the results are as follows: no statistically significant difference (p value 0.278) can be found between the performance of CEB students in the EST and the EoCST. Nevertheless, the mean shows that CEB students pass[3] that part of the exam in both cases as their performance is 64.49% and 63.25% respectively.

[3] Cambridge Assessment English uses the Cambridge English Scale to express language ability. Candidates who achieve an overall score of 160 points of the Cambridge English Scale can certify a B2 level. The score is calculated using statistical analyses carried out before the test and after the test. For the purposes of this study, percentages have been

5.1.1.2. Writing

When looking at the results for the writing component, CEB students perform in a similar way in the Entry Writing Test (EWT) (56.42%) and in the End-of-Course Writing Test (EoCWT) (57.65%) as the difference between the results was not statistically significant (p value 0.356). Another aspect to note is that students did not reach the pass mark in this component.

5.1.1.3. Listening

In terms of CEB students' listening ability, they did not reach the pass mark either in the Entry Listening Test (ELT) or in the End-of-Course Listening Test (EoCLT). In addition, the difference between the ELT (56.89%) and the EoCLT (54.18%) is not statistically significant (p value 0.118).

5.1.1.4. Reading

Reading is the second weakest skill for CEB students after Use of English – see further detail in Table 2. CEB students did not reach the pass mark either in the Entry Reading Test (ERT) (49.88%) or in the End-of-Course Reading Test (EoCRT) (52.84%), and the difference between both tests is not statistically significant (p value 0.098).

5.1.1.5. Use of English

The ability to use grammatical and lexical resources accurately is assessed in different parts of the B2 First exam but it receives a special focus in the Use of English component. Reading and Use of English are one paper in the B2 First exam although the score obtained on the Reading part is independent from the score achieved on the Use of English part of the paper. This research study follows the same procedure and analyses the performance in the Use of English component separately and independently from that in Reading. In this project, the rationale for this separation is to be able to establish a comparison between the results in the Use of English part and the independent tests.

CEB students' grammatical and lexical ability improves significantly from the Entry Use of English Test (EUoET) (44.42%) to the End-of-Course Use of English Test (EoCUoET) (51.29%) (p value

used to express the pass mark in the B2 First exam following the scale score converter available on University of Cambridge Local Examinations Syndicate (2020).

<0.001) (Table 2) although it is still the lowest mean score when compared with all four skills. As we will see later on, Use of English is the only component in which students improved although they do not reach the pass mark in it. In fact, it shows the weakest performance both at the beginning and the end of this project.

Table 2. Results obtained by CEB students on B2 First mock exam.

B2 First mock exam	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
ECT[6]	70	54.00%	12.71	-0.224	0.007	*
EoCCT[7]	70	**56.70%**	11.32			
EST[8]	81	64.49%	9.42	0.111	0.278	
EoCST[9]	81	63.25%	12.65			
EWT[10]	81	56.42%	11.76	-0.093	0.356	
EoCWT[11]	81	57.65%	14.69			
ELT[12]	75	56.89%	16.00	0.168	0.118	
EoCLT[13]	75	54.18%	16.28			
ERT[14]	77	49.88%	14.98	-0.194	0.098	
EoCRT[15]	77	52.84%	15.54			
EUoET[16]	80	44.42%	17.42	0.367	<0.001	*
EoCUoET[17]	80	**51.29%**	19.93			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Cambridge Test; [7] End-of-Course Cambridge Test; [8] Entry Speaking Test; [9] End-of-Couse Speaking Test; [10] Entry Writing Test; [11] End-of-Course Writing Test; [12] Entry Listening Test; [13] End-of-Course Listening Test; [14] Entry Reading Test; [15] End-of-Course Reading Test; [16] Entry Use of English Test; [17] End-of-Course Use of English Test

As we have just seen, the results obtained by CEB students cannot be summarised easily, they require careful analysis as no clear pattern can be observed. The improvement in performance from the ECT to the EoCCT is in line with the fact that most students at CEB (71%) aimed to sit an accreditation exam in the academic year 2017-2018. In addition, to some extent, this finding agrees with those obtained by Xie (2010; cited by Cheng, Sun & Ma, 2015:459), who compared pre- and post-test performance and found that improvement was substantial. The results also agree with the

findings reported by Saif (2006:25), who looked into the washback of a needs-based test of spoken language proficiency which was developed for international teaching assistants. In order to do so, Saif compared two groups of participants: on the one hand, the control group, who attended an old orientation programme for international teaching assistants; and on the other hand, the experimental group, who took a course led by an English as a Second Language teacher.

Nevertheless, these findings contrast with the results reported by Read and Hayes (2003; cited by Ha, 2019:9), who compared score gains on the IELTS listening, reading and writing tests of international students who attended two IELTS courses in New Zealand. The first type of course was more exam-oriented while the second type of course focused not only on test tasks, but also on the development of language knowledge and academic skills. Two versions of IELTS were administered and no statistically significant differences in the overall entry and end-of-course tests were found.

If the performance in each test is analysed independently, the mean scores for CEB students, which includes Year 1 and Year 2, are below the pass mark both at the beginning and at the end of the project. If we look at performance in more detail and analyse results in the different skills, students seem to have reached a plateau in all four skills and the mean scores at entry and end-of-course tests suggest that, in general, students need to work harder to develop the abilities expected of B2 learners. This finding, more precisely the similarity in test scores between the EWT and the EoCWT, is in line with Green's (2007b; cited by Cheng, Sun & Ma, 2015:458) conclusion that test preparation does not contribute much to improving performance. Green's study (2007b; cited by Ha, 2019:9) included three groups of participants. The first group was enrolled on IELTS preparation courses, the second group attended English for Academic Purposes (henceforth, EAP) courses, which had no IELTS component, and the third group attended a mixed course with EAP and IELTS preparation. All participants completed IELTS writing tests at the beginning and at the end of their courses. The results showed that the groups with IELTS preparation did not improve their performance in the test as compared to the EAP group. However, the similarity between EWT and EoCWT results contrasts with Estaji (2013:222)'s research, which studied the washback of IELTS Writing and Academic Writing courses on learners in Iran, and reported score gains in IELTS Writing tests although they were not substantial.

More optimistic trends can be perceived when looking at performance in the Speaking component because CEB students pass the Speaking Paper at the beginning and at the end of the project. This stronger performance, which is compatible with B2 abilities, could be explained by the fact that classroom observation showed that it is the skill that received greater attention in class – 51% of total class time, see Figure 10 – and that the questionnaires indicated that 85% of students – CEB and CEALM – were in favour of practising mainly oral skills, i.e. speaking and listening, in class (Figure 11). The other example of positive trends can be found in the Use of English component, which is the only part where CEB students show a statistically significant improvement. In this sense, we can find a similarity with the study by Read and Hayes (2003, in Ha 2019:9) because they also reported that the students attending a more-exam oriented focus improved in one skill, in this case Listening. Read and Hayes linked the improvement with the large amount of time devoted to this skill in class. A similar finding was obtained by Allen in 2016 (cited by Ha, 2019:10), who also investigated the washback effect of the IELTS test on score gain. The difference with the present study is that participants did not take a test preparation course but prepared for the test independently and sat the exam twice within a one-year period. The results show an improvement in the speaking scores for all the participants.

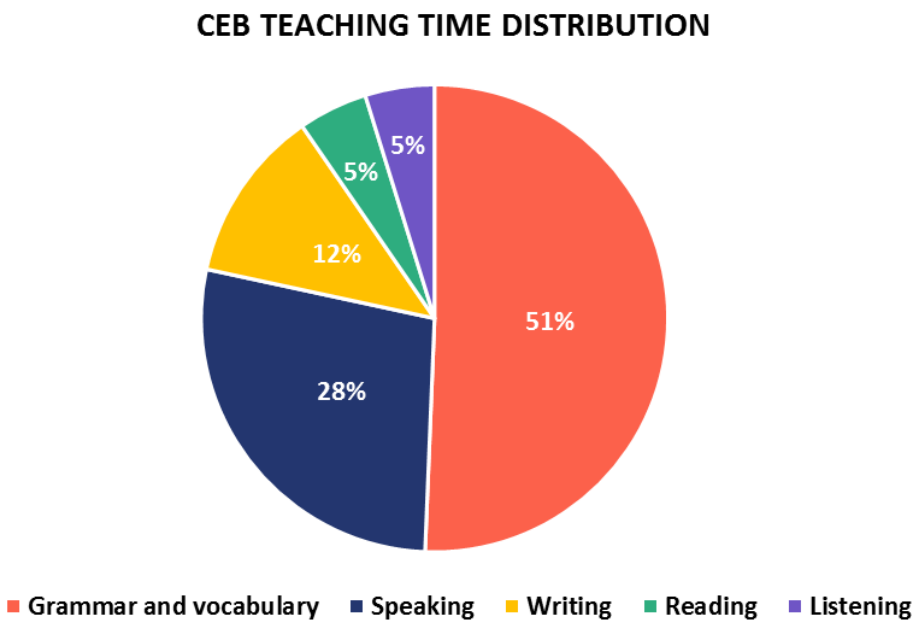


Figure 10. CEB teaching time distribution

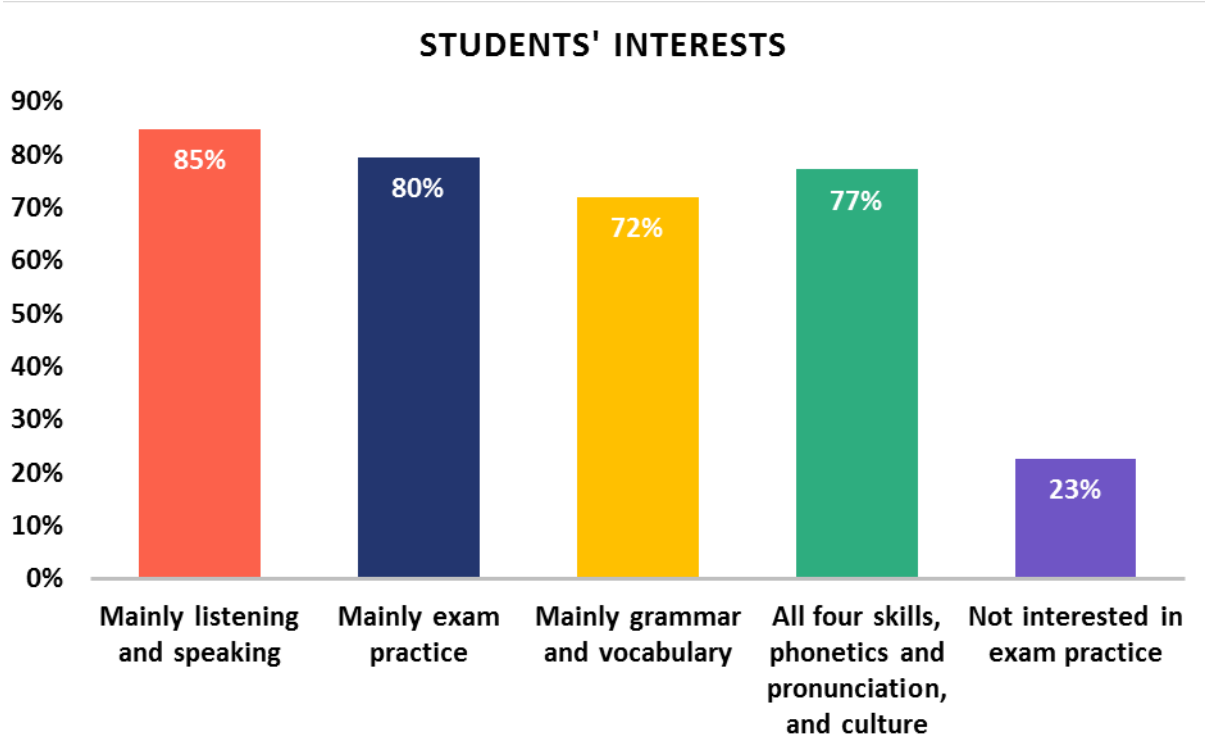


Figure 11. Students' interests in an English course

Low performance in Reading and less tangible score gain when compared to other skills is also reported by Elder and O'Loughlin (2002; cited by Rao et al., 2003:251), who compared performance in reading and writing in the IELTS exam. Interestingly, one of the possible causes for this difference mentioned by the authors was whether reading was more difficult to influence than other skills.

In summary, performance for learners studying at CEB shows that although there is an improvement from the beginning to the end of the project if we look at the overall score; this improvement is not statistically significant when looking at specific components. Although this finding is in line with other studies already discussed such as Green's (2007b; cited by Cheng, Sun & Ma, 2015:458), in the present study, it could suggest that the time between both sets of tests was too little. It will be interesting to see how Year 1 and Year 2 students perform when studied as two independent groups. As already mentioned, the findings obtained agree with some of the literature reviewed and where contrasts are found, the information obtained from the observation schedule and the questionnaires helps to understand the reason for these differences.

5.1.2. Comparison between Year 1 and Year 2 students enrolled on more exam-oriented (CEB) courses

Before describing the results and starting the discussion about the comparison between Year 1 and Year 2 students, it must be noted that this is a comparison within the group of CEB students, who attend more exam-oriented courses. This is because CEALM students have been categorised as Year 1 students while CEB has Year 1 and Year 2 students. So, in order to reduce the presence of construct-irrelevant factors as much as possible it was considered more appropriate to include students from one institution, in this case CEB.

Table 3 and Figure 12 provide the results obtained in this part of the present study. First, when comparing the mean scores obtained on B2 First entry and end-of-course exams by Year 1 and Year 2 students, statistically significant differences appear both in the ECT, with Year 1 students obtaining 46.66% as compared to Year 2 students' 58.76% ($p<0.001$), and in the EoCT, with Year 1 reaching 49.34% and Year 2 reaching the pass mark (61.40%).

Table 3. Year 1 and Year 2 students: Cambridge B2 First mock exam

B2 First mock exam	Year	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
ECT[6]	1	32	46.66%	12.29	-1.109	<0.001	*
	2	50	**58.76%**	9.94			
EoCCT[7]	1	29	49.34%	10.17	-1.262	<0.001	*
	2	43	**61.40%**	9.13			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Cambridge Test; [7] End-of-Course Cambridge Test

5.1.2.1. Speaking

When checking if the performance of Year 1 and Year 2 differs, results show that the performance of Year 1 students is significantly weaker in EST (60.78% vs. 67.68%) (p value 0.001) and in the EoCST (58.89% vs. 66.25%) (p value 0.009).

5.1.2.2. Writing

If we analyse how Year 1 and Year 2 learners did in the Writing Paper, a noticeable difference can be observed both in the EWT (50% vs. 60.39%) (p value <0.001) and in the EoCWT (49.38% vs. 62.80%) (p value <0.001) with Year 2 students being significantly stronger at writing than Year 1 learners.

5.1.2.3. Listening

If the listening ability of Year 1 and Year 2 students is compared, the difference between the former and the latter is statistically significant both in the ELT (50.10% vs. 61.76%) (p value 0.001) and in the EoCLT (48.33% vs. 58.07%) (p value 0.001) with Year 2 students showing a stronger performance.

5.1.2.4. Reading

Statistically significant differences appear when comparing Year 1 and Year 2 results: in the ERT, Year 1 students score 42.71%, while their Year 2 peers reach 53.18% (p value 0.001); in the EoCRT, Year 1 learners achieve 46.39% and Year 2 obtain a mean of 56.59% (p value 0.004).

5.1.2.5. Use of English

Looking at potential differences between Year 1 and Year 2 learners, statistically significant differences can be observed both in the EUoET (36.25% vs. 49.79%) (p value <0.001) and in the EoUoET (41.74% vs. 57.50%) (p value <0.001).

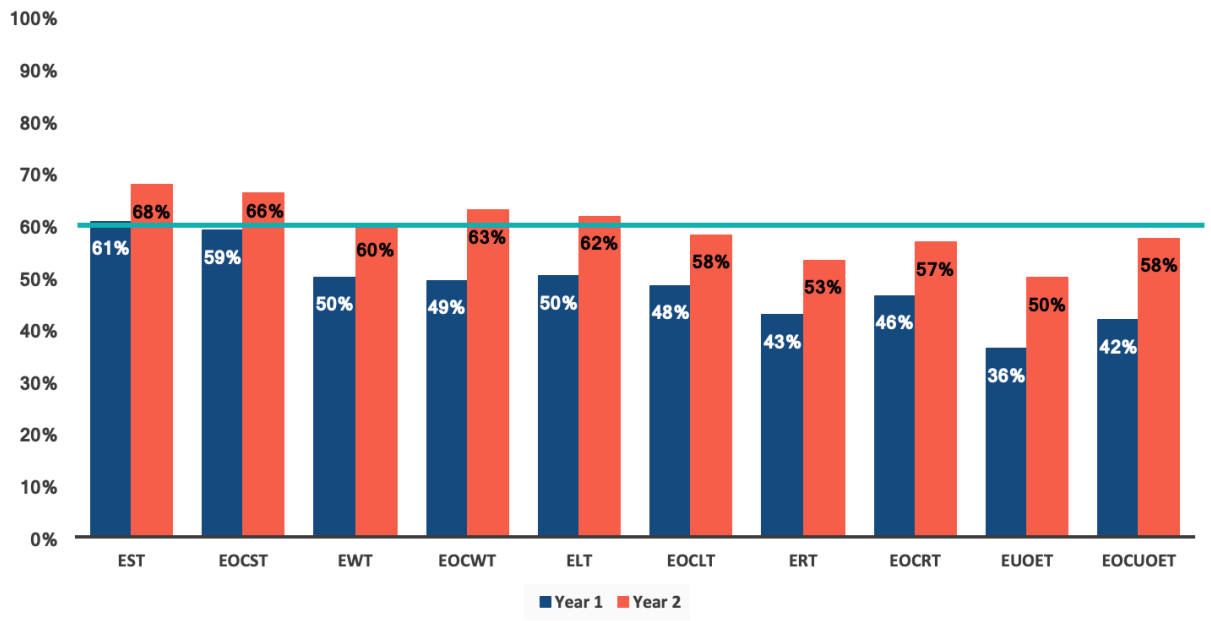


Figure 12. B2 First mock exam: skills profile of Year 1 and Year 2 students

EST - Entry Speaking Test; EoCST - End-of-Course Speaking Test; EWT - Entry Writing Test; EoCWT - End-of-Course Writing Test; ELT - Entry Listening Test; EoCLT - End-of-Course Listening Test; ERT - Entry Reading Test; EoCRT - End-of-Course Reading Test; EUoET - Entry Use of English Test; EoCUoET - End-of-Course Use of English Test

The results obtained from the comparison of Year 1 and Year 2 students show very clear patterns because Year 2 students outperform their Year 1 peers. The mean of Year 2 students is higher than the mean of Year 1 students in the ECT and the EoCCT, and also in the different papers. In addition, Year 2 students are above the pass mark in productive skills and close to it in the rest. If we look at the skills profile, the same trend can be found in Year 1 and Year 2 students, with speaking as the strongest component, followed by writing, then listening and with reading and the component of use of English as the weakest.

It is interesting to analyse the activities or the contents that receive more attention in Year 1 and Year 2 groups to try to contextualise the results and understand why the weaknesses in Year 1 students still affect Year 2. The observation schedule shows that the teaching time distribution follows a similar pattern in Year 1 (Figure 13) and Year 2 (Figure 14). Grammar and vocabulary are the contents that receive the greatest attention, although it must be noted that the time used to practise them is considerably reduced in the second year - from 70% to 44%, followed by speaking, which increases slightly despite showing the strongest performance, and writing in the third place.

Probably the most surprising aspect is that in spite of the importance given to lexical and grammatical contents in the two years, the results in Use of English are the weakest. It is true, however, that it is the only part of the exam where CEB students improve from the beginning to the end of the project as we saw before.

If we zoom in and consider only Year 1 students (Figure 13), their skills profile could be in line, to some extent, with the records of the classroom observation, which show that speaking is the second most frequently practised linguistic aspect, followed by listening and writing with very similar percentages. The fact that reading was not developed in the time the lessons were observed could explain the weaker results in this skill. The oddest aspect is the weak performance in the Use of English component, which contrasts with the time devoted to grammar and vocabulary. It could be argued that the development of the lexico-grammatical ability takes longer to consolidate and could build up and bear fruit in Year 2.

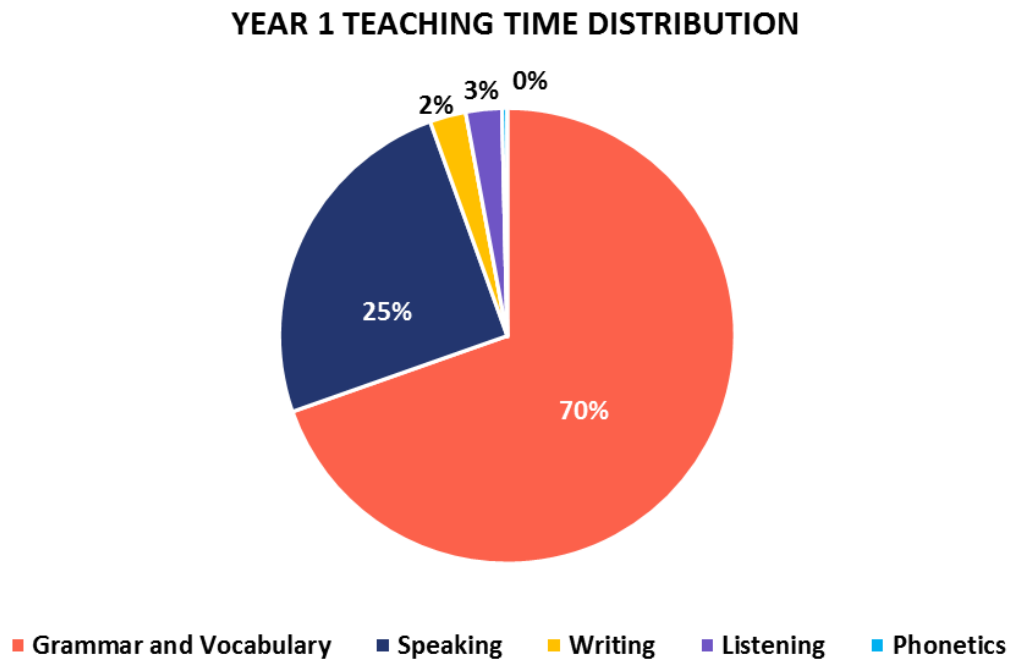


Figure 13. CEB students: Year 1 teaching time distribution

When looking at the observations of Year 2 lessons (Figure 14), the fact that the time distribution is more homogenous than in Year 1 lessons is very positive. Grammar and vocabulary still receive the greatest deal of attention, which could explain the clear improvement experienced by Year 2 students when compared with Year 1 results.

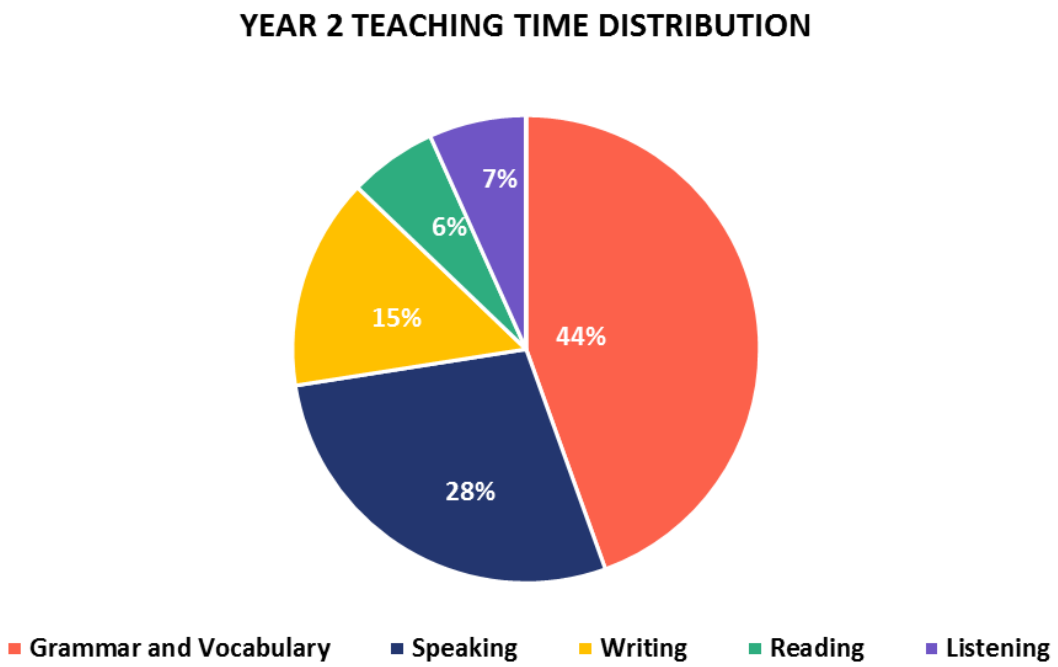


Figure 14. CEB students: Year 2 teaching time distribution

The weak results that CEB students in general show in the Reading and the Use of English components and the similar behaviour displayed in the EoCRT and the EoCUoET, especially by Year 2 students, deserve closer attention. When CEB students were asked about the main obstacle they face when sitting the Reading Paper (Figure 15), they mentioned the complexity and range of vocabulary, which is explicitly tested in the Use of English test. In fact, the student questionnaires showed that lexis was also identified by learners as a serious difficulty in the Use of English test. The second main problem has to do with reading strategies and exam format. Finally, the percentage of learners who did not answer the question or denied experiencing any difficulty in this component is also worth noticing and could be due to lack of awareness of their own strengths and weaknesses or lack of engagement with the questionnaire. The latter is a limitation found in other studies such as Muñoz, Véliz-Campos and Véliz (2019:115)

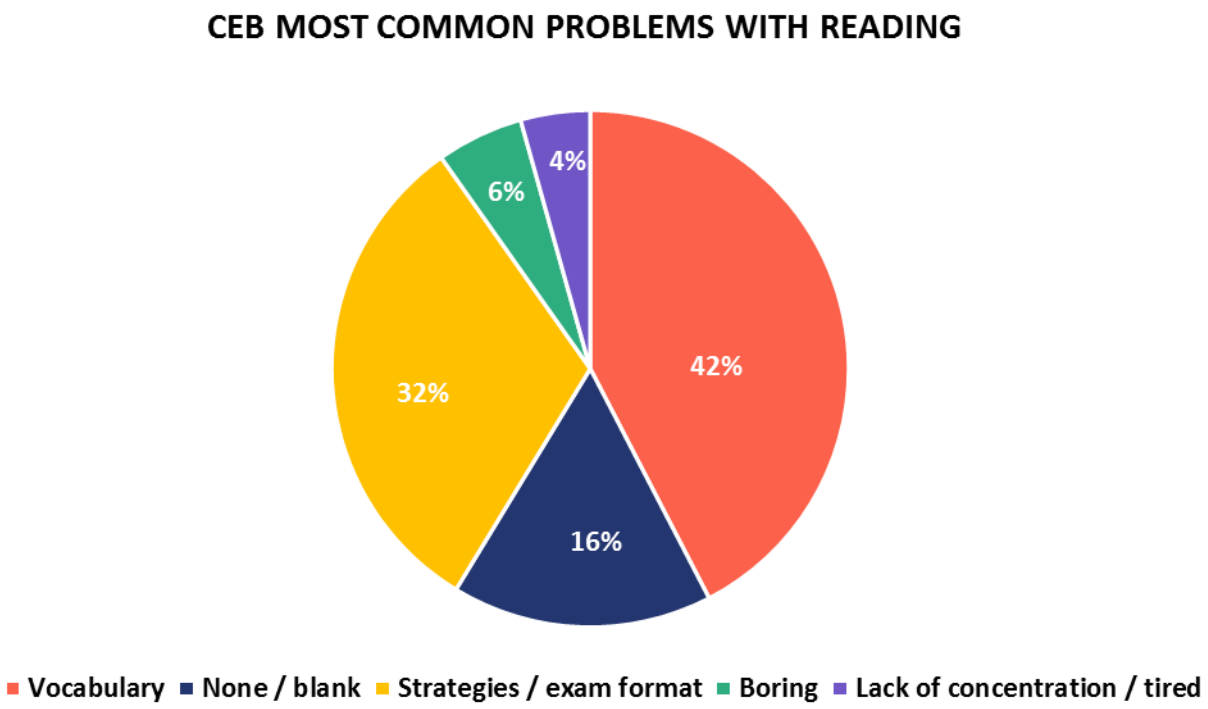


Figure 15. CEB students: most common problems with reading

To sum up, the results seem to confirm that students need more than 120 hours to move from one CEFR level to another, which is in line with the views expressed by the majority of the teachers that took part in this study and with the literature on the topic (Graddol, 2006:96 and Green, 2007a:86). Allen (2016 citing Read & Hayes, 2003, and Green, 2005) adds that, apart from a considerable amount of time, intensive preparation is needed and that these two factors become increasingly important at higher levels. The skills profiles are similar for Year 1 and Year 2 learners and are in line, to some extent, with the teaching time distribution in the institution. Finally, it is interesting to see the relationship between Reading and Use of English not only in terms of mean scores – showing the weakest performance – but also in terms of difficulties experienced by the students.

5.1.3. Students enrolled on general English (CEALM) courses

The results obtained in this part of the present study are presented in Table 4 below. When comparing the overall mark of B2 First mock exams at the beginning and at the end of the project, the difference in performance between ECT (39.72%) and EoCCT (43.72%) is not statistically significant (p value 0.056). As in the case of CEB students, apart from the interest in the overall performance at the beginning and at the end of the project, having a deeper knowledge of how students performed in the different skills was considered useful. In the next sections, the results per skill are described.

5.1.3.1. Speaking

CEALM students improve from the EST (45.98%) to the EoCST (52.35%) (p value < 0.001) although they did not reach the pass mark for this part of the exam[4].

5.1.3.2. Writing

In this paper, an improvement from the EWT (34.76%) to EoCWT (46.96%) can be observed (p value 0.001). However, the students did not reach the pass mark.

5.1.3.3. Listening

The difference between the ELT and the EoCLT is not statistically significant (p value 0.653) and the results do not reach what is expected of B2 candidates, with 33.68% in the ELT and 34.91% in the EoCLT.

5.1.3.4. Reading

CEALM students do not reach the pass mark either in the ERT (40.35%) or in the EoCRT (40.48%) and the difference between both is not statistically significant (p value 0.967).

4 Cambridge Assessment English follows a compensatory approach by which candidates do not need to reach the pass mark in every paper to be awarded a certain CEFR level.

5.1.3.5. Use of English

When analysing CEALM results, the difference between EUoET (35.53%) and EoCUoET (37.78%) is not statistically significant (p value 0.533) and the scores show that together with Listening it is where students need to work harder.

Table 4. Results obtained by CEALM students on B2 First mock exam

B2 First mock exam	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
ECT[6]	11	39.72%	12.1086	-0.357	0.056	
EoCCT[7]	11	43.72%	10.23808			
EST[8]	22	45.98%	11.49	-0.594	<0.001	*
EoCST[9]	22	**52.35%**	9.86			
EWT[10]	23	34.78%	15.63	-0.878	<0.001	*
EoCWT[11]	23	**46.96%**	11.85			
ELT[12]	19	33.68%	14.31	-0.086	0.653	
EoCLT[13]	19	34.91%	14.25			
ERT[14]	19	40.35%	23.29	-0.006	0.967	
EoCERT[15]	19	40.48%	17.46			
EUoET[16]	19	35.53%	14.21	-0.161	0.533	
EoCUoET[17]	19	37.78%	13.85			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Cambridge Test; [7] End-of-Course Cambridge Test; [8] Entry Speaking Test; [9] End-of-Couse Speaking Test; [10] Entry Writing Test; [11] End-of-Course Writing Test; [12] Entry Listening Test; [13] End-of-Course Listening Test; [14] Entry Reading Test; [15] End-of-Course Reading Test; [16] Entry Use of English Test; [17] End-of-Course Use of English Test

If the overall score in the B2 First mock exams are considered, CEALM students' performance is the same in statistical terms and the mean scores show that CEALM learners are not ready for the B2 level although 64% aimed to pass a B2 accreditation exam in the academic year

2017-2018. However, if the skills are analysed independently, learners improve in speaking and writing although the mean scores do not reach the pass mark for those components. The improvement in speaking may be due to the fact that it is the second most often practised linguistic aspect in class, like at CEB, with 35% of the teaching time (Figure 16) and to the students' interest in practising oral skills as 85% were in favour of practising mainly speaking and listening in class (Figure 11 above) – although listening performance does not improve in the light of the results obtained on the ELT and EoCLT. However, classroom observation does not show a special focus on writing, which is set for homework and was not practised in class during the observation period. The mean scores in the rest of the skills and the Use of English component show that students have reached a plateau because no statistically significant differences are observed.

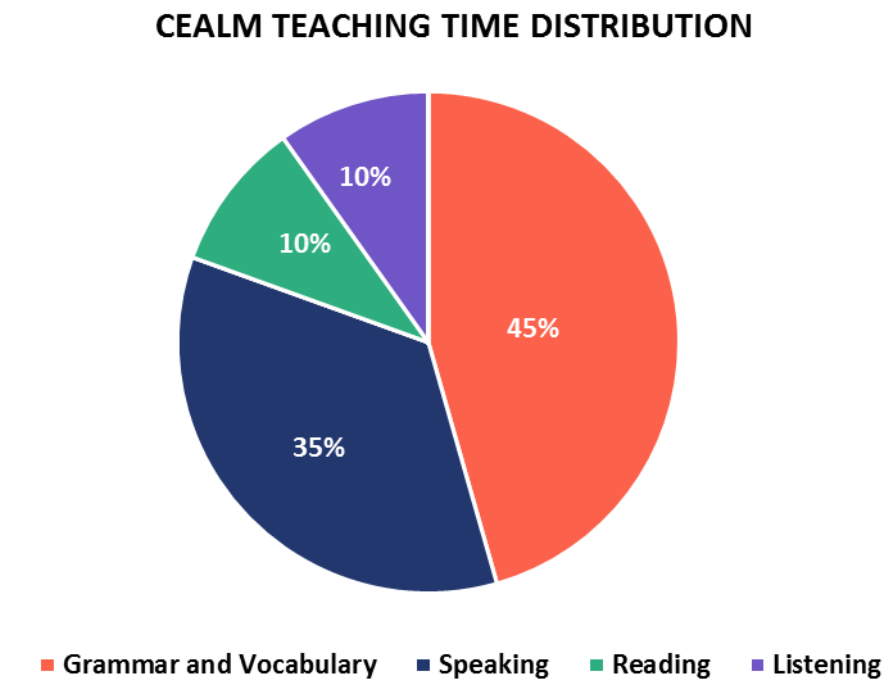


Figure 16. CEALM teaching time distribution

To summarise, the results obtained by CEALM students show that the overall score does not change from Entry to End-of-Course tests although there is an improvement in productive skills. The average score does not reach the pass mark either in the whole exam or in the different papers either in the Entry test or in the End-of-Course test. This would support the position expressed by the majority of the teachers that took part in this study and with the literature on the topic (Graddol, 2006:96 and Green, 2007a:86) that students need more than 120 hours to move from a

CEFR level to another. Alternatively, the results could be influenced by the fact that the last part of the study took place at the end of the academic year. This proximity could have had a greater impact than expected on students' concentration and energy – as reported by Muñoz, Véliz-Campos and Véliz (2019:115) – and this was reflected on their performance. It could also be argued that students did not see these exams as important and, hence, did not show their best performance in them.

5.1.4. Skills profile

This section is going to look into the skills profile to identify the strongest and the weakest skills. The results are reported in Figure 17. The strengths and weaknesses remain the same at the beginning and at the end of this project for participants studying at both institutions. CEB students are stronger at speaking, which is in line with the observations carried out because they show that it is the second most frequently practised linguistic aspect. In addition, students seem to give greater importance to developing their speaking as 85% of the learners – including CEB and CEALM students – wanted their lesson to focus mainly on oral skills, i.e. speaking and listening. The second strongest skills for CEB students are writing and listening. Looking at the results of the observation carried out, writing follows speaking in terms of time devoted in class – 10% of total teaching time. Having writing as the second strongest skill contrasts with the results obtained by Craven (2012; cited by Allen, 2016:10 and by Humprey et al., 2012), who found a weaker performance in writing. As for listening, the same amount of time is devoted to this skill as to reading. The weakest performance is found in the Use of English part although CEB students improved significantly from the beginning to the end of the course. This improvement could be explained by the attention paid to grammar and vocabulary in class as it represents 42% of the total teaching time.

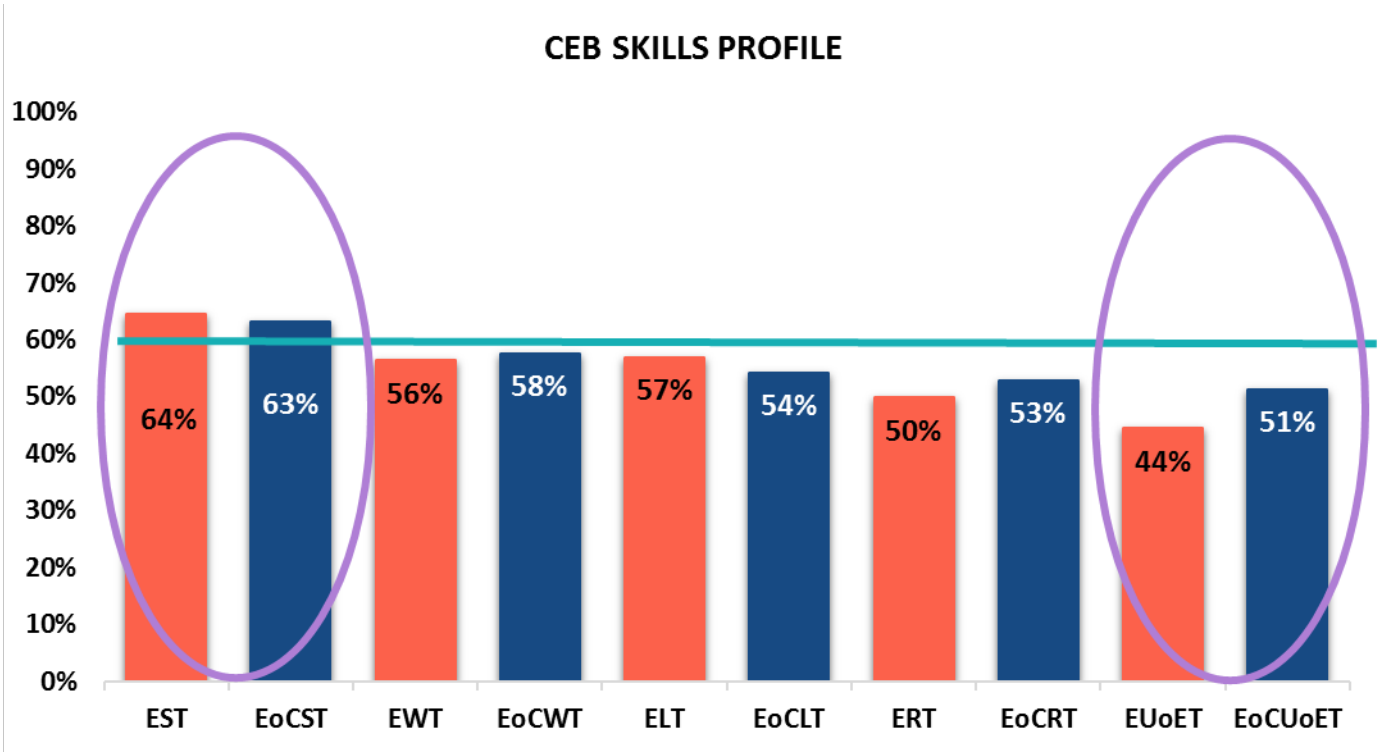


Figure 17. CEB students: skills profile

EST - Entry Speaking Test; EoCST - End-of-Course Speaking Test; EWT - Entry Writing Test; EoCWT - End-of-Course Writing Test; ELT - Entry Listening Test; EoCLT - End-of-Course Listening Test; ERT - Entry Reading Test; EoCRT - End-of-Course Reading Test; EUoET - Entry Use of English Test; EoCUoET - End-of-Course Use of English Test

When asked about the aspects that are more challenging in terms of grammar and vocabulary – language contents explicitly tested in the Use of English paper, the main problem seems to be related to the exam format and the strategies to do the tasks successfully, which can be seen as a clear influence of Cambridge B2 First exam (Figure 18). The second most common problem is how vast grammar and vocabulary are in English and the structures that students at B2 level need to be familiar with. This is a difficulty for learners specially in their first year of preparation because they may not have had the time to cover the syllabus of the level. The most worrying finding is that almost 50% of participants could not provide a specific reason why this component is difficult, which poses a problem for improvement because if students cannot identify what they find difficult, they will not know how to solve the problem, or left the question answered. The latter could suggest that they are not aware of their difficulties or that they do not feel engaged with the questionnaire, which is a limitation that other researchers have experienced

in their own research as reported by Muñoz, Véliz-Campos and Véliz (2019:115). The data was obtained from an open-ended question included in the End-of-Course Questionnaire.

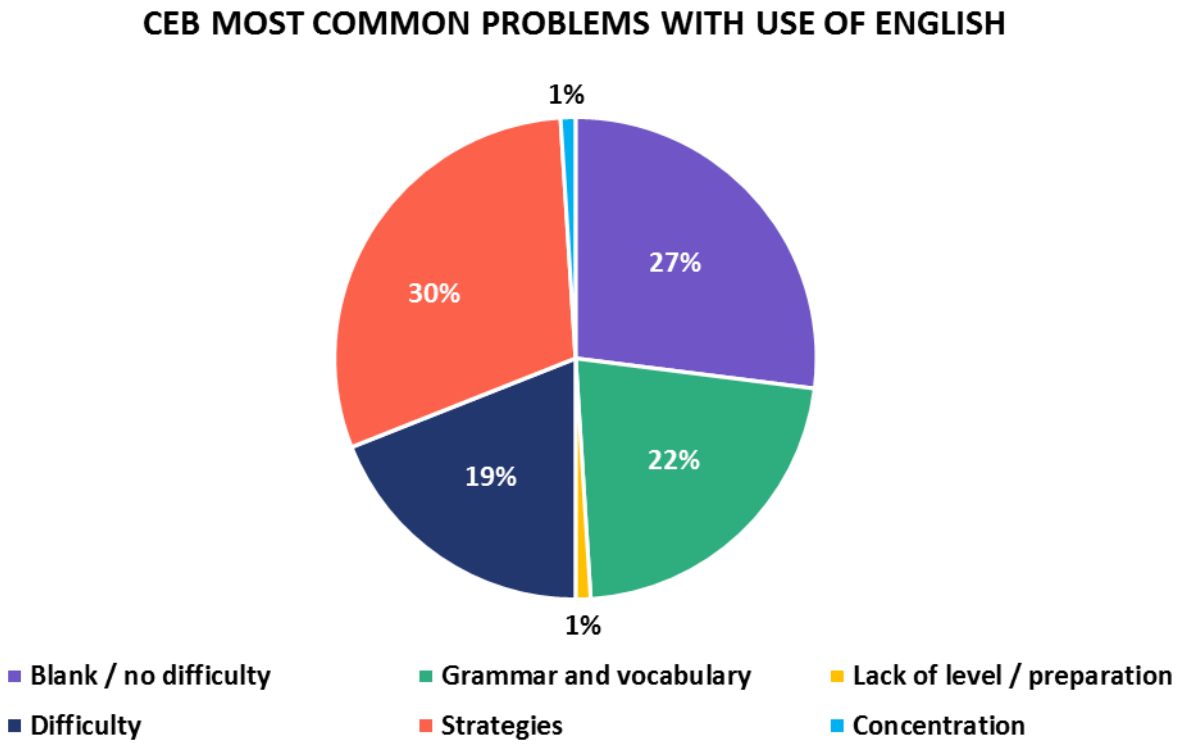


Figure 18. CEB Students: most common problems with Use of English

Looking at CEALM students' skill profile, reported in Figure 19, speaking is also the strongest skill. This, like with CEB, could be explained by the fact that it is the second most frequently practised linguistic aspect in lessons and by the fact that it is an important aspect to master for learners in general. The second strongest skills for CEALM students are writing and reading, which contrasts with the results obtained by Craven (2012; cited by Allen, 2016:10 and by Humprey et al., 2012).

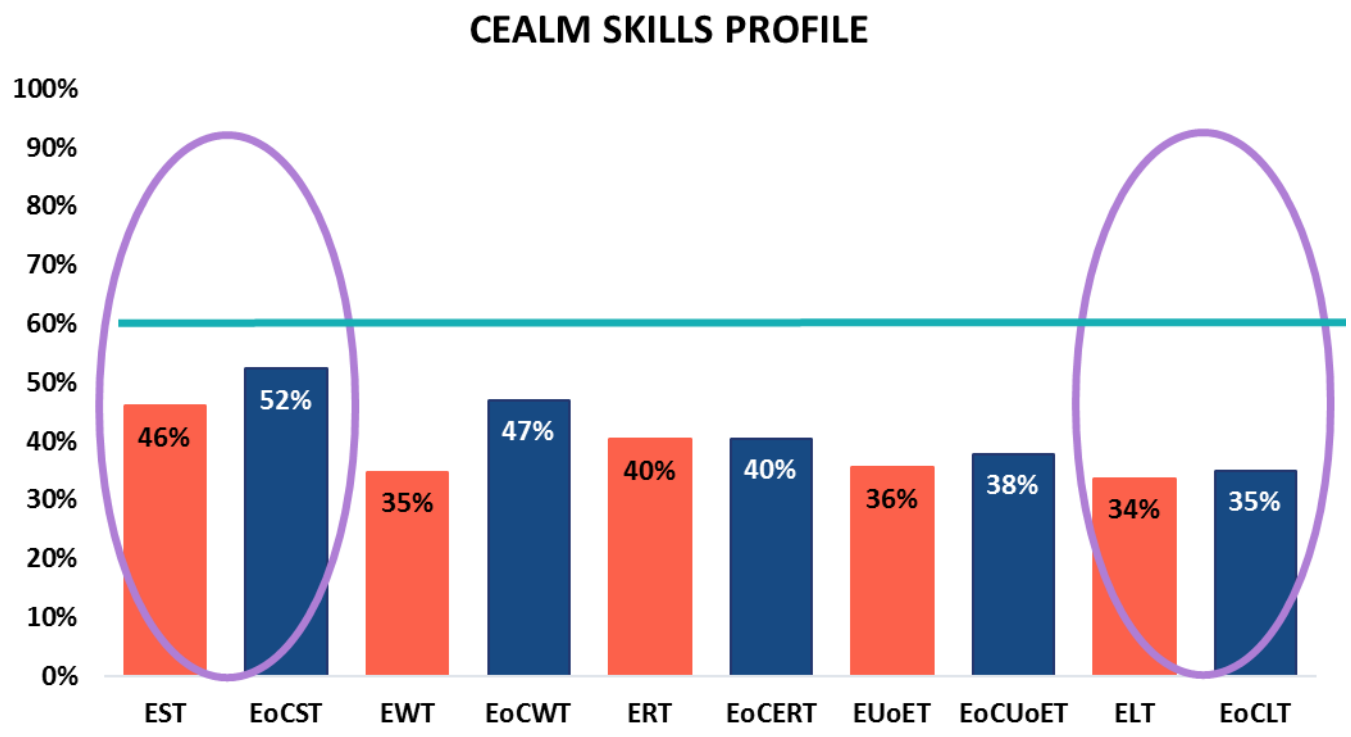


Figure 19. CEALM students: skills profile

EST - Entry Speaking Test; EoCST - End-of-Course Speaking Test; EWT - Entry Writing Test; EoCWT - End-of-Course Writing Test; ELT - Entry Listening Test; EoCLT - End-of-Course Listening Test; ERT - Entry Reading Test; EoCRT - End-of-Course Reading Test; EUoET - Entry Use of English Test; EoCUoET - End-of-Course Use of English Test

These findings cannot be explained by the observation schedule (Figure 16 above) because no teaching time was devoted to writing at CEALM during the observation period and reading received the same attention as listening with 10% of the total class time. Performance in the Use of English part of the B2 First exam is also weak although 45% of the total teaching time is devoted to grammar and vocabulary. Regarding the weakest skill, it is listening. This finding cannot be linked to the amount of time devoted to it in class because, as we have just mentioned, it received the same teaching time as reading. However, it could be connected to the fact that the observation records show that listening was only practised with course book material. Therefore, it may be a good idea to include some authentic material.

To try to have a better understanding of why CEALM students struggle with listening, we looked at the participants' answers to the question enquiring about their main problems with this skill, which are reported in Figure 20. Most CEALM students left the question unanswered, which may suggest, on the one hand, that they are not aware of their difficulties or, on the other hand, that they do not feel engaged with the questionnaire. The latter is a limitation that other researchers have experienced in their own research as reported by Muñoz, Véliz-Campos and Véliz (2019:115). However, those who answered point at pronunciation related aspects such as speed, accent or connected speech as the main obstacle, while a similar percentage of students identify vocabulary and their ability to identify key information as the main problem. Only 3% think that concentration could have an impact on their performance. This information was obtained from an open-ended question included in the End-of-Course Questionnaires.

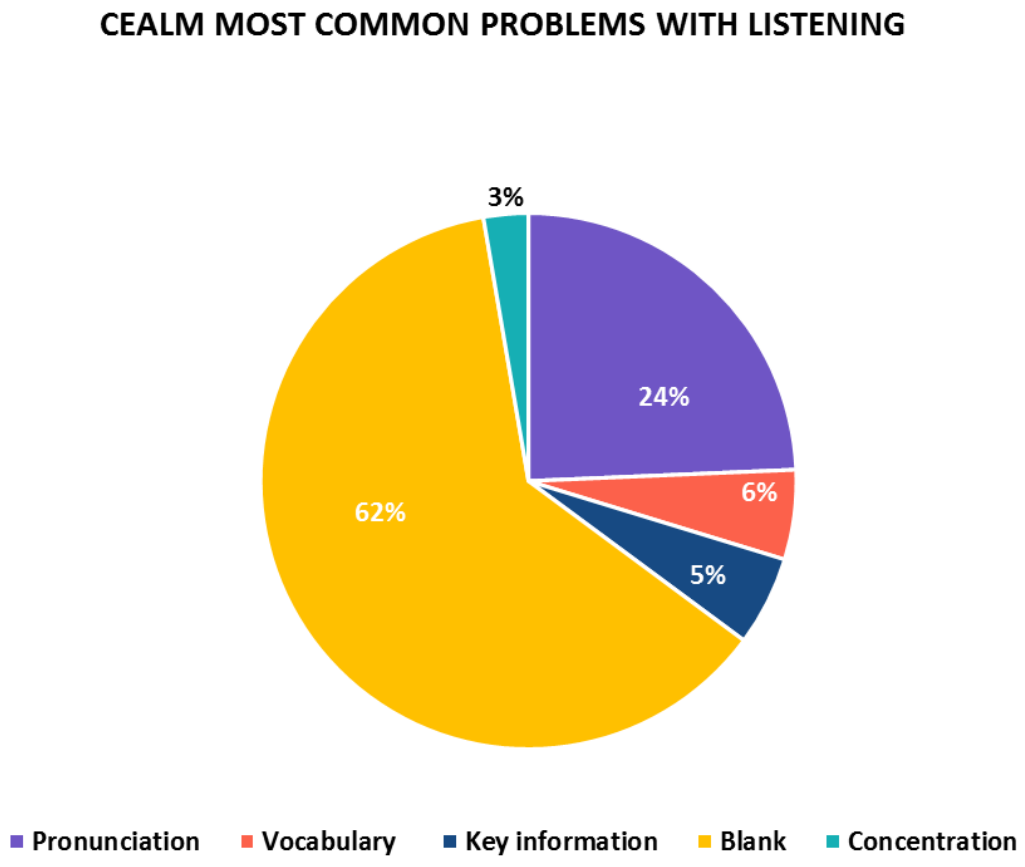


Figure 20. CEALM students: most common problems with listening

To summarise, speaking is the strongest skill for CEB and CEALM students, which can be explained by the time devoted to it in class and by students' interest. As for the weakest skills, they vary from one institution to the other, which could indicate a problem in how institutions teach them.

5.1.5. Comparison between Experimental and Control groups

Results of this part of the present study are provided in Figure 22 below. When looking at potential differences in terms of performance in the Entry B2 First mock exam, the experimental group and the control group perform similarly (p value 0.056) in the ECT. However, a statistical difference can be observed in the EoCCT (p value 0.027) with CEB students' mean reaching 49.34% and CEALM students' achieving 42.50%. None of the groups reaches the pass mark of 60%.

5.1.5.1. Comparisons per skill

5.1.5.1.1. Speaking

In general, the experimental group shows a stronger performance in the Speaking Paper, both in the EST (60.78% vs. 45.81%) (p value <0.001) and in the EoCST (58.89% vs. 53.04%) (p value 0.024). The means also show that speaking is one of the strongest skills and that the experimental group, despite being in their first year of preparation, are almost ready to pass the B2 First Speaking component.

5.1.5.1.2. Writing

The results of the Writing Paper show that in the EWT the experimental group is stronger than the control group (50% vs. 32.78%) (p value <0.001). However, this difference is not statistically significant at the end of the course (p value 0.074). The mean scores in the EWT and the EoCWT (49.38% vs. 43.57%) show that the experimental and the control groups still need to work on this skill to reach the B2 level.

5.1.5.1.3. Listening

In general, the experimental group is stronger at listening, as shown by the mean scores obtained on the ELT (50.10% vs. 30.78%) (p value <0.001) and in the EoCLT (48.33% vs. 32.03%) (p value <0.001). As in the case of writing, the experimental and the control groups need to continue working on this skill to reach the B2 level.

5.1.5.1.4. Reading

The results of the Reading Paper show no statistically significant difference between the experimental and the control groups. This is true for the ERT and for the EoCRT: at the beginning of the study the mean score obtained by the experimental group is 42.71% while the control group achieves 41.71% (p value 0.840); at the end of the study the mean score obtained by the experimental group is 46.39% and the mean score obtained by the control group is 39.01% (p value 0.079). If we look at the mean scores, it is clear that students still need to work on this skill to reach the pass mark.

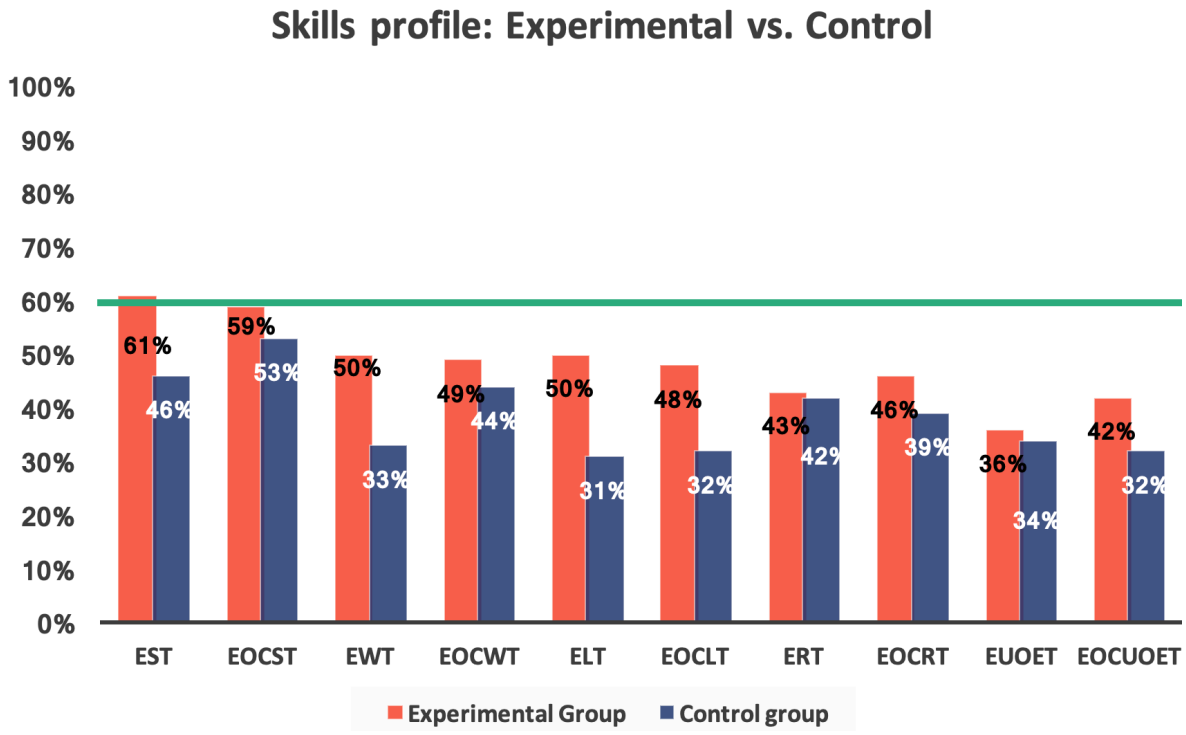
5.1.5.1.5. Use of English

When looking at the results of the Use of English component, it can be seen that in the EUoET the difference between the experimental group and the control group (36.25% vs. 33.57%) is not statistically significant (p value 0.501) although it is clear that neither group shows B2 level abilities yet. When the results of the EoCUoET are analysed, the experimental group shows a stronger performance (41.74%) than the control group (32.42%) (p value 0.036). However, the mean scores are still far from what is expected of B2 candidates.

Looking at performance in Cambridge B2 First mock exam, the mean scores of the experimental and the control group at the beginning of the study are similar as differences are not statistically significant. However, the experimental group shows a stronger performance at the end of the study. This contrasts with the results obtained by Robb and Ercanbrack (1999:18), which evidenced in all instances but one that all students had a similar score gain regardless of their method of study, and with the data reported by Celestine and Su Ming (1999; cited by in Rao et al., 2003:241), who found no significant difference in IELTS results of students who attended preparation and those who did not. This finding also goes against Perrone's (2010 and 2011; cited

by Cheng, Sun & Ma, 2015:458) investigation, which reported that students' mean scores on FCE in the FCE preparation course and the general ETL course were not significantly different. However, it is in line with the results reported by Saif (2006:26), which show that the experimental group performed significantly better than the control group. Going back to this study, the fact that performance does not reach the level expected of B2 candidates supports the idea that students need more than 120 hours to raise one CEFR level to the next and is in line with previous findings and the literature reviewed.

When considering individual skills and components (Figure 21), the skills profile of the experimental and the control group have some similarities because for both groups speaking is the strongest skill. However, the experimental group is stronger at productive skills because writing is their second most solid skill. It is closely followed by listening, then reading, and finally Use of English, which is the component showing the weakest performance. In the case of the control group, reading is the second strongest skill, followed by writing, then use of English, and finally listening, which is the weakest skill.



Figur

e 21. Skills profile: Experimental group vs. Control group

EST - Entry Speaking Test; EoCST - End-of-Course Speaking Test; EWT - Entry Writing Test; EoCWT - End-of-Course Writing Test; ELT - Entry Listening Test; EoCLT - End-of-Course Listening Test
ERT - Entry Reading Test; EoCRT - End-of-Course Reading Test; EUoET - Entry Use of English Test; EoCUoET - End-of-Course Use of English Test

The fact that writing is the second strongest skill for the experimental group while for the control group it is the third may be partly associated with the fact that class time is devoted to writing at CEB (see Figure 13 above) while writing practice was set for homework, at least during the observation period, at CEALM (see Figure 16 above). In terms of reading, CEALM pays more attention to this skill with 10% of the teaching time devoted to it as opposed to CEB, where reading was not practised in class during the observation period. This could be thought to be a factor in the relatively stronger performance of the control group in this skill. To try to understand the reason why listening is the weakest skill for the control group, we would probably need to look at the percentage of students who gave no answer when asked about their main problem with listening (see Figure 20 above) because this may show that students are unaware of their strengths and weaknesses. Classroom observation does not shed much light as more time is devoted to listening

in class at CEALM than at CEB. Nevertheless, the fact that authentic material is used at CEB while at CEALM the only resource used during observation was the course book could also explain the problems that CEALM students have with listening comprehension.

5.2. Do students enrolled on more exam-oriented (CEB) courses improve their language knowledge and abilities?

The Grammar Test is applied as an independent test in order to find out about the grammatical competence of students. The first tests whose results are going to be compared are the Entry Grammar Test and the End-of-Course Grammar Test. Then, the results in the Vocabulary Test, which is applied as an independent test in order to find out about the lexical competence of students, will be discussed. Finally, the results in the independent tests will be compared with those obtained on the Use of English component to understand the potential impact of B2 First preparation on results.

5.2.1. Students enrolled on more exam-oriented (CEB) courses

First, if we take CEB students as a whole and we compare their performance in the EGT and the EoCGT (Table 5), it can be observed that the performance is weaker in the EoCGT as the mean is 64.21% while in the EGT the mean is 71.10% (p value <0.001). When we compare CEB students' performance as a whole in the Entry Vocabulary Test (EVT) (46.17%) and the End-of-Course Vocabulary Test (EoCVT) (44.32%), no statistically significant difference can be observed (p value 0.179).

Table 5. CEB students: independent tests

Independent tests	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EGT[6]	71	**71.70%**	12.68	0.552	<0.001	*
EoCGT[7]	71	64.21%	14.39			
EVT[8]	71	46.17%	15.03	0.122	0.179	
EoCVT[9]	71	44.32%	15.18			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Grammar Test; [7] End-of-Course Grammar Test; [8] Entry Vocabulary Test; [9] End-of-Couse Vocabulary Test

Given that grammar and vocabulary, although tested in all Cambridge exam papers, are tested more explicitly in the Use of English component, it is considered useful to compare the results obtained on this test with the independent grammar and vocabulary tests. Results for CEB students are reported in Table 6 and Table 7 and show that performance in the EGT is stronger than in the EUoET (71.43% vs. 44.80%) (p value <0.001). The same is true for the EoCGT and the EoCUoET (63.88% vs. 52.38%) (p value p value <0.001). As for the vocabulary test and the UoE test, performance is similar in the EVT and the EUoET (46.14% vs. 44.29%) (p value 0.324) and shows that students need to work harder to reach the pass mark. If the end-of-course tests are compared, CEB students show a stronger performance in the EoCUoET as compared to the EoCVT (44.44% vs. 51.61%) (p value <0.001).

Table 6. CEB students: Entry Independent Tests and UoE Tests

Entry Independent Test vs. UoE Test	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EUoET[6]	77	44.80%	17.94	-1.710	<0.001	*
EGT[7]	77	**71.43%**	12.77			
EUoET[8]	80	44.29%	17.76	-0.114	0.324	
EVT[9]	80	46.14%	14.62			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Use of English Test; [7] Entry Grammar Test; [8] Entry Use of English Test; [9] Entry Vocabulary Test

Table 7. CEB students: End-of-Course Independent Tests and UoE Tests

End-of-Course Independent Tests vs. UoE Test	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EoCUoET	73	**51.61%**	18.95	0.42	<0.001	*
EoCVT	73	44.44%	14.99			
EoCUoET	75	52.38%	18.71	-0.689	<0.001	*
EoCGT	75	**63.88%**	14.39			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] End-of-Course Use of English Test; [7] End-of-Course Vocabulary Test; [8] End-of-Course Use of English Test; [9] End-of-Course Grammar Test

When looking at how CEB students performed in the independent tests, i.e. grammar and vocabulary tests, results show that the performance in grammar at the end of the year was weaker than at the beginning of the study although the mean score is still above the pass mark. If the results in the vocabulary test are analysed, no statistically significant difference can be found in performance between the beginning and the end of the project and, in general, performance is

weak as students do not reach the pass mark. It is difficult to explain these negative results if we consider that the lexico-grammatical component was the aspect that received the greatest attention – with 42% of total class time – and only 13% of CEB students seemed to be against a major focus on grammar and vocabulary in the class.

At this point, comparing the results in the independent tests and in the Use of English component of B2 First mock exam may be useful to try to have a better understanding of the results. In this sense, at the beginning of the project students' performance in the EVT and the EUoET is similar in statistical terms and students performed more strongly in the EGT when compared to the EUoET. When looking at results at the end of the project, CEB students did better in the EoCGT than in the EoCUoET but did better in this test when compared with the EoCVT. This may suggest that students struggle mainly with lexis. In fact, the results of the questionnaire show that 42% of CEB students identified vocabulary as the main problem in this case when facing a reading task.

All in all, if we used this data to answer the second research question i.e. whether courses which are more exam-oriented help students to improve their language knowledge and abilities, it would be difficult to establish as the results of CEB students as a whole seem to indicate that either the time between the administration of the entry test and the end-of-course test was too little for students to improve significantly from a statistical point of view or the fact that the last part of the study took place at the end of the academic year had a greater impact than expected on students' concentration and energy – as reported by Muñoz, Véliz-Campos and Véliz (2019:115) – and this was reflected on their performance. Another factor that should be taken into account is the exam stakes: the grammar and vocabulary tests could be perceived as low-stakes because they were administered to CEB students as an "extra"; however, could be seen the B2 First mock exams could be seen as high-stakes since they are used, one the one hand, to decide if students pass to the next level the following year and, on the other hand, to decide if students are ready to sit the B2 First exam.

Regardless of the reasons for this poor performance of CEB students, the fact that students did not improve their performance was also reported by Muñoz and Álvarez (2010), who stated that there are few studies that can report verifiable gains in students' learning.

As for the relationship between performance in the grammar tests, vocabulary tests and Use of English component, it seems to suggest that vocabulary is not acquired successfully by CEB students despite being the linguistic component, together with grammar, that is most frequently covered in class. This weakness not only affects the Use of English results but also impacts the reading skills. Consequently, it would be advisable to look into the reasons for this situation to try to mitigate this problem.

5.2.2. Comparison between Year 1 and Year 2 students enrolled on more exam-oriented (CEB) courses

The results on independent tests of Year 1 and Year 2 students are reported in Table 8. If we look now at how they performed in the EGT and the EoCGT, Year 1 students' performance is weaker both in the EGT (66.28% vs. 74.12%) (p value 0.007) and the EoCGT (56.82% vs. 68.59%) (p value < 0.001). Comparing Year 1 students and Year 2 students' results in the EVT (40.13% vs. 50.16%) and the EoCVT (36.56 vs. 50.26%) a better performance of Year 2 students in both cases is observed.

Table 8. Comparison of results on independent tests of Year 1 and Year 2 students

Independent tests	Year[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EGT[6]	1 2	66.28% **74.12%**	10.22 13.39	-0,639	0.007	*
EoCGT[7]	1 2	56.82% **68.59%**	13.6 13.03	-0.888	<0.001	*
EVT[8]	1 2	40.13% **50.16%**	12.09 14.84	-0.728	0.002	*
EoCVT[9]	1 2	36.56% **50.26%**	14.34 12.75	-1	<0.001	*

[1] Year students are in; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Grammar Test; [7] End-of-Course Grammar Test; [8] Entry Vocabulary Test; [9] End-of-Course Vocabulary Test

If the independent tests are considered, Year 2 students show a stronger performance both at the beginning and at the end of the project and, in fact, they reach the pass mark in all the independent tests. The results in the grammar test are the most solid ones both for Year 1 and Year 2 students as both groups show abilities compatible with what is expected of B2 learners already in the entry test. When looking at vocabulary, the mean scores – just on the pass mark – show that Year 2 learners have acquired the lexical competence typical of the B2 level and show a stronger performance than Year 1, who do not pass the vocabulary tests. These results seem to suggest that, on the one hand, time devoted to grammar and vocabulary in class builds up and is really shown in Year 2 mean scores and, on the other hand, the lexical aspect of language is the main problem for CEB students, as mentioned above, which may suggest that a different approach to teaching vocabulary is required. As already discussed, there might be a connection between reading ability and lexical knowledge since students identified vocabulary as one of the main obstacles to succeed in the reading paper.

5.2.3. Students enrolled on general English (CEALM) courses

In the case of CEALM students again taking them as a whole, an improvement can be observed as the mean for the EGT is 58.86% and for the EoCGT is 63.86% (p value 0.015) (Table 9). However, this is not the case for the vocabulary as the difference between results in the EVT (54.70%) and the EoCVT (50.68%) is not statistically significant (p value 0.091). These results are shown in Table 9.

Table 9. CEALM students: independent tests

Independent tests	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EGT[6]	20	58.86%	11.9331	-0.431	0.015	*
EoCGT[7]	20	**63.86%**	11.27718			
EVT[8]	21	54.70%	11.46602	0.309	0.091	
EoCVT[9]	21	50.68%	14.41226			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Grammar Test; [7] End-of-Course Grammar Test; [8] Entry Vocabulary Test; [9] End-of-Course Vocabulary Test

As with CEB, CEALM students' performance in the independent tests was compared with results in the UoE paper and results are reported in Table 10 and Table 11. CEALM students show better performance in the independent tests both at entry and at the end of the project. The mean score of the EGT (59.52%) show that they have reached the pass mark while they are still far away from it if we look at the EUoET (35.20%) ($p < 0.001$), if we look at the EVT and the EUoET, the results are very similar with students passing the former (54.78%) but failing the latter (34.47%) ($p < 0.001$).

Table 10. CEALM students: Entry Independent Tests and UoE Tests

Entry Independent Test vs. UoE Test	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EUoET[6]	21	35.20%	13.47	2.047	<0.001	*
EGT[7]	21	**59.52%**	10.05			
EUoET[8]	23	34.47%	13.17	1.665	<0.001	*
EVT[9]	23	**54.78%**	11.14			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] Entry Use of English Test; [7] Entry Grammar Test; [8] Entry Use of English Test; [9] Entry Vocabulary Test

The same situation is replicated at the end of the project (Table 11) because learners show a stronger performance in the independent tests (EoCGT 65.21% vs. 33.85% EoCUoET and EoCVT 50% vs. 34.25%) than in the UoE test. It must be noted that CEALM students pass all the independent tests despite being Year 1 students.

Table 11. CEALM students: End-of-Course Independent Tests and UoE Tests

End-of-Course Independent Test vs. UoE Test	N[1]	Mean[2]	SD[3]	Cohen's D[4]	P value[5]	Statistical significance (*)
EoCUoET[6]	23	33.85%	15.15	2.400	<0.001	*
EoCGT[7]	23	**65.21%**	10.58			
EoCUoET[8]	22	34.25%	15.38	1.060	0.011	*
EoCVT[9]	22	**50.00%**	14.32			

[1] Number of students; [2] Mean value of scores (maximum 100%); [3] Standard Deviation of scores; [4] Effect Size; [5] Significance of comparative analysis; [6] End-of-Course Use of English Test; [7] End-of-Course Grammar Test; [8] End-of-Course Use of English Test; [9] End-of-Course Vocabulary Test

When looking at the lexico-grammatical ability tested using the independent grammar and vocabulary tests, CEALM students improve in terms of grammar and pass the EGT and the EoCGT. As for their lexical ability, CEALM students passed both the EVT and the EoCVT although they did not show improvement from the beginning to the end of this study. The positive results in grammar and vocabulary obtained on the independent tests could be explained by the fact that grammar and vocabulary are the most often practised aspects during class time with 45% of the teaching time, and by the interest expressed by 78% of CEALM students, who would like lessons to focus mainly on grammar and vocabulary.

The positive results in the independent tests do not match the weak performance in the Use of English component, where CEALM students do not improve and stay below the pass mark. If we compare the lexical ability of CEALM students measured by Cambridge Use of English test and the vocabulary test, performance in the latter is stronger both at the beginning and at the end of the project. As for grammatical ability, similar results are found, with CEALM students passing both the EGT and the EoCGT but failing the EUoET and the EoCUoET.

The stronger results in the independent tests could evidence the fact that having no frequent practice in the format of Cambridge exams may have a negative impact on performance.

5.2.4. Comparison between Experimental and Control groups

The results obtained after comparing the performance of the experimental group and the control group are reported in Table 12. If the results in the EGT are considered, the experimental group shows a stronger performance than the control group (66.28% vs. 54.94%) (p value <0.001). Moreover, both groups' means show that they reach the 50% pass mark. However, the opposite happens when we look at the results for the EoCGT because the experimental group shows a weaker performance than the control group (56.82% vs. 64.91%) (p value 0.018) although, again, both groups reach the pass mark[5]. If the results in the EVT are analysed, the control group outperforms the experimental group (40.13% vs. 52.76%) (p value <0.001) and the same can be observed in the EoCVT (36.56% vs. 49.21%). In general, the performance in the Vocabulary Test is weak as students do not reach the pass mark in most cases.

Table 12. Comparison of results obtained by the Experimental group and the Control group on the independent tests and the Use of English tests

Independent tests	Population[1]	N[2]	Mean[3]	SD[4]	Cohen's D[5]	P value[6]	Statistical significance (*)
EGT[7]	Exp. Group	31	**66.28%**	10.22	0.916	<0.001	*
	Control Group	35	54.94%	14.00			
EoCGT[8]	Exp. Group	30	56.82%	13.60	-0.661	0.018	*
	Control Group	25	**64.91%**	10.38			
EVT[9]	Exp.Group	34	40.13%	12.09	-1.023	<0.001	*
	Control Group	37	**52.76%**	12.59			
EoCVT[10]	Exp. Group	31	36.56%	14.34	-0.882	0.002	*
	Control Group	24	**49.21%**	14.36			

[1] Population analysed; [2] Number of students; [3] Mean value of scores (maximum 100%); [4]Standard Deviation of scores; [5]Effect Size; [6]Significance of comparative analysis; [7]Entry Grammar Test; [8]End-of-Course Grammar Test; [9] Entry Vocabulary Test; [10]End-of-Course Vocabulary Test

5 The pass mark for the Grammar Tests and the Vocabulary Tests is 50%.

Both the experimental and the control group show a stronger performance in the grammatical competence since both groups pass the EGT and the EoCGT. However, while at the beginning of this project the experimental group had a more solid knowledge of grammar, the situation was the opposite at the end of the study. This finding together with the stronger performance in terms of lexical knowledge of the control group at the beginning and at the end of the project, could suggest that the time devoted to grammar and vocabulary in class at CEB (Figure 22) is used less effectively than at CEALM (Figure 23), even more so if we take into account that the time devoted to grammar and vocabulary in the latter is almost double (compare Figure 22 and Figure 23). Nevertheless, there is a caveat to this statement because CEALM students passed the EVT but their mean score was below the pass mark in the EoCVT.

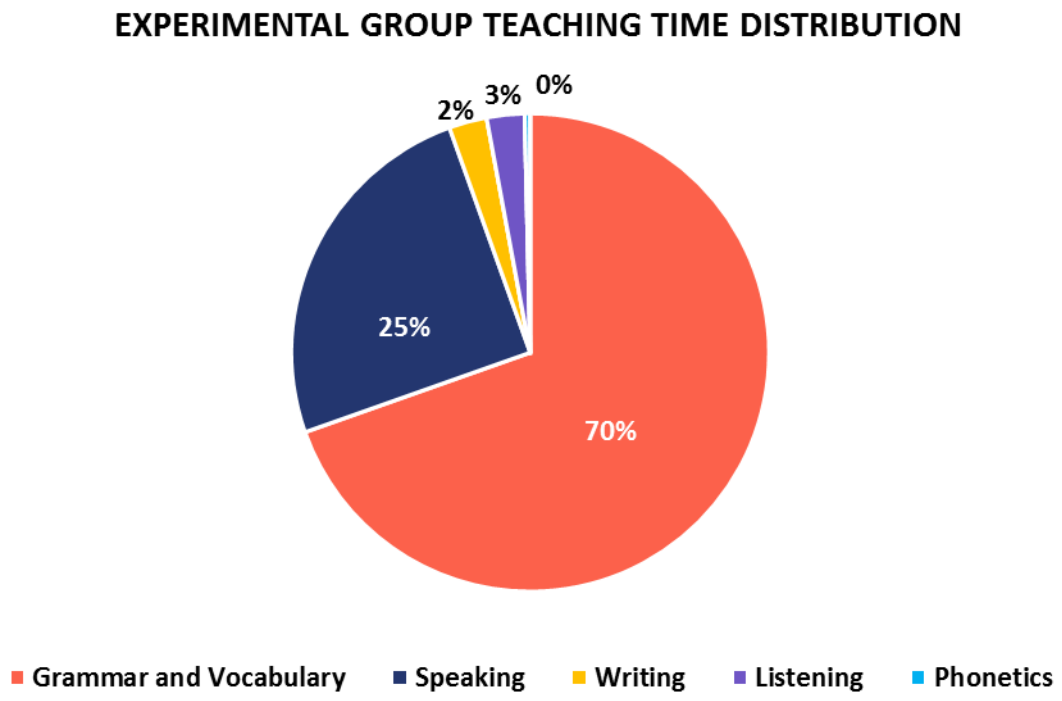


Figure 22. Experimental group: teaching time distribution

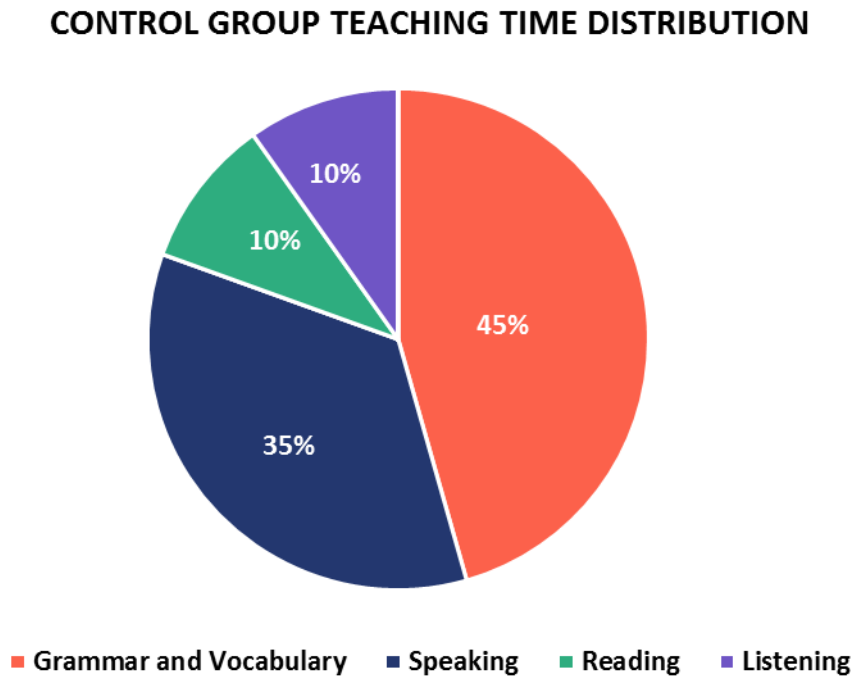


Figure 23. Control group: teaching time distribution

It is difficult to give a clear answer to the second research question because different variables need to be taken into account. If we look at the grammar component, it is clear that CEB students not only do not improve but actually show a weaker performance. Conversely, CEALM students improve their knowledge of grammar in the course. If we focus on vocabulary, neither CEB nor CEALM show a difference in terms of increased lexical knowledge that is statistically significant from the beginning to the end of the study. If we compare the experimental and the control group, although the experimental showed better grammatical knowledge at the beginning of the study, the control group outperformed them at the end. Moreover, the control group had better lexical knowledge at the beginning and at the end of the project. With this information in mind, we could conclude that students enrolled on a more exam-oriented course do not improve their language knowledge and ability although it would be necessary to read also the small print to fully understand the situation.

If we consider the relationship between the independent tests and the Use of English component, CEB students are always stronger at grammar, which supports the idea that the main problem CEB students face is vocabulary. CEALM students always show a stronger performance in the independent tests, which could suggest that the lack of exam practice has a negative effect on test results at least in terms of the ability to use grammar and vocabulary.

5.3. Do students become more autonomous and independent learners as a result of preparing for Cambridge B2 First exam?

This section tries to answer the third research question. For that purpose, some questions from the EQ and the EoCQ have been analysed.

5.3.1. Awareness of their own abilities and difficulties

At the beginning of this project, learners were asked if they knew how to improve their level of English and their answers clearly showed that most of them – 95% of CEALM and 92% of CEB students – know how to do it.

This question appears again in the EoCQ to try to find out if the courses changed the learners' perspective and the results obtained were very similar, as 95% of CEALM and 97% of CEB students gave a positive answer. The only increase can be found in CEB students, since some of those who were unsure about how to improve their English took advantage of the course to identify aspects that help them with their English level.

In a similar vein, the EQ and the EoCQ include a question asking learners about their ability to identify their strengths and weaknesses, and at the beginning of the study students' answers show that they were quite confident about their ability because 86% of CEALM students and 92% of CEB learners answered positively. When students were asked this question again at the end of the project, the results were very similar, as 91% of CEALM students and 92% of CEB state that they know what their strong and weak points are. In this case a slight increase can be observed in CEALM students, who have used the time and the information in the course to become more aware of their abilities. In fact, the EoQC includes five questions enquiring students about what they find difficult in reading, listening, speaking, writing, and Use of English activities and 67% of CEALM students and 70% of CEB students could identify at least one specific obstacle for each type of

activity. Those students who aimed to sit an accreditation exam in the academic year 2017-2018 were asked if they knew how to improve their performance in the test and here again the vast majority of CEALM and CEB students – 86% and 95% respectively – gave a positive answer. This contrasts with the results obtained by Mickan and Motteram (2009:1), who found that a number of candidates admitted not knowing how to improve their scores, which suggests "a lack of personal agency and strategic action in preparing for testing" and with results reported by Reynolds (2010; cited by Sevilla Morales & Chaves Fernández, 2020:210), which suggested that students were uncertain about the activities that better prepared them for the test. In general, it can be said that the results obtained so far are positive because being aware of one's strengths and one's difficulties is key for improvement. However, we cannot conclude that preparing for B2 First exam had a particular effect on these two variables although it is true that the results were already very high for both institutions.

5.3.2. Exam Preparation

In the next lines, the focus is on the person or resources that guide students in their test preparation practice and on the activities and resources that students use to improve their test performance. Results are reported in Figure 24 and Figure 25. The CEALM students who planned to sit an accreditation exam that academic year reflected on their personal experience to try to make progress in terms of test performance while CEB students identified their teacher as the source of guidance. A similar finding was reported by Mickan and Motteram (2009:20) as their data revealed that test takers considered that to succeed in the exam they needed expert help – usually identifying their teacher as the main reference. Although the teacher's role in test preparation could be linked to students' age, since most CEB students are younger than eighteen, results show that there is no difference between adult and teenager students' answers. Also, it could be said to be linked to the amount of teaching hours received, with maybe Year 1 students' relying more heavily on their teacher's advice; nevertheless, the results show 44% of those who approach their teacher for guidance are Year 2 students. The second most popular option for CEB students is their own personal experience. When looking at CEALM, it is interesting to note, the same percentage of CEALM students (20%) rely on their teacher, their peers, and book and exam tips material to improve their performance in the test.

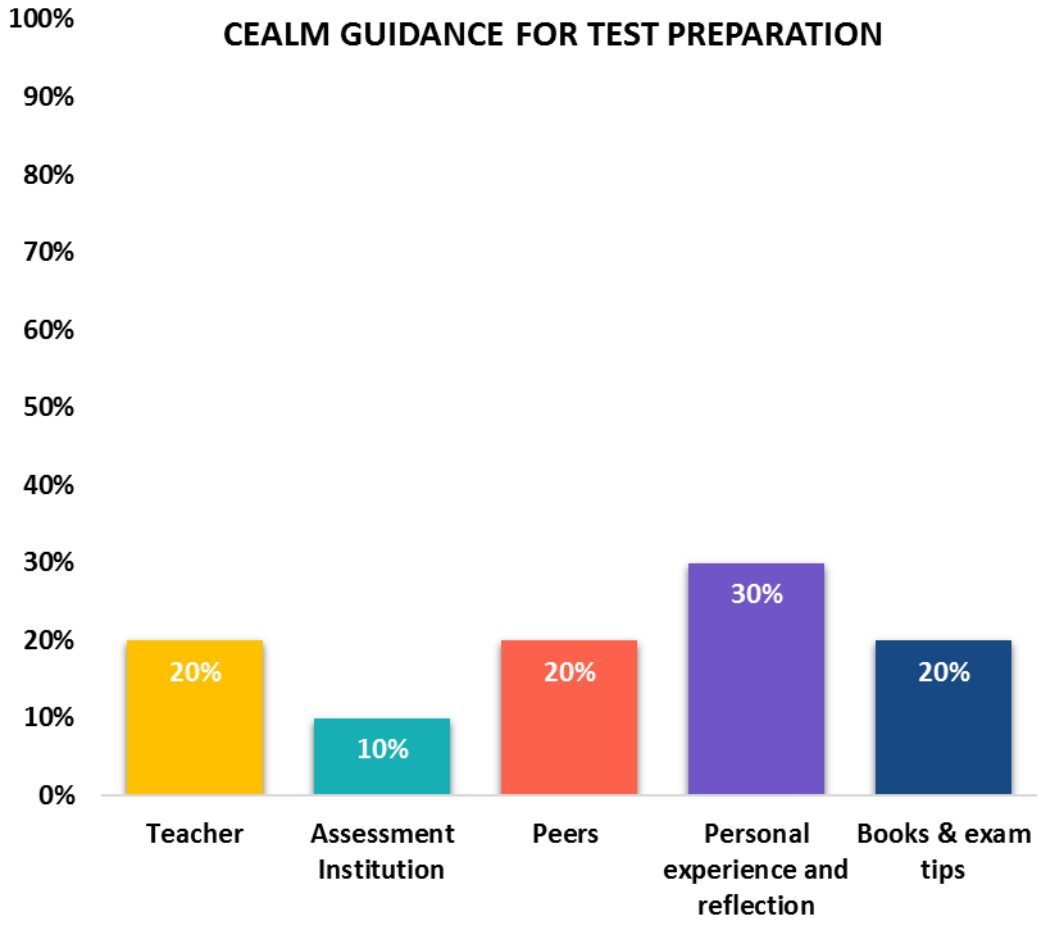


Figure 24. CEALM: guidance for exam preparation

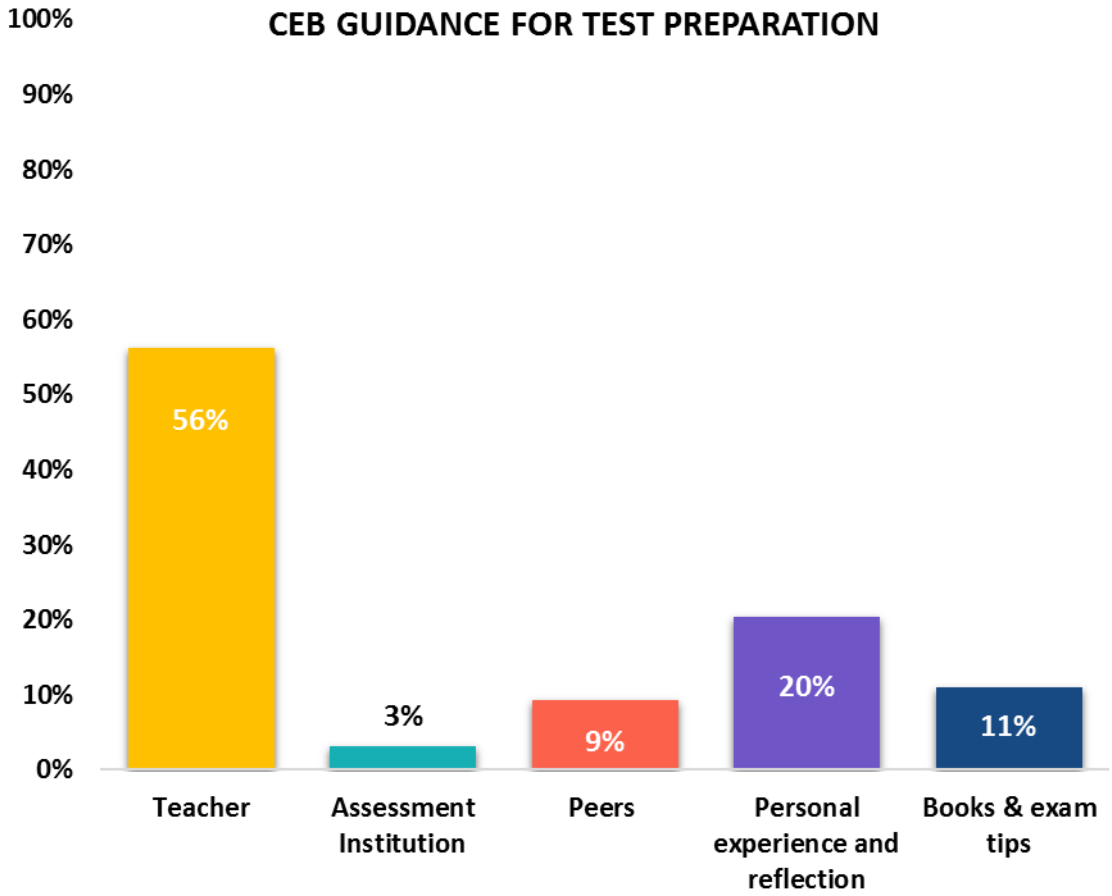


Figure 25. CEB: guidance for exam preparation

When asked if their preparation for Cambridge B2 First exam is different from their preparation for other English exams, the results for CEALM and CEB are very different: only 28% of CEALM students give a positive answer as compared to 61% of CEB students. The questionnaire also enquired students about how their preparation for B2 First differed from preparation for other exams, most of them just added that they study harder. Of those who gave more specific information, most mention doing more exam practice, which is in line with the findings reported by Cheng (1998, 2005), Qi (2004, 2005), and Stoneman (2006), all mentioned by Pan (2014:2). The second most popular option was practising all four skills more. Participants were also asked about how they prepare for a B2 exam – not necessarily B2 First – and their answers showed that for those attending more exam-oriented courses the preferred option was doing exam practice (Figure 26). This coincides with the research studies mentioned above and with the findings obtained by Zhan and Wan (2016:360), and with Mickan and Moterram (2008 and 2009:1), who reported that

most of the participants in their study preferred to use published sample tests to prepare for the exam. As for the rest of the options, they are chosen by a very similar percentage of students. This emphasis on exam practice could be seen as a narrowing of the curriculum and hence as negative washback. Practices such as doing past sample papers and test-like exercises have been reported in previous studies such as Mickam and Motteram (2009), Xie (2013), and Zhan and Andrews (2014), Michaelides (2014), all cited by Zhan and Wan (2016:373), and Külekçi (2016), cited by Toksöz and Kılıçkaya (2017:196). For those students attending general English courses, however, there is a more even spread among different options although the two most frequently chosen ones are studying the grammar and vocabulary that students think could be tested in the exam, and studying English in general, which includes not only class content but also reading books and the press in English, watching TV series or films, and listening to the radio among others (Figure 27). The range in the answers given especially by CEALM students could suggest that the test is not the only factor influencing learner preferences among test preparation activities and that other variables such as prior beliefs about what effective learning also play a role (Green, 2007:75).

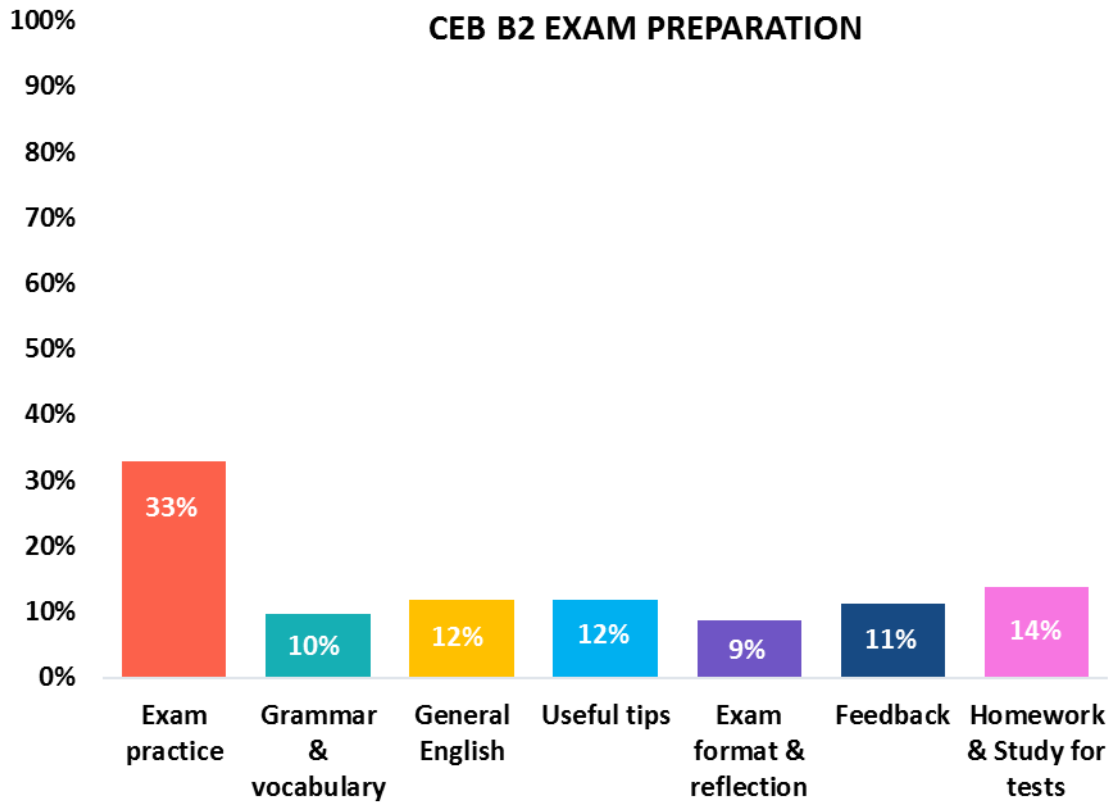


Figure 26. CEB: B2 exam preparation

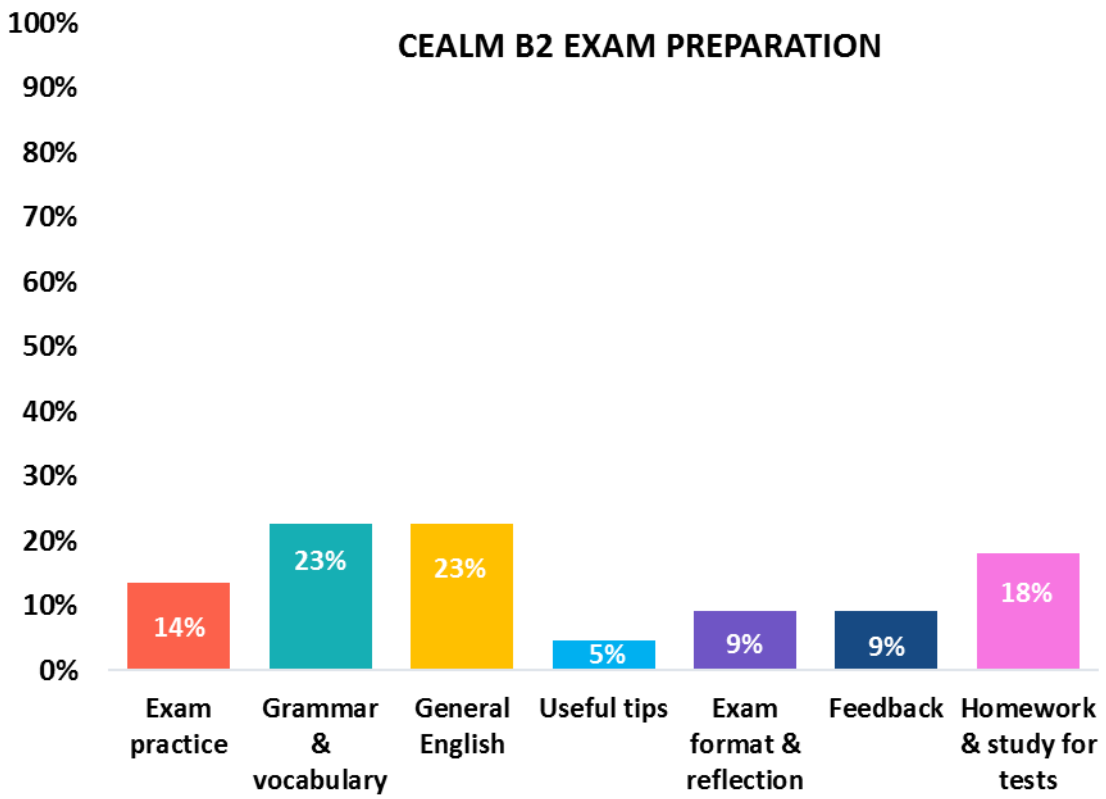


Figure 27. CEALM: B2 exam preparation

To investigate if students' preparation habits were influenced by classroom activities, students were asked about the activities done in class (Figure 28 and Figure 29). The results for both institutions are very similar and actually the top three activities are the same for CEB and CEALM: grammar and vocabulary activities from the course book, activities to practise all four skills and interaction, and exam practice and format activities. This focus on the skills needed to do the test tasks rather than just practising authentic and devised test tasks is called curriculum-teaching by Pompham (2001; cited by Larsson & Olin-Scheller, 2020:4). In this sense, CEB results are in line with the findings reported by Lumley and Stoneman (2000; cited by Tsagari, 2007:50), which show that students are far more exam-oriented than their teachers. From a different point of view, the fact that no major differences can be found between the class activities used in general English courses and in those which are more exam-oriented could contradict the idea expressed by Gabrielatos (1993), Prodromou (1993), Kenny (1995) and reported by Tsagari (2006:7), which pointed to the negative effect of Cambridge B2 First – previously known as First Certificate in English (FCE) – as the authors complained that the exam "enforced traditional ways of teaching, led

to teacher-centred lessons, enforced individualism […] and encouraged considerable expertise in exam and test-taking techniques". However, it agrees with the idea that FCE encourages the teaching of all four skills expressed by Tsagari (2006:9).

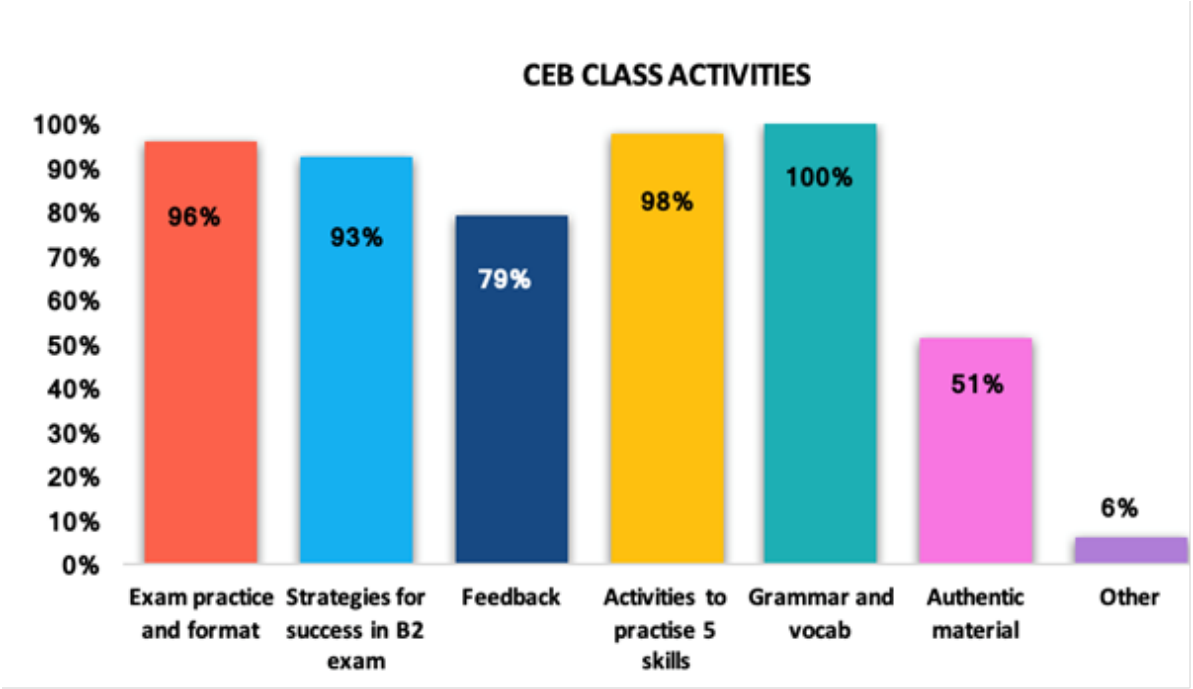


Figure 28. CEB: class activities[6]

[6] Participants could choose more than one option.

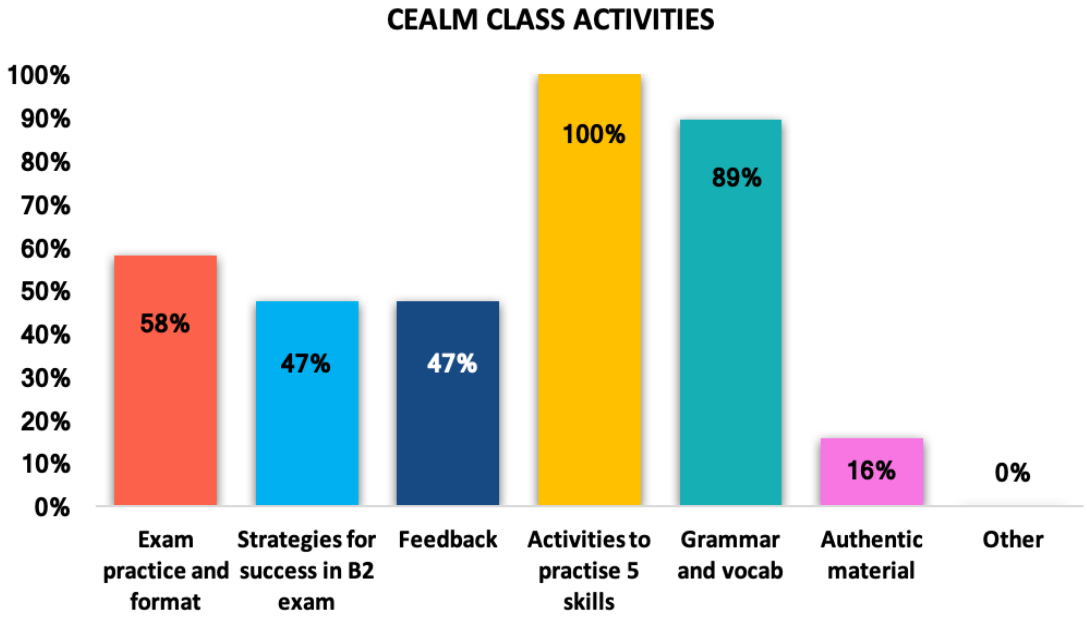


Figure 29. CEALM: class activities[7]

The effect of the exam and the preparation was also analysed from the perspective of motivation. The results of the EQ show that passing a B2 exam is either important or very important for 82% of the participants, which agrees with the findings reported by Tsagari (2006:9) that passing FCE is a source of motivation for students. Also, when participants of CEALM and CEB were asked if aiming to sit a B2 exam encouraged them to work harder, 91% of CEALM and 95% of CEB students gave a positive answer, which is in line with the belief reported by Tsagari (2006:9), Green (2007:6) and Külekçi (2016; cited by in Toksöz & Kılıçkaya, 2017:196). This motivation to work harder is seen by Alderson and Wall (1993) and Ahmad and Rao (2012; cited by Hakim & Tasikmalaya, 2018:63) as an example of positive washback. In addition, the fact that when students are faced with negative exam results only 14% reported experiencing negative feelings goes against the information reported by Harlen and Crick (2003; cited by Tsagari, 2006:304 and by Cheng & Deluca, 2011:107) that point at negative long-term consequences of exams and against the study by Shih (2007:144; cited by Ha, 2019:5), who found detrimental effects on students' self-confidence and motivation for learning English as a result of preparing for the General English Proficiency Test (henceforth GEPT).

[7] Participants could choose more than one option.

However, these results are in line with findings reported by Read and Hayes (2003; cited by Booth & Davis Lee, 2019:19) and Li (1990; cited by Booth & Davis Lee 2019:19), which pointed at positive feelings about the IELTS exam and the Chinese Matriculation English Test and increased motivation to learn English. Cheng (1998; cited by Booth & Davis Lee, 2019:19) also found that Hong Kong Certificate of Education Examination (henceforth HKCEE) encouraged students to work hard to achieve good scores. The results together with participants' answers to the EoCQ about test difficulty also agree with the findings by Watanabe (2001; cited by Booth & Davis Lee, 2019:10), who report that a test could be motivating and have a positive effect on students' preparation if it has the appropriate difficulty for the learner. A total of 64% of CEALM students and 84% of CEB students believed it was possible to pass or were confident that they could pass the B2 First exam after 18 months of preparation. In fact, the results of the EoCQ show that most students (57% of CEALM and 55% of CEB) stated that being faced with negative exam results makes them aware of how to improve, hence increasing their autonomy and independence as learners. Since only 14% of participants mentioned experiencing negative feelings after negative results, the fears reported by Green (2007:6) about potential increases in anxiety – it must be borne in mind that Alderson and Wall (1993) pointed out that we might not want to call anxiety caused by having to take an exam washback – and intimidation were limited in this study. Despite the results of the present study suggesting that negative results make students aware of how to improve, when asked if they had prepared for the second mock exam differently from the first mock exam, most students gave a negative answer (88% in the case of CEALM and 86% in the case of CEB), which could be thought to go against Stoneman's (2006; cited by Cheng, Sun & Ma 2015:447) findings that students' past test-taking experiences have a major influence on their choice of the types of preparation activities, and against Allen's (2016:4), who explains how students' focus changed from the first test to the second from more receptive to more productive skills. Nevertheless, Wesdorp (1982:53; cited by Gosa 2004:42) obtained similar findings when he asked students since the findings suggest that students' habits did not change after the introduction of multiple choice questions. Similarly, Cheng (1998:118; cited by Ha, 2019:6) also found few changes in the learning strategies used by students preparing for the new HKCEE in her two-year research study, and Pan and Newfields (2012; cited by Ha, 2019:6) also discovered only minimal washback of the EFL proficiency Graduation Requirements (henceforth EGR) on university students' learning in Taiwan. Their data showed that the test requirements "did not lead to a noteworthy amount of studying for the test". Finally, the variety of

the activities reported to be used in class, which includes more exam-preparation focused activities, but also authentic material, and activities to work on grammar and vocabulary, but also activities to practise the different skills goes against the information cited by Tsagari (2006:295), which portrayed both teachers and learners as *textbook-bound*. It also goes against the idea expressed by Shepard (1984; cited by Tsagari, 2006:351) that multiple-choice, large-scale standardized tests have mainly negative influences on the quality of teaching and learning although it is true that B2 First tasks are not limited to multiple-choice.

To conclude, the results discussed above show that students are aware of their abilities and can direct their learning towards the aspects that they find more challenging. This could be a positive consequence of the familiarity with evaluation and assessment, as 95% of CEALM and 96% of CEB students stated that they are familiar with the objectives of their courses and 86% of CEALM and 94% of CEB students are familiar with the assessment criteria and methods used in the course and, according to Muñoz and Álvarez (2010:48), this knowledge helps students establish a direction for learning, become critical of their own progress, perform better, and develop their ability to self-assess. A similar view is expressed by Hughes (1989:46 in Muñoz and Álvarez, 2010:38), who points out that "the rationale for (a) test, its specifications, and sample items should be made available to everyone concerned with preparation for the test". Nevertheless, it cannot be concluded that preparing for B2 First exam made students more autonomous and independent learners since the percentage of students in both institutions were already very high at the beginning of the project and the increase, when it happened, was very small.

No major differences have been found that could lead us to think that CEB students are better prepared for autonomous learning than CEALM. In fact, CEALM students seem more autonomous as their main point of reference to decide on their preparation and learning is their own experience, while CEB students rely more heavily on their teacher. Finally, if we look at the activities used for exam preparation, while CEALM students tend to use more general resources such as grammar and vocabulary and contact with the language in general, CEB students use exam practice to boost performance in the exam. In this sense, the washback of the accreditation exam is stronger and could be said to be negative on CEB students as far as test preparation is concerned.

6. CONCLUSIONS AND CONCLUDING REMARKS

Exams have traditionally been used as instruments to assess knowledge and for selection purposes, and speaking a foreign language has always been considered as an asset. Nevertheless, in the last decades being able to speak a language such as English has become crucial for people in general to improve academic, professional or even immigration prospects, and the role of exams to prove or certify this ability to communicate in a foreign language has rocketed. These factors have contributed to the interest in testing and assessment and the consequences they have for the individuals involved.

The research into washback has focused mainly on contexts in which a new exam was introduced and teachers were the main subjects under study. This research line has enhanced the understanding of washback and has raised awareness of the potential consequences that exams may have in the learning and teaching context. Nevertheless, scholars agree that it is time to look into the washback that could take place in contexts in which an exam has been used for some time and to analyse the role played by students. More precisely, scholars point at the need to analyse aspects such as a score gain, students' attitude towards learning and exam success, and their experience in exams and exam preparation among others. Given the complexity of the phenomenon, most of the studies on the subject have offered limited findings, which has led to the belief that future research should collect empirical data through a range of instruments.

This research project aimed to have a better understanding of how preparing for Cambridge Assessment English B2 First exam may affect language learners. More precisely, it aimed to answer three research questions. The first one looked at whether students attending a more exam-oriented (CEB) course showed a better performance in B2 First exam than students attending a more general English (CEALM) course. The second research question looked at whether students attending a more exam-oriented course improved their language knowledge and abilities. The third question analysed students' autonomy and independence when learning and whether they increased as a result of preparing for B2 First exam.

In order to answer these three research questions, the project analysed data from 130 students and 8 teachers from two different institutions in Jaén (Spain) using qualitative and quantitative methods. The data collection instruments included four sets of exams, two B2 First

mock exams and two grammar and vocabulary tests applied at the beginning and at the end of the project to measure students' progress. These results were contextualised using the information of two sets of questionnaires for students, which were applied at the beginning and at the end of study. To have a more complete perspective of the situation under study, the teachers answered a questionnaire, and a sample of their lessons was observed.

The quantitative data obtained from the exams, the questionnaires and the classroom observation were studied by means of statistical analyses, and contextualised using the qualitative data obtained from the questionnaires. The results were discussed considering relevant and recent research into washback.

In the light of this, the following conclusions can be drawn, which answer the three research questions:

1.- Do students enrolled on more exam-oriented (CEB) courses show a better performance in Cambridge B2 First mock exam than those enrolled on general English (CEALM) courses?

Students attending more exam-oriented courses show a stronger overall performance in Cambridge B2 First mock exam than students attending a general English course. This finding is in line with some of the literature reviewed while it contrasts with most of the studies on the subject, which did not report statistically significant differences between the experimental and control groups.

If performance in each skill is considered independently, at the end of the project, the experimental group is stronger at speaking, listening, and at the knowledge of grammar and vocabulary tested in the Use of English component. However, no statistical differences could be found when comparing writing and reading skills. Such findings could be explained by the different methodologies followed in each institution, the time devoted to each skill in class, and the material and resources used.

2.- Do students enrolled on more exam-oriented (CEB) courses improve their language knowledge and abilities?

Students enrolled on a more exam-oriented course do not show an improvement in their language knowledge and ability if the results in the independent tests are considered. However, it is necessary to consider the data commented on in the discussion and the caveats, some of which will be discussed in the next few lines, to fully understand this conclusion.

3.- Do students become more autonomous and independent learners as a result of preparing for Cambridge B2 First exam?

Preparing for B2 First exam has only a limited effect on students' ability to know how to improve their English and on their awareness of their strengths and weaknesses. It must be considered that the results in this sense were already extremely positive at the beginning of the study and only increased slightly in some situations at the end of the project. Such positive findings at the end of the study contrast with the literature reviewed.

When considering students' reference for learning and exam preparation, it can be concluded that the students enrolled on general English courses show more independent and autonomous approaches to learning than those students enrolled on more exam-oriented courses because students in more general English courses rely mainly on their own personal experience and reflection while participants attending more exam-oriented courses identify their teacher as their main reference.

If the ability to adapt their learning to different situations and exams is considered, the findings obtained show that most of the students enrolled on more exam-oriented courses adapt their studying habits, hence showing a more independent and autonomous behaviour than the participants from general English courses, where it is a minority of students.

In addition to the conclusions mentioned above as related to the initial research questions, this project has offered interesting findings:

1) The results obtained from comparing Year 1 and Year 2 students enrolled on more exam-oriented courses support the idea that learners need more than 120 teaching hours to move from one CEFR level to the next.

2) Exam preparation seems especially relevant for the Use of English component because although the control group showed a better performance in the independent tests, the experimental group outperformed them in the Use of English test.

3) The washback of B2 First is evidenced by the fact that most of the students preparing for it adapted their studying habits. The washback on motivation can be said to be positive because it encouraged students to work harder and a majority of the students maintained a positive perspective towards the exam even when they faced negative results. However, the washback on

learning habits can be perceived as negative for students attending more exam-focused courses because their preferred activity was to do more exam practice. This finding agrees with the literature review.

4) The activities used in the two types of courses are very similar with the top most frequently practised activities being the same. This suggests, on the one hand, that students are more exam-oriented than their teachers and, on the other hand, that the washback on classroom practice is very limited and it is not negative.

Finally, the scope of this research project is limited for several reasons. (i) The results cannot be used to make general assumptions about the washback effect of B2 First and more exam-focused courses because the participants belonged just to two institutions, both of them in Jaén. (ii) The point in time at which the study was conducted together with the number of tests that students had to do posed obstacles in terms of students' attention and commitment, which affected the different tests in different ways. Students at CEB were mainly focused on the preparation for B2 First exam and their end-of-course exam, and usually did the independent tests, which were low stakes for them, rather quickly. Similarly, students at CEALM considered that the B2 First mock exams were above their level and low stakes for them and did them quite quickly. Moreover, (iii) the lack of statistically significant differences between Entry tests and End-of-course tests could suggest that the time between them was too little for students to make significant improvements.

In the light of the conclusions obtained and limitations observed in the present study, future research on the matter is considered necessary. It would be useful to conduct a larger and broader research project with more students and institutions involved to have a representative sample that could allow the author to generalise the findings. This is considered vital to minimise construct irrelevant factors and to draw more solid conclusions about the washback of Cambridge B2 First exam and more exam-oriented courses, which could help make informed decisions in the teaching and learning of English. Similarly, it would be desirable to analyse performance over a longer period of time to obtain more solid evidence of significant improvements in performance. In doing that, the time factor would not seem so relevant and it is possible that clearer patterns could be observed. From a slightly different angle, another potential research line could be to look into the role that teachers have and the potential influence they may exert on their students to be able to

decide to what extent the washback to the learner is due to the exam and to what extent the teacher's influence impacts on students' learning practice and test-taking strategies and even their motivation. Students and teachers usually share a special relationship based on trust and support and it is interesting to analyse the teacher factor but doing so as one more variable to understand the effect of exams on students.

The idea of conducting this research project was conceived in the classroom after having taught students for a number of years and after having carried out a more limited research project on washback in the framework of a Master's Thesis. Teachers' main concern is to help their students to improve and learn in the best possible manner and, as a professional, it became necessary to have a reflective attitude and obtain evidence about the learning process for students preparing for B2 First. The process of designing and carrying out the project has been enriching and has fostered a reflective attitude. Reading about different studies and the views shared connects professionals from very different contexts and has encouraged the author to take nothing for granted. The results obtained and their limitations support the idea about the complexity of the phenomenon under study but are also a useful contribution in terms of students' progression and autonomy based on empirical research. When analysed in the light of the research available, in some cases it is in line and hence supports some views expressed by relevant authors and in some other cases offers an interesting contrast, which can inform future research.

7. Sections in Spanish

7.1 Título

EL EFECTO DE LOS EXÁMENES EN LOS ESTUDIANTES: ¿SON LOS CURSOS ORIENTADOS A LA PREPARACIÓN DE EXÁMENES REALMENTE EFECTIVOS?

7.2 Tabla de contenidos

7.3. Introducción

Cuéntame algo y lo olvidaré. Enséñame y lo recordaré. Hazme partícipe de ello y aprenderé.[8]

(Benjamin Franklin)

7.3.1. Aprender

El aprendizaje es fundamental para el desarrollo y es la base de la sociedad y el progreso. Aprender, y aprender de la mejor manera, es el pilar de esta Tesis Doctoral. Pero, ¿a qué nos referimos exactamente cuando hablamos de aprendizaje?

El aprendizaje cambia la forma en la que vemos el mundo. Este cambio puede producirse a nivel del conocimiento, de la actitud o del comportamiento (Queen's University, n.d.). El aprendizaje es un proceso *activo* porque los estudiantes han de explorar el mundo que los rodea, observarlo e interactuar con todo lo que sucede a su alrededor, manipular objetos, participar en conversaciones y relacionarse con otros individuos, dado su componente social. El mayor o menor éxito de este proceso de aprendizaje se basa en tres factores fundamentales: los procesos cognitivos, la atención y participación, y los comportamientos de aprendizaje (Knight, 2020).

Los *aspectos cognitivos del aprendizaje* están relacionados con el pensamiento y los procesos mentales y son más profundos que memorizar o recordar información. De hecho, tienen que ver con el *aprendizaje constructivo,* es decir, con construir conocimiento y desarrollar destrezas basadas en lo que ya sabemos, lo que se considera la base de todo aprendizaje futuro, así como con ser capaces de pensar de forma independiente y crítica para transferir conocimiento a contextos nuevos y diferentes (Queen's University, n.d.). Además, hay una serie de factores que influyen en el aprendizaje y que son cruciales para la enseñanza. Entre ellos encontramos la *carga cognitiva*, es decir, la cantidad de información nueva que los estudiantes son capaces de procesar; la *consolidación*, que incluye técnicas como la repetición espaciada, la elaboración, el ensayo y la capacidad de relacionar la nueva información con la experiencia personal; y *la presentación de la información de manera dual*, lo que refuerza la capacidad de recordar al presentarse la información de forma visual y verbal.

[8] Las traducciones de las citas que aparecen en esta sección las ha realizado la autora.

El carácter activo del aprendizaje hace que la *atención* y la *participación* sean ingredientes básicos, y para sacarles el máximo partido debemos tener en cuenta una serie de aspectos fundamentales como es la expectativa de conseguir un objetivo, concepto al que aluden Knight (2020) y Svinicki (2004). Los estudiantes prestan más atención y se muestran más participativos cuando creen que pueden realizar satisfactoriamente una tarea que es *relevante* para ellos porque les interesa. En este sentido, resulta fundamental encontrar el equilibrio entre el reto que supone la tarea y la creencia de que se trata de un reto asequible. Por otro lado, los estudiantes aprenden mejor cuando sienten que tienen el control de lo que aprenden, de cuándo lo aprenden y de cómo lo aprenden. Knight (2020) también se refiere a esta capacidad de dirigir el propio aprendizaje como un elemento clave. El tercer aspecto fundamental es el *componente social* del aprendizaje, que refuerza la voluntad de los estudiantes para prestar atención y participar activamente en el aprendizaje, es decir, si el estudiante se siente cómodo interactuando con otros estudiantes, probando recursos lingüísticos y evaluando a sus compañeros, aprenderá mejor. Por último, *disfrutar* y *sentir curiosidad* son, sin duda, elementos fundamentales a la hora de aprender porque influyen en la dirección, intensidad, persistencia y calidad de los comportamientos de aprendizaje que los alumnos emplean (Ambrose et al., 2010:69).

Finalmente, debemos tener en cuenta que el aprendizaje no es algo que los estudiantes reciben, sino que es algo que los estudiantes consiguen, de ahí la importancia de los *comportamientos de aprendizaje. Marcarse objetivos* es vital puesto que se ha demostrado una y otra vez que los estudiantes consiguen más cuando se centran en un objetivo específico que es alcanzable (Crooks, 1988; citado por Green, 2007:23) y trabajan para conseguirlo. Además, si los estudiantes perciben que el objetivo es valioso, estarán más dispuestos a dedicar tiempo y esfuerzo. Para ello deberán poner en práctica una serie de estrategias específicas, lo que Knight (2020) llama *gestión del aprendizaje*, y reflexionar sobre ellas, evaluando lo que están haciendo bien y los aspectos que deben mejorar.

El aprendizaje es la base del desarrollo y la evolución y lo es incluso más en el mundo en el que vivimos, que se encuentra en continuo cambio. Uno de los principales desafíos a los que se enfrentan el aprendizaje y la enseñanza en la actualidad es hacer que el aprendizaje a lo largo de toda la vida sea una realidad.

7.3.2. El aprendizaje a lo largo de toda la vida, las lenguas y el inglés como *lingua franca*

Organizaciones e instituciones de reconocido prestigio como las Naciones Unidas o la Comisión Europea consideran el aprendizaje a lo largo de toda la vida como un componente clave para el desarrollo cultural, económico y medioambiental. Las Naciones Unidas, por ejemplo, afirman que el aprendizaje a lo largo de toda la vida "puede ayudar a erradicar la pobreza, proteger el planeta, defender los derechos humanos, construir sociedades igualitarias, inclusivas en las que sus individuos vivan en paz, así como promover el progreso social, económico, cultural y tecnológico" (UNESCO, 2016). De la misma manera, la Comisión Europea ha elaborado una lista con ocho competencias fundamentales que se enmarcan en el aprendizaje a lo largo de toda la vida y que preparan a los ciudadanos para desenvolverse en la sociedad actual (European Commission, 2017). Una de estas competencias fundamentales tiene que ver con mejorar el conocimiento de idiomas y el número de lenguas que se aprenden. El Consejo Europeo del 14 de diciembre de 2017 señaló en sus conclusiones que las lenguas juegan un papel social, ya que permiten a las personas participar de forma más activa como ciudadanos y les ayudan a estar más y mejor preparados para lidiar con los desafíos de las sociedades actuales, que son multilingües y diversas. Asimismo, las lenguas refuerzan el entendimiento entre culturas y la paz porque unen a las personas y hacen a países y culturas extranjeras accesibles. Desde un punto de vista económico, la capacidad de hablar un idioma extranjero mejora las perspectivas de movilidad y empleo, por lo que aumenta la competitividad de los países.

El inglés se ha posicionado como una *lingua franca* gracias a la globalización. Cada vez más personas aprenden inglés porque es algo muy demandado en el entorno laboral, ya sea para encontrar trabajo o para mejorar las condiciones laborales (Chávez Zambano, Saltos Vivas & Saltos Dueñas, 2017:761). La importancia del inglés es tal que aprender y hablar inglés ya no se considera un lujo sino una necesidad en cualquier parte del mundo o cualquiera que sea nuestro ámbito profesional (Jaimechango, 2009; citado por Chávez Zambano, Saltos Vivas & Saltos Dueñas, 2017:761).

En vista de la importancia de las lenguas en general y del inglés en particular, la enseñanza, el aprendizaje y la evaluación de las lenguas han suscitado un gran interés. Actualmente, existen proyectos de colaboración entre gobiernos y expertos de reconocido prestigio en este ámbito para modernizar la enseñanza y hacerla más eficiente a través del desarrollo de métodos innovadores y

promoviendo una metodología de evaluación común porque se ha consolidado la idea de que enseñar y evaluar idiomas deben ir de la mano. Como ejemplo podemos mencionar la iniciativa RELANG (por sus siglas en inglés, *Relating Language Curricula, tests and examinations to the Common European Framework of Reference for Languages*) que conecta el currículo lingüístico y la evaluación al *Marco Común Europeo de Referencia para las Lenguas* y que permite a las autoridades educativas alinear los exámenes de idiomas con los niveles de competencia del *Marco Común Europeo de Referencia para las Lenguas* (European Commission, n. d.).

7.3.3. El aprendizaje y la evaluación

El papel de la evaluación y de los exámenes ha evolucionado con el tiempo. En algunos casos se han usado como instrumento de poder y de control. En la China Imperial, hace más de mil años, el proceso de selección de los oficiales de más alto rango se realizaba mediante lo que probablemente fueran las primeras oposiciones jamás celebradas (Lai, 1970; Hu, 1984; Arnove, Altback, & Kelly, 1992). De esta manera, las autoridades estaban seleccionando a sus trabajadores y a su vez estaban estableciendo y controlando el sistema educativo a través de un examen de gran relevancia que sería determinante no solo en las vidas de los candidatos sino también en el futuro del Imperio (Spolsky, 1995a, 1995b). Sin embargo, las primeras nociones de los principios de la evaluación ya estaban presentes en estos exámenes ya que se implementaron algunas medidas para producir un examen justo y para evitar la corrupción. Este uso de los exámenes como forma de evitar la corrupción y promover la excelencia y el talento también ha sido documentada por Eckstein y Noah (1992), así como por Bray y Steward (1998) e inspira el uso actual de los exámenes, de manera que es tan poderoso hoy como lo era hace siglos.

Los exámenes también se han utilizado como instrumentos de reforma educativa (Linn, 2000, citada por Cheng, Watanabe & Curtis, 2008:6) y este uso ha conllevado un cambio en la dirección del aprendizaje, la enseñanza y la evaluación. Tradicionalmente, los exámenes se realizaban al final del proceso de enseñanza y aprendizaje. Sin embargo, algunos tienen tal importancia para los candidatos que influyen en las actitudes, comportamientos y motivación de los profesores, los alumnos y los padres, de manera que el efecto del examen se percibe hacia atrás, de ahí que en inglés se utilice el término *washback* (Pearson, 1988:98; citado por Cheng, Watanabe & Curtis, 2008:7) que incluye en la misma formación de la palabra la idea de que el efecto precede al examen. Sin embargo, al igual que Davies (1985), Pearson pensaba que el efecto

de los exámenes se siente en realidad a posteriori, ya que cuando se introduce un nuevo examen cuyos resultados son transcendentales para los candidatos, los materiales de formación se alinean con el nuevo examen y tanto profesores como alumnos se adaptan a él, a menudo esforzándose más para conseguir mejores calificaciones (Cheng, Watanabe & Curtis, 2008:12). El poder para transformar y cambiar las actitudes y comportamientos ha sido el principal tema de la investigación sobre el efecto de los exámenes. Como ejemplo se puede citar a Pearson (1988), Shohamy (1992), Cheng (1997), Andrews (2002), Read & Hayes (2003) o a Saif (2006), que han sido mencionados por Tsagari (2007:13). Algunos académicos han visto en los exámenes y en su efecto una oportunidad para innovar y favorecer la evolución de la enseñanza y el aprendizaje. Otros, sin embargo, lo han visto como una forma de empobrecer el aprendizaje pues perciben que tanto la enseñanza como el aprendizaje se basan principalmente en la realización de exámenes de muestra (Davies, 1968:125) y piensan que las clases de preparación se limitan a entrenar a los alumnos para los exámenes en lugar de enseñarles contenidos relevantes, que, en el caso que nos ocupa, son los idiomas (Wiseman, 1961:159; citado por Cheng, Watanabe & Curtis, 2008:9).

7.3.4. La evaluación como una oportunidad para aprender

> Los exámenes se han visto como un mal recurrente y necesario, una dosis de medicina desagradable, cuyo sabor hay que eliminar tan pronto como sea posible. (Cheng, Watanabe & Curtis, 2008:14)

El valor de los exámenes con fines de selección o como motores de innovación se encuentra arraigado y su uso en el día a día del aula no es nada nuevo. Sin embargo, la utilidad de los exámenes en el aula parece no estar clara si tenemos en cuenta que los profesores suelen quejarse de que su elevado número les deja poco tiempo para enseñar y que los alumnos ven los exámenes como una forma de control hasta el punto de que lo único que les preocupa es la calificación que obtendrán. Ante esta situación, no sorprende que la evaluación no goce de una gran popularidad en este contexto. Resulta evidente que si los exámenes se emplean como meras instantáneas de las habilidades y conocimiento de los alumnos reflejadas en una calificación, se podría estar de acuerdo con que hay demasiados exámenes.

Afortunadamente, la evaluación puede ofrecer información de interés sobre los alumnos y para los alumnos y ayudar a tomar decisiones informadas en el ámbito de la enseñanza y el

aprendizaje. No obstante, para que esto sea posible la evaluación debe estar conectada con los principios del aprendizaje y estar plenamente integrada en él. En la enseñanza de idiomas, el *Marco Común Europeo de Referencia para las lenguas* (MCER) se ha convertido en *el* documento de referencia. Sus declaraciones de capacidad lingüística (*Can-Do Statements*) para cada destreza y situación comunicativa son fácilmente adaptables para crear objetivos de aprendizaje acordes a cada nivel. El hecho de que se encuentren formulados en positivo, subrayando lo que los estudiantes son capaces de hacer en cada nivel, es una forma motivadora de diseñar los distintos peldaños que los estudiantes de idiomas deben subir. Al alinear los objetivos educativos con el MCER, las instituciones crean un itinerario de aprendizaje más coherente y contribuyen a la internacionalización de la enseñanza y el aprendizaje. Una vez que los objetivos de aprendizaje están claros, hay que conectarlos con la práctica y las actividades del aula explicando esta conexión para que los estudiantes entiendan la utilidad de las actividades ya que de esta manera sentirán que tienen un mayor control de su aprendizaje, lo que a su vez, aumentará su motivación. Además, conectar los objetivos de aprendizaje con la evaluación familiarizando a los estudiantes con los criterios y métodos de evaluación hace que el proceso de enseñanza y aprendizaje sea más transparente. Esto se debe por una parte a que los alumnos saben cómo se les va a evaluar mientras, por otra parte, la calificación tendrá un significado mayor que el mero aprobado o suspenso. La información obtenida de esa evaluación permitirá a los alumnos identificar sus debilidades y los aspectos que requieren un mayor esfuerzo. No obstante, también les hará ver lo que están haciendo bien y supondrá una fuente de motivación para seguir trabajando. Para que la evaluación pueda ofrecer dicha información, los docentes tendrán que realizar una reflexión meticulosa empleando los criterios de evaluación, lo que les permitirá tener un conocimiento más objetivo del rendimiento de sus alumnos y les ayudará a realizar una programación más fiable, con unas etapas y objetivos más realistas.

La conexión entre aprendizaje y la evaluación que se produce en el aula y que realizan los docentes es probablemente más fácil de entender que la que se produce entre el aprendizaje y la evaluación sumativa, que a menudo sucede al final del curso, ya que ésta última se percibe como algo que marca el final de una etapa. Esta conexión es todavía más difícil de percibir cuando la evaluación la realiza una entidad evaluadora externa, que no conoce a los alumnos, sus circunstancias o sus objetivos vitales. Sin embargo, la equivalencia entre los exámenes de certificación y el MCER, la importancia que tiene el aprendizaje a lo largo de toda la vida, y el

interés que muestran las entidades evaluadoras por el efecto y el impacto de los exámenes en las personas, la educación y la sociedad en general ha reforzado la conexión entre la evaluación sumativa y el aprendizaje.

El hecho de que los exámenes de certificación se hayan alineado con el MCER y con los principios de la evaluación de idiomas ha aumentado la transparencia de los métodos y los instrumentos de evaluación, así como de las calificaciones, y ha permitido que las titulaciones se internacionalicen. Añadir un nivel del MCER a cualquier calificación le otorga automáticamente el contexto necesario para interpretar los resultados y tomar decisiones en función de lo que una persona es capaz de hacer y las situaciones en las que se puede desenvolver con sus habilidades lingüísticas. Asimismo, ha abierto un sinfín de oportunidades para las universidades y programas de internacionalización como el programa Erasmus, facilitando la admisión de estudiantes extranjeros. En el entorno profesional, también ha facilitado los procesos de selección y lo mismo se puede decir en los casos en los que hablar un idioma es un factor determinante para migrar.

No obstante, estas posibilidades se basan en la premisa de que los resultados obtenidos por los candidatos cumplan con todos los requisitos, es decir, que sean válidos, fiables y justos. Las instituciones evaluadoras se encuentran en constante evolución para diseñar exámenes basados en la investigación y que se ajusten a los principios de aprendizaje y a las tendencias más novedosas en evaluación, ya que son conscientes de la relevancia que tienen sus exámenes debido al reconocimiento del que gozan. Asimismo, existen asociaciones como Ofqual y ALTE que garantizan que los exámenes de idiomas cumplen con los necesarios estándares de calidad para que sean justos, fiables, prácticos y válidos, para que produzcan el mayor impacto positivo posible tanto a nivel individual como en la sociedad en general, reduciendo asimismo los efectos no deseados.

La relevancia del aprendizaje a lo largo de toda la vida, el hecho de que nunca dejamos de aprender un idioma junto con el poder de los exámenes ha hecho que las pruebas de acreditación tengan un nuevo papel y que ya no se las considere como la última etapa del viaje lingüístico sino como un punto y seguido, como un nuevo hito alcanzado. Los exámenes de acreditación son herramientas de análisis exhaustivas y deben serlo porque sus resultados se emplean para tomar decisiones que tienen la capacidad de cambiar la vida de las personas (Raban, 2008:x y University of Cambridge Local Examinations Syndicate, 2016). Por ello, resulta útil aprovechar todos los datos que los exámenes recogen de los candidatos para ayudarlos a entender lo que saben hacer bien y

las dificultades que tienen, es decir, para que conozcan sus habilidades y para que utilicen ese conocimiento para seguir aprendiendo. Esto es posible si la calificación obtenida viene acompañada de una descripción de lo que los estudiantes son capaces de hacer en cada nivel. Con esto volvemos a aludir a la conexión entre la calificación obtenida, lo que representa en términos de habilidades reales y su influencia en el aprendizaje. Probablemente esta conexión sea un ejemplo del efecto positivo de los exámenes.

7.3.5. Es complicado

> Cito esto como un ejemplo de la importancia que tiene investigar las creencias personales en lugar de aceptar lo que parece ser verdad (Cheng, Watanabe & Curtis, 2008:x)

La sección anterior acababa con la palabra probablemente como limitador del efecto positivo que puede tener la conexión entre exámenes, habilidades lingüísticas reales y aprendizaje. Aunque uno podría pensar que no hay inconvenientes en esta conexión, autores de reconocido prestigio como Cheng, Watanabe y Curtis (2008:11) sugieren que probablemente sea más seguro añadir esta limitación especialmente si tenemos en cuenta que la naturaleza positiva o negativa del efecto de los exámenes puede verse influenciada por muchos factores, entre los que los autores citan los *factores relacionados con el examen,* como los métodos de evaluación, el contenido, las destrezas evaluadas, el objetivo del examen, y las decisiones que se toman en función de sus resultados. Asimismo, algunos exámenes y entidades evaluadoras tienen un cierto prestigio dentro del sistema educativo, lo que junto con la relevancia del examen puede influir en el efecto que éste tiene. Por otra parte, el *contexto de aprendizaje*, ya sea el *micro-contexto*, es decir, el colegio o el *macro-contexto*, es decir, la sociedad, la ciudad o la región donde el examen se usa, también influye en el efecto del examen.

Los principales actores en el proceso de enseñanza y aprendizaje son los profesores y los alumnos. La formación, las creencias sobre evaluación, enseñanza y aprendizaje de los docentes, así como la percepción personal que el profesor tiene de un examen en concreto influyen también en el efecto que provocan los exámenes. Por último, los estudiantes han recibido poca atención en la investigación, a pesar de la creencia de autores fundamentales como Green (2007:314) de que la forma en la que un alumno responde a las demandas de un examen y a otros aspectos del

aprendizaje es más determinante para el aprendizaje que las clases a las que asiste o los materiales que utiliza. Algunos estudios han tratado de conocer al estudiante a través de sus profesores. Sin embargo, se ha probado que la forma y la intensidad en la que los alumnos sienten el efecto de los exámenes no tiene por qué equivaler a cómo los profesores sienten ese efecto.

La complejidad del efecto de los exámenes se ha puesto de manifiesto en el trabajo de expertos fundamentales en la materia. Alderson y Wall (1993), que marcaron el camino de la investigación con su trabajo en el que se preguntaban si el efecto de los exámenes realmente existía y enumeraban una series de hipótesis, llegaron a la conclusión de que se tenía que investigar más sobre el tema. Las preguntas que han de responderse ahora tienen que ver con conocer mejor el efecto de los exámenes, lo que lo produce y por qué (Cheng, Watanabe & Curtis, 2008:ix). Por su parte, Alderson y Wall señalaron en 1993 (citados por Cheng, Watanabe & Curtis, 2008:12) dos líneas de investigación: por un lado, estudiar el papel de los exámenes como un motor de cambio y motivación – lo que ha recibido una atención considerable – y por otro lado, indagar en la motivación y el rendimiento.

7.3.6. Qué esperar

> El efecto de los exámenes se encuentra en la relación entre la preparación para superar con éxito el examen y la preparación para tener éxito más allá del examen, en el ámbito que el examen quiere reflejar y cuyo acceso puede controlar. (Green, 2007:1)

La complejidad del efecto de los exámenes y de los factores específicos que influyen en él explican la importancia de que los estudios que se realicen tengan un contexto y objetivos muy bien definidos, en los que nada se dé por sentado. El motor de esta Tesis Doctoral es entender la forma en la que los profesionales de la enseñanza y de la evaluación pueden ayudar a los estudiantes a aprender mejor y también dar a los alumnos las herramientas para que tengan un papel más activo y eficaz en su propio aprendizaje. Este estudio está basado en la experiencia reciente de la autora, que ha trabajado enseñando inglés y preparando a sus alumnos para las titulaciones de Cambridge English y en el conocimiento adquirido con la elaboración de su Trabajo Fin de Máster en 2015 (Peña Jaenes, 2015), que trató el efecto de los exámenes desde la perspectiva de los cursos de idiomas y de la evaluación de la expresión escrita. Este proyecto de investigación se centra en una

de las principales líneas de investigación identificadas por Alderson y Wall (1993): los estudiantes de idiomas y analiza su rendimiento y motivación. Su objetivo es poder entender mejor el progreso de los alumnos gracias a datos empíricos y dar respuesta a tres preguntas de investigación:

i) ¿Demuestran los alumnos matriculados en cursos orientados a la preparación de exámenes (Centro de Estudios Británicos, CEB) un mejor rendimiento en la prueba de B2 First que los alumnos matriculados en cursos de inglés general (Centro de Estudios Avanzados en Lenguas Modernas, CEALM)?

ii) ¿Mejoran sus conocimiento lingüístico y sus destrezas los alumnos matriculados en cursos orientados a la preparación de exámenes (CEB)?

iii) ¿Aumenta la autonomía e independencia de los alumnos gracias a la preparación del examen de Cambridge B2 First?

El principal objetivo de esta Tesis y las tres preguntas de investigación han guiado el estudio y han permitido obtener conclusiones basadas en los datos que, aunque sorprendentes a veces – lo que viene a apoyar la cita de Cheng, Watanabe y Curtis (2008:x) – permiten a la autora tener un mayor conocimiento del efecto de los exámenes, del rendimiento de los estudiantes y de su motivación. Asimismo, también aportan luz a la hora de entender el equilibrio entre preparar para el éxito en un examen y preparar para el éxito más allá del mismo, lo que debería ser el objetivo último de todo curso de idiomas. Como creo que sucede con la mayoría de los estudios, nos abren los ojos a nuevas líneas de investigación. Este proyecto ha animado a la autora a continuar estudiando y trabajando en el ámbito de la evaluación. Este aprendizaje se narra en este proyecto, y su estructura se describe en las siguientes líneas:

El capítulo que sigue a la Introducción es la Revisión del Estado de la Cuestión en la materia. Comienza explorando la evaluación, su evolución y sus tipos y objetivos, así como las cualidades que todo examen debería tener. El capítulo presta especial atención a las cuatro destrezas, de una perspectiva más teórica a un enfoque más práctico cuando analiza sus implicaciones para la evaluación. La segunda parte de la Revisión del Estado de la Cuestión se centra en el efecto de los exámenes y describe y comenta la investigación realizada hasta la fecha para intentar encontrar lagunas en las que la contribución de esta Tesis pueda resultar de utilidad y estudia cómo la

percepción del efecto que provocan los exámenes y su definición han evolucionado hasta tener un conocimiento más maduro del término. Se habla de la complejidad del efecto y de la dificultad a la hora de identificar lo que lo causa y los aspectos que influyen en él. La tercera parte se centra en los exámenes de acreditación en general como introducción al examen que está en el centro de esta investigación: el examen de Cambridge B2 First. En esta sección se describe la institución que desarrolla el examen, la evolución del examen y su diseño actual, ya que todos estos aspectos son fundamentales a la hora de entender el valor que tiene la prueba.

Antes de pasar a hablar de la metodología del estudio, y basándose en la literatura revisada, se justifica la relevancia de la investigación y se identifican las lagunas que se han encontrado en los estudios consultados así como el interés académico y personal de la autora, lo que enlaza con el principal objetivo del proyecto, sus objetivos específicos y las preguntas de investigación. Estos aspectos determinan la metodología seguida y su descripción conforma el tercer capítulo del proyecto.

El hecho de que la fase de recogida de datos se realizara cuando la autora trabajaba como profesora y los objetivos del proyecto fueron claves a la hora de decidir realizar investigación basada en el aula. La literatura consultada y, una vez más, el objetivo – conseguir un mayor conocimiento basado en datos – requería el uso de una amplia gama de herramientas y métodos de recogida de datos, como la observación, los cuestionarios para alumnos y profesores, y los distintos tipos de pruebas para conseguir datos sobre el rendimiento de los alumnos – datos cuantitativos – y también información para contextualizar y explicar los hallazgos – datos cualitativos. De esta manera, la autora trataba de solucionar uno de los problemas encontrados en su anterior proyecto de investigación – aunque con seguridad nuevos aparecerían – e intentaba crear varias vías para canalizar los distintos tipos de información. Este capítulo ofrece una descripción detallada de los instrumentos utilizados, de los procesos de validación empleados y de las medidas de control de calidad implementadas para aumentar la validez y utilidad de los resultados. Los participantes, los centros y la ciudad donde el estudio se realizó se describen, ya que es un aspecto fundamental para contextualizar el efecto de los exámenes y los resultados de la investigación. Por último, se analiza el proceso de recogida de datos y las variables que vertebran la Discusión y los Resultados.

El quinto capítulo del proyecto describe los resultados obtenidos de los análisis estadísticos y los analiza y los comenta a la luz de los datos cualitativos y cuantitativos obtenidos de los

cuestionarios, la observación de las clases y de los estudios más relevantes en la materia. La descripción, análisis y comentarios de los resultados intentan responder a las preguntas de investigación y explicar la posible contribución de este proyecto.

El último capítulo está dedicado a las conclusiones obtenidas del análisis de los resultados, que dan respuesta a las tres preguntas de investigación. Asimismo, también destaca algunos hallazgos que, a pesar de no estar directamente relacionados con las preguntas de investigación, podrían considerarse de interés o de relevancia. El capítulo finaliza con las principales limitaciones encontradas durante el proyecto, que en muchos casos inspira a la autora para realizar investigaciones futuras para intentar sortearlas y continuar estudiando el tema.

A continuación, algunas elecciones lingüísticas y convenciones se explicarán para mayor claridad. El principal objetivo de este proyecto de investigación es el efecto que los exámenes tienen en los alumnos, este efecto recibe el nombre en inglés de *washback* o *backwash* aunque, como veremos, existen otros términos que algunos autores emplean. Si bien ambos términos son aceptados, el término *washback* se ha empleado en este proyecto, principalmente por la influencia de Alderson y Wall (1993), pues lo usaron en su trabajo "Does Washback Exist?", que marcó el inicio de la investigación en el tema; y por la influencia de Anthony Green (2007), que también utiliza el término *washback* en su investigación. El mismo Alderson explica su elección y la relación entre los dos términos en Cheng, Watanabe y Curtis (2008:xii) de la siguiente manera:

> Si puedo permitirme el lujo de hacer una nota a pie de página, en referencia al uso de dos términos que se refieren al mismo fenómeno, *backwash* y *washback*, debería explicar que una de las razones por las que el artículo de Alderson y Wall se tituló "Does Washback Exist?" fue que la palabra *washback* era más frecuente en los debates, presentaciones de congresos y en la formación de profesorado. Cuando yo estudiaba en la Universidad de Edimburgo (Escocia), por ejemplo, Alan Davies, el decano de la evaluación de idiomas en Reino Unido, usaba con frecuencia el término *washback* y no recuerdo que jamás utilizara la palabra "*backwash*". Sin embargo, en la literatura disponible en aquel momento, la palabra "*backwash*" parecía ser prevalente. He aquí una nueva razón para hacernos nuestra pregunta sobre la existencia del efecto de los exámenes llamado *washback*. Para aclarar la diferencia entre los términos *backwash* y *washback*, podemos afirmar que no hay

> diferencia alguna. La única diferencia es que si una persona ha estudiado en la Universidad de Reading (Reino Unido), donde Arthur Hughes enseñaba, entonces probablemente utilizaran el término *backwash*. Si una persona ha estudiado evaluación de idiomas en cualquier otra universidad, pero especialmente en Edimburgo o Lancaster en Reino Unido, entonces casi con toda certeza utilizará el término *washback* (Cheng, Watanabe y Curtis, 2008:xii).

Las personas que participaron en este proyecto estudiaban o trabajaban en dos conocidos centros de idiomas de Jaén (España). Sin embargo, el objetivo de este proyecto no es comparar ambos centros, cuyos directores amablemente permitieron a la autora realizar su investigación, sino entender mejor el progreso de los alumnos que están matriculados en cursos más orientados a la preparación de exámenes y el de los alumnos que están matriculados en cursos de inglés más general. No obstante, por razones prácticas, la autora a menudo se refiere a los primeros como estudiantes de CEB, porque todos los estudiantes del Centro de Estudios Británicos (CEB) estaban matriculados en cursos más orientados a la preparación de exámenes, mientras que se refiere a los segundos como estudiantes del CEALM ya que todos los participantes en esta investigación que estudiaban en el Centro de Estudios Avanzados en Lenguas Modernas (CEALM) de la Universidad de Jaén estaban matriculados en cursos de inglés más general.

La posible contribución de esta Tesis Doctoral es entender el progreso de los estudiantes mediante la obtención de datos. Para ser más precisos, busca medir la eficacia en términos de mejora en la puntuación en las pruebas tipo de B2 First de los cursos más orientados a la preparación del examen de Cambridge B2 First en comparación con los cursos de inglés más general. Asimismo, busca medir la eficacia en términos de mejora de la habilidad lingüística en general de los cursos más orientados a la preparación del examen de Cambridge B2 First en comparación con los cursos de inglés más general. Por último, busca conocer mejor el posible efecto del examen B2 First en la autonomía e independencia de los alumnos y para ello también se basará en datos empíricos.

Por último, las figuras y las tablas están numeradas siguiendo el orden en el que aparecen en el texto. El índice de figuras y el índice de tablas aparecen después de la Tabla de Contenidos para facilitar la lectura. Asimismo, los acrónimos se desarrollan únicamente la primera vez que aparecen en el texto y figuran en el índice de acrónimos para referencia del lector. Finalmente, se ha incluido un índice de apéndices para facilitar su consulta.

7.4. Resumen

Tradicionalmente, los exámenes y la evaluación se han utilizado para seleccionar y para premiar el talento y la excelencia por encima de otros criterios, de modo que, con frecuencia, se han considerado como un punto y final que cierra una determinada etapa. Los exámenes y la evaluación se veían en términos de éxito o fracaso, aprobado o suspenso, lo que limitaba en gran medida su valor y podía generar sentimientos negativos en los individuos implicados: los candidatos, los profesores, las familias, los centros educativos y la sociedad en general.

La investigación en la materia nos permite diferenciar distintos tipos de evaluación, de los cuales los más conocidos probablemente sean la evaluación formativa y la sumativa. En gran medida es posible que nos hayamos reconciliado, al menos en parte, con la evaluación formativa. Este tipo de evaluación la llevan a cabo los profesores en el aula y nos resulta más fácil ver su valor para marcar hitos que se van consiguiendo a más corto plazo y que resultan motivadores para el alumnado y permiten al docente fijar unos objetivos y etapas más realistas. Sin embargo, la evaluación sumativa sigue viéndose con recelo, principalmente cuando la realizan entidades externas, que no conocen las circunstancias de los alumnos, y cuyos resultados solo parecen interesarnos en términos de éxito o fracaso.

En un mundo interconectado como el actual el inglés se ha convertido en un instrumento de comunicación sin fronteras que nos permite disfrutar de la cultura y el entretenimiento, abriéndonos un mundo de oportunidades en ese ámbito pero también en esferas más académicas, así como económicas y en el mundo profesional, sin importar casi a lo que uno se dedique. La capacidad de comunicarse en inglés y poder demostrarlo es fundamental. Tener un título de idiomas reconocido se ha convertido en una tarjeta de embarque que nos permite viajar por todo el mundo ya sea para estudiar o trabajar y la clave está en la palabra *reconocido*. Las instituciones y las empresas buscan evidencias de lo que un futuro alumno o trabajador es capaz de hacer en una situación comunicativa determinada y, por tanto, necesitan tener certeza del significado de un aprobado, una calificación o una puntuación. Nuevamente volvemos a la evaluación y el valor que aporta, pero en este caso hablamos de la evaluación sumativa realizada por organismos independientes y externos que certifican de forma objetiva lo que un candidato es capaz de demostrar. Las entidades evaluadoras cumplen con esa función. De hecho, sus exámenes están basados en la investigación, han superado estrictos controles de calidad tanto internos como

externos, y se encuentran alineados con estándares de calidad internacionales, como el *Marco Común Europeo de Referencia para las Lenguas*, lo que por un lado les aportan un reconocido prestigio y por otro permite a empresas e instituciones conocer de forma certera lo que un futuro empleado o alumno es capaz de hacer. En este punto vemos que la evaluación sigue cumpliendo con esa función de selección y de reconocimiento del talento y la excelencia.

Por otro lado, la sociedad en la que vivimos, en continuo cambio, ha hecho necesario seguir aprendiendo a lo largo de toda la vida. Este aprendizaje no tiene que ser formal o tutelado, puede ser autónomo y basado en el interés o las prioridades. La evaluación sumativa, cuando la calificación viene acompañada de información que el candidato y otras personas interesadas puedan entender, permite guiar este tipo de aprendizaje pues aporta una gran cantidad de datos objetivos que se pueden utilizar para entender lo que uno sabe hacer razonablemente bien y lo que todavía está por conseguir. De esta manera, la evaluación sumativa viene a complementar otros tipos de evaluación y permite seguir marcando hitos en el proceso de aprendizaje que, bien entendido, nunca acaba. Así podemos considerar que la evaluación tiene un efecto positivo como guía y motivación para continuar aprendiendo.

Una vez justificada la relevancia del tema, la literatura en la materia nos ayuda a entender el estado de la cuestión y las posibles oportunidades para contribuir. La investigación en evaluación ha prestado especial atención a las cualidades que deben tener los exámenes para garantizar que sus resultados sean válidos, fiables, prácticos y justos. Sin embargo, la publicación del artículo de Alderson y Wall en 1993 marcó un antes y un después en la investigación sobre el efecto de los exámenes y provocó un creciente interés en los exámenes y cómo afectan a la formación, y más concretamente en cómo afectan a los docentes y el día a día en el aula. Hasta la fecha, la investigación centrada en los estudiantes ha sido escasa y ha aportado hallazgos limitados debido a la complejidad y el carácter impredecible del fenómeno que nos ocupa. En cuanto a los contextos estudiados, si bien hay numerosos ejemplos de proyectos realizados en lugares donde se ha introducido un nuevo examen, el número de estudios en lugares donde un examen estaba bien asentado era más reducido, lo que abría una ventana de oportunidad para realizar una posible contribución. El examen que se incluye en este estudio, B2 First, goza de un amplio reconocimiento y es aceptado por un gran número de empresas e instituciones para certificar el nivel de inglés.

Esta Tesis Doctoral se centra en una de las dos principales líneas de investigación identificadas por Alderson y Wall (1993): los estudiantes de idiomas, y analiza su rendimiento y motivación. Así, el foco principal se proyecta sobre la evaluación y el aprendizaje, y se persigue entender cómo los profesionales de la enseñanza y de la evaluación pueden ayudar a los alumnos a aprender mejor, proporcionándoles las herramientas necesarias para tener un papel más activo y efectivo en su propio aprendizaje. Con esta idea en mente, el objetivo principal es obtener evidencia científica que ayude a comprender mejor el progreso de los alumnos. Para este fin, se plantean tres preguntas básicas de investigación:

1) ¿Demuestran un mejor rendimiento en el examen de Cambridge B2 First los estudiantes matriculados en cursos más orientados a la preparación de examen (CEB) que los matriculados en cursos de inglés general (CEALM)?
2) ¿Mejoran los estudiantes matriculados en cursos más orientados a la preparación de exámenes (CEB) su conocimiento lingüístico y sus habilidades?
3) ¿Aumenta la autonomía y la independencia de los alumnos gracias a la preparación para el examen de Cambridge B2 First?

A través de estas preguntas se estudian las diferencias y similitudes entre dos tipos de cursos de idiomas, los de inglés más general y los de inglés más enfocado a la preparación de un examen. No obstante, el objetivo no es comparar ambos cursos sino entender el efecto que la preparación para un examen y dicho examen de reconocido prestigio pueden tener en los estudiantes de idiomas. Para dar respuesta a las preguntas se ha realizado una investigación basada en el aula, que se ha desarrollado en dos centros de idiomas de Jaén (España) y que ha contado con la participación de 130 estudiantes y ocho profesores. Para la recogida de datos se han empleado pruebas de B2 First y tests de gramática y vocabulario que se han aplicado al principio y al final del proyecto y que previamente habían sido validados estadísticamente. Asimismo, los estudiantes han contestado a dos cuestionarios, uno de entrada y otro de salida. Por último, la autora ha recabado información de la práctica docente gracias a un extenso cuestionario que han respondido los profesores y a la observación de una muestra de clases. De esta manera, se han obtenido datos cuantitativos, que se han analizado estadísticamente, y datos cualitativos que han permitido contextualizar y, en la mayoría de los casos, explicar los resultados obtenidos en los distintos exámenes.

Los resultados se han comentado y analizado teniendo en cuenta un amplio número de estudios en la materia. En algunos casos han ido en consonancia con los hallazgos publicados, con lo que sirven de apoyo y les dan una mayor solidez, mientras que en otros casos han contrastado con la literatura consultada, de manera que dan pie a realizar futuras investigaciones para obtener más evidencia.

7.5. Conclusiones

Los exámenes se han utilizado como instrumentos para evaluar el conocimiento y con fines de selección. Hablar un idioma extranjero siempre se ha considerado como un valor añadido en la formación de una persona. Sin embargo, en las últimas décadas el dominio del inglés como lengua adicional se ha convertido casi en un requisito curricular para mejorar las oportunidades laborales, académicas o de migración. De ahí que el papel de los exámenes para certificar esta habilidad de comunicarse en una lengua extranjera haya despuntado. Todos estos factores han contribuido al interés que la evaluación y sus consecuencias tienen para las personas que participan de una u otra manera en este proceso.

La investigación del efecto que provocan los exámenes se ha centrado principalmente en aquellos contextos en los que se introducía una nueva prueba y los docentes eran los principales sujetos de análisis. Esta línea de investigación ha mejorado nuestro conocimiento del efecto que provocan los exámenes y nos ha hecho más conscientes de sus consecuencias en el aprendizaje y la enseñanza. No obstante, los expertos coinciden en que ha llegado el momento de investigar el efecto que provocan los exámenes en aquellos contextos en los que se han estado utilizando durante cierto tiempo y de analizar el papel que juegan los estudiantes. Para ser más exactos, los autores señalan la necesidad de analizar aspectos como la mejora en las puntuaciones, la actitud de los estudiantes hacia el aprendizaje y el éxito en los exámenes, así como su experiencia a la hora de realizarlos y prepararse para ellos. Dada la complejidad del fenómeno que nos ocupa, la mayoría de los estudios realizados ha ofrecido resultados limitados, lo que ha llevado a la conclusión de que la investigación debe recoger datos empíricos a través de distintos instrumentos.

Este proyecto de investigación se ha realizado para mejorar nuestro conocimiento sobre cómo la preparación para el examen de Cambridge Assessment English B2 First puede afectar a los estudiantes de idiomas, para lo cual trata de dar respuesta a tres preguntas de investigación. La primera se centra en si los estudiantes matriculados en cursos más orientados a la preparación de examen (CEB) tienen un mejor rendimiento en el examen B2 First que los estudiantes que asisten a un curso de inglés más general (CEALM). La segunda pregunta tiene que ver con la posibilidad de que los estudiantes que asisten a cursos más orientados a la preparación de examen (CEB) mejoren su conocimiento lingüístico y sus habilidades. La última pregunta analiza la autonomía y la

independencia de los alumnos e intenta dilucidar si éstas aumentan como consecuencia de la preparación para el examen B2 First.

Para responder a estas tres preguntas de investigación, el proyecto analizó datos de 130 estudiantes y 8 profesores de dos centros de idiomas en Jaén (España) usando métodos cuantitativos y cualitativos. Los instrumentos de recogida de datos incluían cuatro series de exámenes, dos exámenes de prueba de B2 First y dos exámenes de gramática y vocabulario que se aplicaron al comienzo y al final del proyecto para medir el progreso de los estudiantes. Estos resultados se han contextualizado usando la información de dos series de cuestionarios para estudiantes, que también se aplicaron al comienzo y al final del estudio. Además, para tener una perspectiva más completa de la situación de estudio, los profesores respondieron a un cuestionario y se llevó a cabo una observación de una muestra de las clases.

Los datos obtenidos a partir de los exámenes realizados por el alumnado se analizaron estadísticamente, se contextualizaron teniendo en cuenta los resultados de los cuestionarios y los de las observaciones en el aula, y se comentaron y discutieron de acuerdo con la información disponible en la literatura publicada sobre la materia.

Una vez realizado el estudio, podemos extraer las siguientes conclusiones que dan respuesta a las tres preguntas de investigación.

1.- ¿Demuestran un mejor rendimiento en el examen de Cambridge B2 First los estudiantes matriculados en cursos más orientados a la preparación de examen (CEB) que los matriculados en cursos de inglés general (CEALM)?

Los alumnos que asisten a cursos más orientados a la preparación de examen muestran un mejor rendimiento general en la prueba de B2 First que los alumnos matriculados en cursos de inglés más general. Este hallazgo coincide con parte de la literatura revisada pero contrasta con la mayoría de los estudios en la materia pues no ofrecen diferencias estadísticamente significativas entre el grupo experimental y el grupo de control.

Si analizamos el rendimiento en cada destreza de manera independiente, al final del proyecto el grupo experimental presenta niveles significativamente más altos en expresión oral, comprensión auditiva y el conocimiento de gramática y vocabulario evaluado en el componente de Uso del Inglés. Sin embargo, no se encontraron diferencias estadísticamente significativas al

comparar la expresión escrita y la comprensión de lectura de ambos grupos. Estos resultados pueden explicarse por las distintas metodologías seguidas en cada centro, el tiempo dedicado a cada destreza en clase, así como por el material y los recursos empleados.

2.- ¿Mejoran los estudiantes matriculados en cursos más orientados a la preparación de exámenes (CEB) su conocimiento lingüístico y sus habilidades?

Los alumnos matriculados en cursos más orientados a la preparación de examen no demuestran haber mejorado su conocimiento lingüístico y sus habilidades si consideramos los resultados en las pruebas independientes. No obstante, resulta necesario considerar todos los aspectos tratados en la discusión y las limitaciones, algunas de ellas se comentarán en las próximas líneas, para entender bien la situación.

3.- ¿Aumenta la autonomía y la independencia de los alumnos gracias a la preparación para el examen de Cambridge B2 First?

La preparación para el examen B2 First tiene un efecto parcial en la capacidad de los alumnos para mejorar su inglés y conocer sus fortalezas y debilidades. Debe tenerse en cuenta que los resultados ya eran muy positivos al principio del estudio y solo han mejorado levemente en algunos casos al final del mismo. Los resultados positivos obtenidos al final del estudio contrastan con la literatura consultada.

Si valoramos la autonomía y la independencia de los alumnos en función de qué o a quién consultan a la hora de aprender y prepararse de cara al examen, podemos concluir que los estudiantes matriculados en cursos de inglés general demuestran una actitud más independiente y autónoma hacia el aprendizaje que aquellos alumnos matriculados en cursos más orientados a la preparación de examen porque los primeros emplean principalmente su experiencia y reflexión personal mientras que los segundos tienen en su profesor o profesora a la principal referencia.

Además de las conclusiones ya mencionadas y que responden directamente a las preguntas de investigación, este proyecto aporta algunos hallazgos de interés:

1) Los resultados obtenidos al comparar los estudiantes de primer y segundo año matriculados en cursos más orientados a la preparación de exámenes apoyan la creencia de que los alumnos necesitan más de 120 horas para pasar de un nivel del MCER al siguiente.

2) Se aprecia un efecto positivo de la preparación del examen sobre el componente del Uso de Inglés. Así, si bien el grupo de control demostró un mejor rendimiento en los tests independientes, el grupo experimental lo tuvo en este componente.

3) El efecto del examen B2 First se pone de manifiesto en el hecho de que los alumnos que se preparan para este examen adaptan sus hábitos de estudio. El examen beneficia la motivación puesto que anima a los alumnos a esforzarse más y la mayoría de los alumnos mantuvo una actitud positiva hacia el examen aun cuando obtenían resultados negativos. Sin embargo, el efecto de los exámenes en los hábitos de aprendizaje puede verse como algo negativo, ya que los alumnos de cursos más enfocados al examen prefieren realizar más práctica de examen. Este hallazgo coincide con los resultados de los estudios consultados.

4) Las actividades empleadas en ambos tipos de cursos son muy similares y, de hecho, las actividades más frecuentes son las mismas, lo que sugiere, por un lado, que los estudiantes prefieren realizar actividades más orientadas al examen que sus profesores y, por otra parte, que el efecto en la práctica de clase es muy limitado y no es negativo.

Por último, el alcance de este estudio se encuentra limitado por distintas razones. (i) Los resultados no permiten generalizar el efecto del examen B2 First y los cursos más orientados a la preparación de examen puesto que los participantes pertenecían a solo dos centros y ambos se encontraban en Jaén (España). (ii) El momento en el que se realizó el estudio junto con el número de tests que los estudiantes tuvieron que realizar han limitado la atención y el compromiso de los estudiantes, lo que ha afectado a los tests de distinta manera. Los estudiantes de CEB estaban principalmente enfocados en la preparación del examen B2 First y en su examen de fin de curso, y a menudo se percibió que realizaban los tests independientes, que no tenían ninguna relevancia académica para ellos, bastante rápido. Los estudiantes de CEALM realizaron de una forma similar en este caso la prueba de B2 First porque consideraban que estaba por encima de su nivel y no tenía repercusión académica para ellos. (iii) Por último, la falta de diferencias estadísticamente significativas entre los tests de principio y los tests de final de curso pueden sugerir que el tiempo entre ambos fue demasiado corto para que los estudiantes pudieran conseguir unas mejoras significativas.

En vista de las conclusiones obtenidas y de las limitaciones observadas, se hace necesario continuar investigando en el futuro. Resultaría de utilidad realizar un estudio más amplio y extenso con más estudiantes y centros de idiomas con el fin de poder obtener una muestra representativa que permita a la autora generalizar los hallazgos. De hecho, la obtención de una muestra más amplia reduciría el efecto distorsionador que algunos factores pueden tener sobre los resultados e incrementaría la robustez de las conclusiones. Además, sería deseable analizar el rendimiento en un periodo de tiempo más extenso para poder obtener datos más fiables sobre mejoras de rendimiento significativas. De esta manera, el factor tiempo no sería tan determinante y se detectarían patrones más evidentes. Desde un punto de vista diferente, otra línea de investigación podría centrarse en el papel de los docentes y su posible influencia en su alumnado. De esta manera se podría decidir hasta qué punto el efecto de los exámenes en los estudiantes se debe al examen propiamente dicho y en qué medida la influencia del profesor se siente en el aprendizaje de los alumnos, sus estrategias de examen e incluso su motivación. Los profesores y sus alumnos comparten una relación especial basada en la confianza y el apoyo, y resultaría interesante analizar el factor docente pero como una variable más para entender el efecto del examen en los estudiantes.

La idea de realizar este proyecto de investigación nació en el aula después de haber enseñado a alumnos durante algunos años y de haber realizado una investigación más limitada sobre el efecto de los exámenes en el marco de un Trabajo Fin de Máster. La principal preocupación de los docentes es ayudar a sus alumnos a aprender de la mejor manera posible y, como profesional, sentí la necesidad de tener una actitud más reflexiva y conseguir datos sobre el proceso de aprendizaje de los estudiantes que se preparan para el examen B2 First. El diseño y la realización de este proyecto han resultado enriquecedores y han favorecido esa actitud reflexiva. Leer distintos estudios y los puntos de vista compartidos por profesionales de contextos muy diferentes ha animado a la autora a no dar nada por sentado. Los resultados obtenidos, con sus limitaciones, confirman la idea de que el fenómeno que nos ocupa es complejo. Así, los nuevos datos empíricos aportados por el presente trabajo en parte concuerdan con algunos resultados previamente publicados, pero en parte también contrastan con los de otras contribuciones aparecidas, aun siendo todos ellos de autores relevantes en la materia. En todo caso, constituyen una fuente de inspiración para investigar.

8. REFERENCES

Aftab, A., Qureshi, S. & William, I. (2014). Investigating the Washback Effect on the Pakistani Intermediate English Examination, *Academic Journals*, *5*(7): 149-154.

Ahmad, S. & Rao, C. (2012). Examination Washback Effect: Syllabus, Teaching Methodology and the Learners' Communicative Competence, *Journal of Education and Practice, 3*(15).

Alcaraz, N. (2015). Aproximación histórica a la evaluación educativa: De la generación de la medición a la generación ecléctica, *Revista Iberoamericana de Evaluación Educativa*, *8*(1): 11-25.

Alderson, J. C. (1990). Testing Reading Comprehension Skills, *Reading in a Foreign Language, 6*(2): 425-438.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2004). Foreword. In: L. Cheng & Y. Watanabe with A. Curtis, eds., *Washback in Language Testing. Research Contexts and Methods*, 1st ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc. Publishers, ix-xii.

Alderson, J. C. & Banerjee, J. (2001). Impact and washback research in language testing. In: C. Elder et al., ed., *Experimenting with Uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press, 150-161.

Alderson, J. C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback, *Language Testing, 13*(3): 280-297.

Alderson, J. C. & Wall, D. (1993). Does Washback Exist?, *Applied Linguistics, 14*(2): 115-129.

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context, *Language Testing in Asia, 6*(7): 1-20.

Allwright, D. & Bailey, K. M. (1991). *Focus on the Language Classroom*. Cambridge: Cambridge University Press.

Ambrose, S., Bridges, M., DiPietro, M., Lovett, M. & Norman, M. (2010). *How Learning Works: 7 Research-Based Principles for Smart Teaching*. San Francisco: Jossey-Bass.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning.

[online] Available at: http://epaa.asu.edu/epaa/v10n18/ [Accessed 12 November 2005].

Andrews, S. J., Fullilove, J. & Wong, Y. (2002). Targeting washback: a case study, *System, 30*, 207–33.

Arnove, R. F., Altback, P. G. & Kelly, G. P. (1992). *Emergent Issues in Education: Comparative Perspectives*. Albany, NY: State University of New York Press.

Association of Language Testers in Europe (ALTE) (2001). *Principles of Good Practice for ALTE Examinations*. Unpublished internal document. ALTE.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667350.

Bachman, L. F. & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing, *Language Testing, 13*(3): 257-279.

Bailey, K. M. (1999). Washback in language testing, *TOEFL Monograph Series. Report Number: RM-99-04, TOEFL-MS-15*. Available at https://www.ets.org/Media/Research/pdf/RM-99-04.pdf [Accessed 27 July 2017].

Bailey, K. M. (1999). *Washback in Language Testing. TOEFL Monograph Series. Report Number: RM-99-04, TOEFL- MS-15*. Princeton, NJ: Educational Testing Service.

Bailey, K. M. & Masuhara, H. (2013). Language testing washback: The role of materials. In: B. Tomlinson, ed., *Applied Linguistics and Materials Development*, 1st ed. London; New York: Bloomsbury Academic. DOI: http://dx.doi.org/10.5040/9781472541567.ch-019.

Baker, F. (2013). Using corpora to design assessment. In: A.J. Kunnan, ed, *The Companion to Language Assessment,* 1st ed. Hoboken: Wiley Blackwell.

Beikmahdavi, N. (2016). Washback in language testing: review of related literature first, *International Journal of Modern Language Teaching and Learning, 1*(4): 130-136.

Bell, J. & Thomas, A. (2014). *Gold First Coursebook*. Essex: Pearson Education Limited.

Bestard Monroig, J. & Pérez Martínez, M. C. (1992). *La didáctica de la lengua inglesa. Fundamentos lingüísticos y metodológicos*. Madrid: Síntesis.

Black, P. J. & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment*. London: Kings College London School of Education.

Booth, D. K. (2012). Exploring the washback of TOEIC in South Korea. Unpublished Doctoral Dissertation: University of Auckland.

Booth, D. K. & Lee, N. D. (2019). Learner Perceptions and Washback of the Paper-Based TOEFL Test on Student Affect at one Japanese University, *KSAALT-TESOL Academic Journal, 1*(1): 9-23.

Bray, M. & Steward, L. (1998). *Examination Systems in Small States: Comparative Perspectives on Policies, Models and Operations*. London: Commonwealth Secretariat.

British Council (2017). *IELTS* [online]. Exámenes. Available at https://www.britishcouncil.es/examenes/ielts [Accessed 13 July 2017].

Brook-Hart, G. & Owen, D. (2011). *Complete First Certificate*. Madrid: Cambridge University Press.

Brown, B. L. (2008). The Information Processing Model of Memory. Lecture Handout for Psychology 1101. In: *Introduction to General Psychology Course* [online] Georgia. Available at: http://facstaff.gpc.edu/~bbrown/ psyc1101/memory/3boxmodel.htm [Accessed 18 October 2020].

Brown, G. & Yule, G. (1983). *Discourse Analysis*. Cambridge: Cambridge University Press.

Brown, H. D. (1987). *Principles of Language Learning and Teaching*. New Jersey: Prentice-Hall Regents.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. San Francisco, CA: Pearson ESL.

Brown, J. D. (1998). An investigation into approaches to IELTS preparation, with particular focus on the academic writing component of the test. In: S. Wood, ed., *IELTS Research Reports 1998, Volume 1.*, 1st ed. Sydney: ELICOS Association / Canberra: IELTS Australia.

Brownell, J. (2002). *Listening: Attitudes, Principles, and Skills*, 2nd ed. Boston: Allyn and Bacon.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations, *JALT Journal, 10*: 12-42.

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

Bueno González, A. (1996). Testing English as a foreign language: an overview and some methodological considerations, *RESLA,* 11: 17-49.

Bueno González, A. (2015). Data collection and analysis: procedures, instruments, and basic statistical techniques. In: M.L. Pérez Cañado & B. Pennock-Speck, eds., *Writing and Presenting*

a Dissertation on Linguistics, Applied Linguistics and Culture Studies for Undergraduates and Graduates in Spain, 1st ed. Valencia: Universidad de Valencia.

Burrows, C. (1998). Searching for Washback: an investigation of the impact on teachers of the implementation into the Adult Migrant English Program of the assessment of the Certificates in Spoken and Written English. Unpublished Doctoral Dissertation: Macquarie University.

Burrows, C. (2004). Washback in Classroom-Based Assessment: A Study of the Washback Effect in the Australian Adult Migrant English Program. In: L. Cheng, Y. Watanabe & A. Curtis, eds., *Washback in Language Testing: Research Context and Methods*, 1st ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Cambridge English (2014). *Cambridge English: First (from 2015), Victoria and Edward* [online]. Available at https://www.youtube.com/watch?v=EdeZp0n0JHw [Accessed 11 August 2017].

Cambridge English De (2012). *Cambridge English First DOs and DON'Ts* [online]. Available at https://cambridgeesolde.wordpress.com/2012/11/12/cambridge-english-first-dos-and-donts/ [Accessed 31 August 2017].

Cambridge International Examinations (2014). *Cambridge Handbook 2015 (International). Regulations for Conducting Cambridge Examinations* [online]. Available at https://abisoman.com/staff/wp-content/uploads/sites/16/2014/05/Cambridge-Handbook-2015-international-chapter-5.pdf [Accessed 1 September 2017].

Cambridge University Press (2017). *Cambridge English. Resources* [online]. Available at https://www.cambridge.org/us/cambridgeenglish/resources?webSubjAll%5B%5D=Adult+Courses&webSubjAll%5B%5D=Business%2C+Professional+and+Vocational&webSubjAll%5B%5D=Cambridge+English+Exams+%26+IELTS&webSubjAll%5B%5D=Dictionaries&webSubjAll%5B%5D=Grammar%2C+Vocabulary+and+ [Accessed 8 August 2017].

Cambridge University Press (n.d.). *English Vocabulary Profile*. Available at: https://www.englishprofile.org/wordlists [Accessed 15 July 2020].

Cambridge University Press and University of Cambridge Local Examinations Syndicate (2016). *First 2. Authentic Examination Papers*. Cambridge University Press.

Campbell, R. & Wales, R. (1970). The study of language acquisition. In: J. Lyon ed. *New Horizons in Linguistics*, 1st ed. Harmondsworth, UK: Penguin.

Canale, M. (1983). On some dimensions of language proficiency. In: J.W. Oller, Jr., ed., *Issues in*

Language Testing Research. Rowley: Newbury House, 333-342.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics, 1*(1): 1-47.

Capel, A. & Sharp, W. (2014). *Objective First 4th Edition*. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Carroll, J. B. (1961). The nature of data, or how to choose a correlation coefficient, *Psychometrika, 26*(4): 347-72.

Carroll, J. B. (1968). The psychology of language testing. In: A. Davies, ed., *Language Testing Symposium. A Psycholinguistic Perspective*, 1st ed. London: Oxford University Press, 46-69.

Carroll, J. B. (1980). *Testing Communicative Performance*. Oxford: Pergamon.

Carroll, J. B. Carton, A. S. & Wilds, C. P. (1959). *An Investigation of Cloze Items in the Measurement of Achievement in Foreign Languages*. Cambridge, MA: College Entrance Examination Board.

Celestine, C. & Su, M. (1999). The effect of background disciplines on IELTS scores. *IELTS Research Report, 2*.

Centro de Estudios Avanzados en Lenguas Modernas (n.d.). Examinadores y Correctores de Pruebas de Acreditación. Available at: https://cealm.ujaen.es/sites/centro_cealm/files/uploads/Idoneidad%20de%20los%20Examinadores_CEALM%20UJA.pdf [Accesssed 7 July 2020].

Chávez-Zambrano, M., Saltos-Vivas, M. A. & Saltos-Dueñas, C. M. (2017). La importancia del aprendizaje y conocimiento del idioma inglés en la enseñanza superior, *Dominio de las Ciencias, 3*: 759-771. DOI: http://dx.doi.org/10.23857/dom.cien.pocaip.2017.3.mono1.ago.759-771.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong, *Language and Education, 11*(1): 38-54.

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning, *Studies in Educational Evaluation, 24*(3): 279-301.

Cheng, L. (2001). Washback studies: Methodological considerations, *Curriculum Forum, 10*(2): 17-32.

Cheng, L. (2005). *Changing Language Teaching through Language Testing: A Washback Study. Studies in Language Testing 21.* Cambridge: Cambridge ESOL/Cambridge University Press.

Cheng, L. (2008). Washback, impact and consequences. In: E. Shohamy & N. H. Hornberger, eds.,

Encyclopedia of Language and Education, 2nd ed. *Language Testing and Assessment* (vol. 7), New York, NY: Springer, 349-364.

Cheng, L. (2010). How does washback influence teaching? Implications for Hong Kong, *Language and Education*, *11*(1): 38-54.

Cheng, L. & DeLuca, C. (2011). Voices from test-takers: Further evidence for test validation and test use. *Educational Assessment 16*(2): 104-122.

Cheng, L. & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework, *Language Assessment Quarterly, 12*(1): 50-66.

Cheng, L., Sun, Y. & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework, *Language Teaching, 48(*4): 436-470. DOI:10.1017/S0261444815000233.

Cheng, L. Y., Watanabe, Y. & Curtis, A. (2004). *Washback in Language Testing Research: Contexts and Methods*. Mahwah: Lawrence Erlbaum.

Chenoweth, A. & Hayes, J. (2001). Fluency in writing: Generating text in L1 and L2, *Written Communication*, 18: 80-98.

Clapham, C. (2000). Assessment and testing, *Annual Review of Applied Linguistics, 20*, 147-161. https://doi.org/10.1017/S0267190500200093.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing?, *System, 25*: 201-315.

Collins, S., Reiss, M. & Stobart, G. (2010). What happens when high-stakes testing stops? Teachers' perceptions of the impact of compulsory national testing in science of 11-year-olds in England and its abolition in Wales, *Assessment in Education: Principles, Policy and Practice*, *17*(3): 273-286.

Cooper, R. L. (1968). An elaborated language testing model. In: J. A. Upshur & J. Fata, eds., *Problems in Foreign Language Testing. Language Learning Special Issue*, Nº 3. Ann Arbor, MI: Research Club in Language Learning, 57-72.

Council of Europe (1970). *A Compendium of Studies Commissioned by the Council for Cultural Co-operation: A Contribution to the United Nations' International Education Year*. Strasbourg: Council of Europe.

Council of Europe (2001). *Common European Framework of Reference for Languages. Learning, Teaching and Assessment*. Cambridge: Cambridge University Press.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Available at https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 [Accessed 5 April 2019].

Craven, E. (2012). The quest for IELTS Band 7.0: Investigating English language proficiency development of international students at an Australian university. In: J. Osborne, ed., *IELTS Research Reports* (Vol. 13). Canberra: IELTS Australia and Manchester: British Council, 1-61.

Crombach, L. & Meehl, P. (1995). Construct validity in psychological tests, *Psychological Bulletin, 52*: 281-302.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students, *Review of Educational Research, 58*(4): 438-481.

Cruz Trapero, J. M. (2016). *Protocol to design a CEFR-linked proficiency rating scale for oral production and app implementation* [online]. Available at: http://ruja.ujaen.es/jspui/bitstream/10953/795/1/9788491590286.pdf [Accessed 4 September 2020].

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters, *Working Papers on Bilingualism, 19*: 121-29.

Cutler, A. (2005). *Twenty-first Century Psycholinguistics: Four Cornerstones*. Hillsdale: Erlbaum.

Damankesh, M. & Babaii, E. (2015). The washback effect of Iranian high school final examinations on students' test-taking and test-preparation strategies, *Studies in Educational Evaluation, 45*: 62-69. https://doi.org/10.1016/j.stueduc.2015.03.009.

Davidson, F. & Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven: Yale University Press.

Davies, A. (1968). *Language Testing Symposium: A Psycholinguistic Approach*. Oxford: Oxford University Press.

Davies, A. (1985). Follow my leader: Is that what language tests do? In: Y. P. Lee, C. Y. Y. Fok, R. Lord & G. Low, eds., *New Directions in Language Testing*. Oxford: Pergamon Press, 1-12.

Davies, A. (1997). Demands of Being Professional in Language Testing, *Language Testing*, 14(3): 328-339.

Davies, A. (2008a). Textbook trends in teaching language testing, *Language Testing, 25*(3): 327-347.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *A Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Davies, B. (2008b). *Assessing Academic English: Testing English Proficiency 1950-1989: The IELTS Solution*. Cambridge: Cambridge University Press.

Denton, P. (1992). Seating arrangements for better classroom management, *Adventist Education, 54*(5): 29-32.

Denzin, N. K. (1970). *Sociological Methods: A Source Book*. Chicago: Aldine.

Dhanapala, R. M. (2019). *Significance of Listening skill in the ESL and EFL context*. Available at: https://www.researchgate.net/publication/336196565_Significance_of_Listening_skill_in_the_ESL_and_EFL_context [Accessed 16 October 2020].

Díez-Bedmar, M. B. (2010). *Análisis de la expresión escrita en inglés en la prueba de acceso a la Universidad de Jaén*. Jaén: Universidad de Jaén.

Dong, M. (2020). Structural relationship between learners' perceptions of a test, learning practices, and learning outcomes: A study on the washback mechanism of a high-stakes test, *Studies in Educational Evaluation*, *64*: 1-11. https://doi.org/10.1016/j.stueduc.2019.100824.

Eckstein, M. A. & Noah, H. J. (1992). *Examinations: Comparative and International Studies*. Oxford: Pergamon Press.

Educational Testing Services (2020). *Why Choose the TOEFL® Test?* [online]. TOEFL Ibt. Available at: https://www.ets.org/toefl/test-takers/ibt/why/ [Accessed 19 September 2020].

Elder, C., Iwashita, N. & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test taker have to offer?, *Language Testing, 19*(4): 347-368.

Elder, C. & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. In: R. Tulloh, ed., *IELTS Research Reports (Vol. 4),* 1st ed. Canberra: IELTS Australia, 207-254.

Estaji, M. (2013). Demystifying the Complexity of Washback Effect on Learners in the IELTS Academic Writing Test, *Study in English Language Teaching, 1*(1): 211-226.

European Commission (2017). Developing Key Competences for all throughout Life. Available at: https://ec.europa.eu/education/sites/education/files/document-library-docs/factsheet-key-competences-lifelong-learning_en.pdf [Acccessed 18 November 2020].

European Commission (n. d.). About Multilingualism Policy [online]. Education and Training.

Available at: https://ec.europa.eu/education/policies/multilingualism/about-multilingualism-policy_en [Accessed 18 November 2020].

Exam English Ltd. (2014). *Cambridge English: First (FCE)* [online]. Available at http://www.examenglish.com/FCE/index.php [Accessed 11 August 2017].

Ferman, I. (2004). The Washback of an EFL National Oral Matriculation Test to Teaching and Learning. In: L. Cheng, Y. Watanabe & A. Curtis, eds., *Washback in Language Testing: Research Context and Methods*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Field, J. (2008). *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. In: L. Taylor, ed., *Studies in Language Testing. Examining Speaking,* 1st ed. Cambridge: Cambridge University Press.

Flanders, N. A. (1970). *Analyzing Teaching Behaviour*. Boston: Addison-Wesley.

Frederiksen, J. R. & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher, 18*(9): 27-32.

Fulcher, G. (2003). *Testing Second Language Speaking*. Harlow: Longman.

Fulcher, G. (2010). *Practical Language Testing.* London: Hodder Education.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment*. Routledge.

Gabrielatos, C. (1993). Learning How to Fish: Fostering Fluency and Independence. *TESOL Greece Newsletter*, *38*: 23-26.

Galaczi, E. & A. Ffrench. 2011. Context validity. In: L. Taylor, ed., *Studies in Language Testing. Examining Speaking*. Cambridge: Cambridge University Press.

Gest, S. D. & Rodkin, P. C. (2011). Teaching practices and elementary classroom peer ecologies, *Journal of Applied Developmental Psychology, 32*: 288-296. DOI:10.1016/j.appdev.2011.02.004.

Gosa, C. M. C. (2004). Investigating Washback: A Case Study Using Students' Diaries. Unpublished PhD thesis. Lancaster: Lancaster University.

Gough, P. B., Hoover, W. A. & Peterson, C. L. (1996). Some observations on a simple view of reading. In: C. Cornoldi & J. Oakhill, eds., *Reading Comprehension Difficulties*. Mahwah: Erlbaum.

Grabe, W. (1991). Current Developments in Second Language Reading Research. *TESOL Quarterly, 25*(3): 375-406.

Graddol, D. (2006). *English Next*. London: British Council.

Gray, W. S. (1960). The major aspects of reading. In: J. Robinson, ed., *Sequential Development of Reading Abilities*. Chicago: Chicago University Press.

Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test, *Assessing Writing, 10*: 44-60.

Green, A. (2006a). Watching for washback: Observing the influence of the International English Language Testing System academic writing test in the classroom, *Language Assessment Quarterly, 3*(4): 333-368.

Green, A. (2006b). Washback to the learner: learner and teacher perspectives on IELTS preparation course expectations and outcomes, *Assessing Writing, 11*: 113-134.

Green, A. (2007a). *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*. Cambridge: Cambridge University Press.

Green, A. (2007b). Washback to learning outcomes: A comparative study of IELTS preparation and university pre- sessional language courses, *Assessment in Education, 14*(1): 75-97.

Green, A. (2013). Washback in Language Assessment, Achieving Beneficial Backwash, *International Journal of English Studies, 13*(2): 39-51. DOI: https://doi.org/10.6018/ ijes.13.2.185891.

Green, A. (2014a). *Exploring Language Assessment and Testing*. Language in Action. Oxon: Routledge.

Green, R. (2014b). *Language Testing at Lancaster* [course].

Gremmen, M. C., van den Berg, Y. H. M., Segers, E. et al. (2016). Considerations for classroom seating arrangements and the role of teacher characteristics and beliefs, *Social Psychology of Education 19*: 749-774. https://doi.org/10.1007/s11218-016-9353-y.

Ha, N. T. T. (2019). A literature review of washback effects of assessment on language learning, *Journal of Science Ho Chi Minh City Open University, 9*(5): 3-15.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: in search of the evidence, *Educational Measurement: Issues and Practices, 18*(4): 5-9.

Hakim, L. N. & Tasikmalaya, U. P. (2018). Washback Effect in Language Testing: What Do We Know and What Is Its Effect?, *Jurnal Forum Didaktik, 2*(1): 59-68.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. Harlow: Longman.

Hamp-Lyons, L. (1991). *Assessing Second Language Writing in Academic Contexts*. Norwood: Ablex Publishing Corporation.

Hamp-Lyons, L. (1997). Washback, impact and validity: ethical concerns, *Language Testing, 14*(3): 295-303.

Hamp-Lyons, L. (2001). Ethics, Faimess(Es), and Developments in Language Testing. In: C. Elder, 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.

Hamp-Lyons, L. & Brown, A. (2007). *The Effect of Changes in the New TOEFL Format on the Teaching and Learning of EFL/ESL: Stage 2 (2003–2005), Entering Innovation*. Princeton, NJ: Educational Testing Service.

Harlen, W. & Crick, R. D. (2003). Testing and motivation for learning, *Assessment in Education: Principles, Policy and Practice, 10*(2): 169-207.

Hasselgren, A. (1998). Smallwords and good testing. Studia Humanitas Bergensia. Unpublished PhD Thesis, University of Bergen.

Hawkey, R. A. (2006). *Studies in Language Testing: Vol. 24. Impact Theory & Practice: Studies of the IELTS test and Progetto Lingue 2000.* Cambridge: Cambridge University Press.

Hawkey, R. A. (2011). Consequential validity. In: L. Taylor, ed., *Studies in Language Testing. Examining Speaking*. Cambridge: Cambridge University Press.

Hawkey, R. A. & Milanovic, M. (2013). *Cambridge English Exams. The first Hundred Years*. Cambridge: Cambridge University Press.

Hawkins, J. A. & Filipovic, L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. Cambridge: Cambridge University Press.

He, A. W. & Young, R. (1998). Language Proficiency interviews: A discourse approach. In: R. Young & A. W. He, eds., *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam: Benjamins.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation and Research*. Cambridge: Newbury House.

Hewings, M. (1999). *Advanced Grammar in Use*. Cambridge: Cambridge University Press.

Hu, C. T. (1984). The historical background: Examinations and controls in pre-modern China, *Comparative Education, 20*, 7–26.

Huerta-Macías, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5*(1): 8-11.

Hughes, A. (1988). Introducing a needs-based test of English language proficiency into an English medium university in Turkey. In: A. Hughes, ed., *Testing English for University Study (ELT Documents #127).* London: Modem English Publications in association with the British Council, 134-146.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hughes, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript, University of Reading, England.

Hughes, A. (2003). *Testing for Language Teachers*, 2nd ed. Cambridge: Cambridge University Press.

Hughes, R. (2010). *Teaching and Researching: Speaking*, 2nd ed. London: Longman.

Humphreys, P., Haugh, M., Fenton-Smith, M., Lobo, A., Michael, R. & Walkinshaw, I. (2012). Tracking international students' English proficiency over the first semester of undergraduate study. In: J. Osborne & G. Lim, eds., *IELTS Research Report Series* (Vol. 1) Canberra: IDP: IELTS Australia, 1-41.

Hymes, D. H. (1972). On communicative competence. In: J. B. Pride & J. Holmes, eds., *Sociolinguistics*. Harmondsworth: Penguin, 269-293.

Instituto de Estadística y Cartografía de Andalucía (2020). *Jaén*. Available at: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/ficha.htm?mun=23050 [Accessed 7 July 2020].

International English Language Testing System (2020). IELTS numbers rise to three million a year [online] *News*. Available at: https://www.ielts.org/news/2017/ielts-numbers-rise-to-three-million-a-year [Accessed 19 September 2020].

International Language Testing Association (2000). *Code of Ethics*. Birmingham: ILTA.

International Language Testing Association (2007). *ILTA Guidelines for Practice*. Birmingham: ILTA.

Jaimechango. (2009). *Importancia del inglés en la educación*. Available at: https://es.slideshare.net/jaimechango/importancia-del-ingles-en-la-educacon [Accessed 18 November 2020].

Jung, E. H. (2003). The role of discourse signalling cues in second language listening comprehension, *The Modern Language Journal, 87*: 562-577. Available at http://dx.doi.org/10.1111/1540-4781.00208 [Accessed 18 October 2020].

Karabulut, A. (2007). Micro Level Impacts of Foreign Language Test (University Entrance Examination) in Turkey: A Washback Study. (MA thesis). Retrospective Theses and

Dissertations. Available at: http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=15883&context=rtd [Accessed 18 October 2020].

Kenny, N. (1995). The New FCE Reading Paper. *ELT News*, 74(15).

Khalifa, H. (2014). *Research Notes. Issue 58*. [online]. Available at http://www.cambridgeenglish.org/images/182921-research-notes-58-document.pdf [Accessed 12 February 2015].

Khalifa, H. & Weir, C. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge: Cambridge University Press.

Kim, E. Y. J. (2017). The TOEFL iBT writing: Korean students' perceptions of the TOEFL iBT writing test, *Assessing Writing, 33*: 1-11.

Kiomrs, R., Abdolmehdi, R. & Naser R. (2011). On the Interaction of Test Washback and Teacher Assessment Literacy: The Case of Iranian EFL Secondary School Teachers, *English Language Teaching, 4*(1): 156-161.

Klatzky, R. 1980. *Human Memory: Structures and Processes*. San Francisco: W.H. Freeman & Co.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F. & Stecher, B. M. (2000). What do test scores in Texas tell us? [online] Available at http://epaa.asu.edu/epaa/v8n49/ [Accessed 12 November 2018].

Klinger, D., DeLuca, C. & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy* 76: 1-34.

Knight, B. (2020). *Principles of Language Learning* [online].

Kremmel, B. & Schmitt, N. (2017). *Vocabulary Level Test* [online]. Available at http://onlinelibrary.wiley.com/doi/10.1002/9781118784235.eelt0499/full [Accessed 3 August 2017].

Külekçi, E. (2016). A concise analysis of the Foreign Language Examination (YDS) in Turkey and its possible washback effects, *International Online Journal of Education and Teaching, 3*(4): 303-315. Available at http://iojet.org/index.php/IOJET/article/view/141/143 [Accessed 18 October 2020].

Lado, R. (1961). *Language Testing*. New York: McGraw-Hill.

Lai, C. T. (1970). *A Scholar in Imperial China*. Hong Kong: Kelly & Walsh.

Larsen, R. P. & Feder, D. D. (1940). Common and differential factors in reading and hearing comprehension, *Journal of Educational Psychology, 31*(4): 241-

252. https://doi.org/10.1037/h0060424.

Larsson, M. & Ollin-Scheller, C. (2020). Adaptation and resistance: washback effects of the national test on upper secondary Swedish teaching, *The Curriculum Journal.* Available at: https://www.researchgate.net/publication/339092698_Adaptation_and_resistance_washback_effects_of_the_national_test_on_upper_secondary_Swedish_teaching [Accessed 31 July 2020].

Latham, H. (1877). *On the Action of Examinations Considered as a Means of Selection*. Cambridge: Deighton, Bell and Company.

Latham-Koenig, C. & Oxenden, C. (2014*). English File Upper Intermediate Student's Book.* Oxford: Oxford University Press.

Laufer, B. & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context, *Language Learning, 48*: 365–391.

Leung, C. & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment, *TESOL Quaterly, 40*(1): 211-234.

Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Li, X. (1990). How Powerful Can a Language Test Be? The MET in China, *Journal of Multilingual and Multicultural Development, 11*(5): 393-404.

Linn, R. L. (2000). Assessments and accountability, *Educational Researcher, 29*(2), 4-16.

Lumley, T. & Stoneman, B. (2000). Conflicting perspectives on the role of test preparation in relation to learning, *Hong Kong Journal of Applied Linguistics, 5*(1): 50-80.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test, *Language Testing, 22*(2): 142-173.

Madrid, D. (1995). Internal and external factors affecting foreign language teaching. In: C. Medina and L. García, eds., *I Jornadas de Estudios Ingleses, 1st ed*. Jaén: Servicio de Publicaciones de la Universidad de Jaén, 59-82.

Madrid, D. (1998a). El área curricular de los idiomas: epistemología y desarrollo. In: L. Rico & D. Madrid, eds., *Las didácticas Especiales,* 1st ed. Madrid: Síntesis.

Madrid, D. (1998b). *Guía para la investigación en el aula de idiomas*. Granada: Grupo Editorial Universitario.

Madrid, D. (n. d.). Observation and research in the language classroom. *Master in Linguistics Applied to the Teaching of English as a Foreign Language.* Barcelona: Fundación Universitaria Iberoamericana.

Madrid, D. & Bueno, A. (2005). Classroom research. In: N. McLaren, D. Madrid & A. Bueno, eds., *TEFL in Secondary Education,* 1st ed. Granada: Editorial Universidad de Granada, 641-677.

Madsen, H. (1983). *Techniques in Testing*. New York: Oxford University Press.

Manchón, R. M., Murphy, L., Roca de Larios, J. & Aguado, P. (2005). Learning and teaching writing in the EFL classroom. In: N. McLaren, D. Madrid, A. Bueno, eds., *TEFL in Secondary Education,* 1st ed. Granada: Universidad de Granada.

Mason, J. (2002). *Qualitative Researching 2nd ed.* London, Thousand Oaks, New Delhi: Sage.

McEwen, N. (1995). Educational accountability in Alberta, *Canadian Journal of Education, 20*: 27-44.

McKeown, S., Stringer, M. & Cairns, E. (2015). Classroom segregation: Where do students sit and how is this related to group relations?, *British Educational Research Journal, 42*(1): 40–55. doi:10.1002/berj.3200.

McNamara, T. (1996). *Measuring Second Language Performance*. Harlow: Longman.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Mehrens, W. A. (1998). Consequences of assessment: what is the evidence?, *Education Policy Analysis Archives, 6*(13): 1-30.

Messick, S. (1996). Validity and washback in language testing, *Language Testing, 13*(3): 241-256.

Messick, S. (1998). Test validity: A matter of consequence, *Social Indicators Research, 45*(1): 35-44.

Michaelides M. P. (2014). Validity considerations ensuring from examinees' perceptions about high-stakes national examination in Cyprus, *Assessment in Education: Principles, Policy and Practice, 21*(4): 427-41.

Mickan, P. & Motteram, J. (2008). An ethnographic study of classroom instruction in an IELTS preparation program, *IELTS Research Reports, 8*: 17-43.

Mickan, P. & Motteram, J. (2009). The preparation practices of IELTS candidates: Case studies, *IELTS Research Reports, 10:* 4-39.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In: C. J. Brumfit & K. Johnson, eds., *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press, 143-157.

Morrow, K. (1986). The Evaluation of Tests of Communicative Performance. In: M. Portal, ed.,

Innovations in Language Testing: Proceedings of the IUS/NFER Conference. London: NFER/Nelson.

Moskowitz, G. (1971). Interaction analysis –a new modern language for supervisors, *Foreign Language Annals, 1(3)*: 218-235.

Muñoz, A. P. & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom, *Language Testing 27*(1): 33-49.

Muñoz, P., Véliz-Campos, M. & Véliz, L. (2019). Assessment in the English language classroom in Chile: Exploring the washback effect of traditional testing and alternative assessment on seventh grade students, *Paideia, 64*: 97-118, https://doi.org/10.29393/Pa64-4APM30004.

Murphy, R. (2004). *English Grammar in Use*. Cambridge: Cambridge University Press.

Murray, J. C., Riazi, A. M. & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia, *Language Testing, 29*(4): 577-595.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.

Nunan, D. (2002). Listening in Language Learning. In: J.C. Richards & W.A. Renandya, eds., *Methodology in Language Teaching*. Cambridge: Cambridge University Press.

Ochs, E. (1979). Transcription as theory. In: E. Ochs & B. B. Schieffelin, eds., *Developmental Pragmatics*. New York: Academic Press.

Oller, J. W. Jr. (1979). *Language Tests at School*. Harlow, UK: Longman.

O'Loughlin, K. & Arkoudis, S. (2009). Investigating IELTS exit score gains in higher education. In: J. Osborne, ed., *IELTS Research Reports (Vol. 10*), 1st ed. Canberra: IELTS Australia and Manchester: British Council, 95-180.

Pan, Y. C. (2014). Learner Washback Variability in Standardized Exit Tests, *Teaching English as a Second or Foreign Language, 18*(2).

Pan, Y. C. & Newfields, T. (2013). Student washback from tertiary standardized English proficiency exit requirements in Taiwan, *Journal of Teaching and Learning, 9*(1): 1-16.

Pearson, I. (1988). Tests as Levers for Change. In: D. Chamberlain & R. J. Baumgardner, eds., *ESP in the Classroom: Practice and Evaluation. ELT Documents Volume 128.* London: Modern English Publications.

Peña Jaenes, V. (2015). *Testing Writing: Washback Effect on Language Courses.* MA Thesis. Universidad de Jaén.

Perrone, M. (2010). *The impact of the First Certificate of English (FCE) on the EFL classroom: A washback study*. Unpublished doctoral dissertation. Teachers College: Columbia University.

Perrone, M. (2011). The effect of classroom-based assessment and language processing on the second language acquisition of EFL students, *Journal of Adult Education, 40*: 20-33.

Petre, A. L. (2017). The impact of Alternative Assessment Strategies on students, *Scientific Research & Education in the Air Force-AFASES*, 2: 157-160.

Polesel, J., Rice, S. & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia, *Journal of Education Policy*, *29*(5): 640-657.

Popham, W. J. (1987). The merits of measurement-driven instruction, *Phi Delta Kappa, 68*: 679-682.

Popham, J. W. (2001) Teaching to the test?, *Educational Leadership*, *58*(6): 16-20.

Powers, D. E. (2010). *The Case for a Comprehensive, Four-Skills Assessment of English Language Proficiency*. [online]. ETS. Available at https://www.ets.org/research/policy_research_reports/publications/report/2010/itkc [Accessed 12 February 2015].

Prodromou, L. (1995). The Washback effect: from testing to teaching, *ELT Journal, 49*(1): 13-25.

Qi, L. (2004). Has a high-stakes test produced the intended changes? In: L. Cheng, Y. Watanabe & A. Curtis, eds., *Washback in Language Testing: Research Contexts and Methods*. Mahwah, NJ: Lawrence Erlbaum, 147-170.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test, *Language Testing, 22*(2): 142-173.

Queen's University (n. d.). Learning Challenges [online]. Getting Started. Available at: https://www.queensu.ca/teachingandlearning/modules/students/06_learning%20challenges.html [Accessed 18 November 2020].

Raban, S. (2008). *Examining the World. A History of the University of Cambridge Local Examinations Syndicate*. Cambridge: Cambridge University Press.

Rahimi, F., Esfandiari, M.R. & Amini, M. (2016). An overview of studies conducted on washback, impact and validity, *Studies in Literature and Language*, 13(4): 6-14.

Rao, Ch., McPherson, K. Chand, R. & Khan, V. (2003). Assessing the impact of IELTS preparation programs on candidates' performance on the General Training reading and writing test modules, *IELTS Research Reports, 5.*

Read, J. (1988). Measuring the vocabulary knowledge of second language learners, *RELC Journal,*

*19(*2): 12-25.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. & Hayes, B. (2003). The Impact of IELTS on preparation for academic study in New Zealand, *Washback in Language Testing: Research Contexts and Methods*, *4*: 153-206. DOI: 10.4324/9781410609731.

Reynolds, J. (2010). *An Exploratory Study of TOEFL Students as Evaluators of Washback to the Learners*. Research for the Master of Applied Linguistics. The University of Queensland.

Riley, J. (1996). *Getting the most from your Data. A Handbook on Practical Ideas on how to Analyse Qualitative Data, 2nd ed*. Bristol: Technical and Educational Services Ltd.

Riswandi, D. & Wahyudi, D. T. (2018). Can we find the washback effects of International Testing System on IELTS learners?, *Leksika,* 12(1): 15-19.

Robb, T. N. & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students, *Teaching English as a Second or Foreign Language*, 3(4): 1-22.

Rost, M. (2002). *Teaching and Researching Listening*. London: Longman.

Rubin, J. (1995). An overview to a guide for the Teaching of Second Language Listening. In: D. J. Mendelsohn & J. Rubin, eds., *A Guide to the Teaching of Second Language Listening*. San Diego: Dominie.

Saglam, G. L. A. (2018). Can Exams Change How and What Teachers Teach? Investigating the Washback Effect of a University English Language Proficiency Test in the Turkish Context, *Eurasian Journal of Applied Linguistics, 4*(2): 155-176. DOI: http://dx.doi.org/10.32601/ejal.464094.

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants, *Language Testing, 23*(1): 1-34 DOI: https://doi.org/10.1191/0265532206lt322oa.

Salkind, N. J. (2000). *Tests and Measurement for People who (Think they) Hate Tests and Measurement*. Los Angeles: Sage Publications Limited.

Schmitt, N. & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes, *Studies in Second Language Acquisition, 19*: 17-36.

Schmitt, N. Schmitt, D. & Clapham (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Level Test, *Language Testing, 18*(1): 55-88.

Schön, D. (1983). *The Reflective Practitioner. How Professionals Think in Action*. New York: Basic

Book Inc. Publishers.

Schön, D. (1987). *Educating the Reflective Practitioner. Towards a New Design for Teaching and Learning in the Profession*. San Francisco: Jossey-Bass.

Schön, D. (1990). *The Reflective Turn: Case Studies in and on Educational Practice*. New York: Teachers College Press.

Seliger, H. W. & Shohamy, E. (1989). *Second Language Research Methods*. Oxford: Oxford University Press.

Servicio Público de Empleo Estatal (2019). *Informe del Mercado de Trabajo de Jaén. Datos de 2018*. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiusurXrbvqAhUExoUKHdpCAdUQFjAAegQIBBAB&url=https%3A%2F%2Fwww.sepe.es%2FSiteSepe%2Fcontenidos%2Fque_es_el_sepe%2Fpublicaciones%2Fpdf%2Fpdf_mercado_trabajo%2F2019%2FMercado-de-Trabajo-Provincial-2019%2FMercado-de-Trabajo-2019-Ja-n--Datos-2018-.pdf&usg=AOvVaw2pYya8Hd7OOH2QfCjYJk_I [Accessed 7 July 2020].

Sevilla Morales, H. & Chaves Fernández, L. (2020). Washback Effects of Board-Based Speaking Tests, *Letras, 68*: 199-238. DOI: https://doi.org/10.15359/rl.2-68.8.

Shephard, W. G. (1984). The Cambridge Examinations- an Exercise in Public Relations. In: J. B. Heaton, ed., *Language Testing*. Great Britain: Modern English Publications Ltd, 28-31.

Shih, C. M. (2007). A new washback model of students' learning, *The Canadian Modern Language Review, 64*(1): 135-162.

Shih, C. M. (2009). How tests change teaching: A model for reference, *English Teaching, 8*(2): 188-206.

Shiotsu, T. & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance, *Language Testing, 24*(1): 99-128.

Shohamy, E. (1993). *The Power of Test: The Impact of Language Testing on Teaching and Learning. National Foreign Language Center Occasional Papers.* Washington, DC: National Foreign Language Center.

Shohamy, E. (1997). Critical Language Testing and Beyond, *Studies in Educational Evaluation, 24*(4): 331-345.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. New

York: Pearson Education.

Shohamy, E. (2008). Language Policy and Language Assessment: The Relationship, *Current Issues in Language Planning*, 9(3): 363-373. DOI: 10.1080/14664200802139604.

Shohamy, E., Donitsa-Schmidt, S. & Ferman, I. (1996). Test Impact Revisited: Washback Effect over Time, *Language Testing, 13*(3): 298-317.

Shohamy, E., Or, I. & May, S. (2017). *Language Testing and Assessment. Encyclopedia of Language and Education, 3rd ed.* Auckland: Springer.

Smith, M. L. (1991). Put to the test: the effects of external testing on teachers, *Educational Researcher, 20*(5): 8-11.

Smith, P. & King, J. R. (2013). An Examination of Veridicality in Verbal Protocols of Language Learners, *Theory & Practice in Language Studies*: 709-720.

Spada, N. & Fröhlich, M. (1995). *COLT Observation Scheme: Coding Conventions and Applications*. Sydney: National Centre for English Language.

Spolsky, B. (1995a). The examination of classroom backwash cycle: Some historical cases. In: D. Nunan, V. Berry & R. Berry, eds., *Bringing about Change in Language Education*. Hong Kong: University of Hong Kong, Department of Curriculum Studies, 55-66.

Spolsky, B. (1995b). *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.

Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams, *Language Teaching Research, 9*(1): 5-29.

Stoneman, B. W. H. (2006). *The impact of an exit English test on Hong Kong undergraduates: A study investigating the effects of test status on students' test preparation behaviors.* Unpublished doctoral dissertation. Hong Kong Polytechnic University.

Struyven, K., Dochy, F. & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review, *Assessment and Evaluation in Higher Education, 30*: 331-347.

Suchithra, N., Yew, L. K. & Kesumawati, A. B. (2014). Exploring the Listening Processes of Pre-University ESL Students, *Procedia - Social and Behavioral Sciences 118*: 475 - 482. Available at: https://www.researchgate.net/publication/273850602_Exploring_the_Listening_Processes_of_Pre-university_ESL_Students [Accessed 16 October 2020] DOI: https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1016%2Fj.sbspro.201

4.02.065.

Sukyadi, D. & Mardiani, R. (2011). The Washback Effect of the English National Examination, *K@ta, 13*(1): 96-111.

Sultana, N. (2018). Investigating the Relationship between Washback and Curriculum Alignment: A Literature Review, *Canadian Journal for New Scholars in Education, 9*(2): 151-158.

Sun, Y. (2016). *Context, construct, and consequences: washback of the college English test in China.* Unpublished doctoral dissertation. Queen's University.

Svinicki, M. D. (2004). *Learning and Motivation in the Postsecondary Classroom*. Bolton: Anker Publishing Company.

Swain, M. (1985). Large-scale communicative testing: A case study. In: Y. P. Lee, A. C. Y. Fok, R. Lord & G. Low, eds., *New Directions in Language Testing*. Oxford: Pergamon Press, 35-46.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores, *Language Testing, 18*(3): 275-302.

Takei, A. (2002). What should be identified about listening? In: A. Takei, ed., *Consideration of Listening in English: Inquiring Scientifically Listening Comprehension and Instruction,* 1st ed. Tokyo: Kagensha.

Taylor, L. (2005). Washback and impact, *ELT Journal, 59*(2): 154-155. DOI: https://doi.org/10.1093/eltj/cci030.

Templer, B. (2004). High-stakes testing at high fees: notes and queries on the international English proficiency assessment market, *Journal for Critical Education Policy Studies, 2*(1): 189-226.

Toksöz, I. & Kılıçkaya, F. (2017). Review of Journal Articles on Washback in Language Testing in Turkey (2010-2017), *Lublin Studies in Modern Languages and Literature, 41*(2): 184-204. DOI: 10.17951/lsmll.2017.41.2.184.

Towell, R., Hawkins, R. & Bazergui, N. (1996). The Development of fluency in advanced learners of French, *Applied Linguistics, 17*(1): 84-119.

Traynor, R. (1985). The TOEFL: an appraisal, *ELT Journal*, 39(1): 43-47.

Tsagari, D. (2006). *Investigating the Washback Effect of a High-stakes EFL exam in the Greek Context: Participants' perceptions, material design and Classroom applications*. Unpublished Doctoral Dissertation. University of Lancaster.

Tsagari, D. (2007). Review of washback in language testing: How has been done? What more needs doing? [online]. ResearchGate. Available at

https://www.researchgate.net/publication/234600418_Review_of_Washback_in_Language_Testing_What_Has_Been_Done_What_More_Needs_Doing [Accessed 17 February 2017].

Turner, C. (2001). The need for impact studies of L2 performance testing and rating: identifying areas of potential consequences at all levels of the testing cycle. In: A. Brown et al., eds., *Experimenting with Uncertainty: Essays in Honour of Alan Davies, Studies in Language Testing 11*. Cambridge: University of Cambridge Local Examinations Syndicate / Cambridge University Press, 138-149.

UNESCO (2016). UNESCO Institute for Lifelong Learning. [online]. UNESCO Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000246598 [Accessed 18 November 2020].

Ungerleider, C. (2003). *Failing our Kids: How we are Ruining our Public Schools*. Toronto: McClelland & Stewart.

Universidad de Jaén (n. d.). *Memoria Académica 2017-2018*. Available at: https://www.ujaen.es/gobierno/secgen/sites/gobierno_secgen/files/uploads/memorias/CG2_01905_anexo08_Memoria%20Academica%2017-18.pdf [Accessed 7 July 2020].

University of Cambridge Local Examinations Syndicate (2013a). *Cambridge English First*. Cambridge: University of Cambridge Local Examinations Syndicate.

University of Cambridge Local Examinations Syndicate (2013b). *Cambridge English First for Schools*. Cambridge: University of Cambridge Local Examinations Syndicate.

University of Cambridge Local Examinations Syndicate (2016a). *Cambridge English. Regulations*. Available at http://www.cambridgeenglish.org/images/84609-faq-regulations.pdf [Accessed 1 September 2017].

University of Cambridge Local Examinations Syndicate (2016b). *Principles of Good Practice. Research and Innovation in Language Learning and Assessment* [online]. Cambridge English. Available at: https://www.cambridgeenglish.org/Images/22695-principles-of-good-practice.pdf [Accessed 19 September 2020].

University of Cambridge Local Examinations Syndicate (2017a). *¿Por qué Cambridge English?* [online]. Cambridge English. Available at https://www.cambridgeenglish.org/es/why-cambridge-english/ [Accessed 13 August 2020].

University of Cambridge Local Examinations Syndicate (2017b). *Cambridge English First (FCE). Exam Format*. [online]. Cambridge English. Available at http://www.cambridgeenglish.org/exams/first/exam-format/ [Accessed 30 August 2017].

University of Cambridge Local Examinations Syndicate (2017c). *Code of Practice – Our Approach to Assessment* [online]. Available at: https://www.cambridgeinternational.org/Images/416992-code-of-practice.pdf [Accessed 19 September 2020].

University of Cambridge Local Examinations Syndicate (2017d). *Our Heritage* [online]. Cambridge English. Available at: http://www.cambridgeassessment.org.uk/about-us/who-we-are/our-heritage/ [Accessed 26 October 2017].

University of Cambridge Local Examinations Syndicate (2017e). *Producing Exams*. [online]. Cambridge English. Available at http://www.cambridgeenglish.org/why-cambridge-english/producing-exams/ [Accessed 13 July 2017].

University of Cambridge Local Examinations Syndicate (2019a). *B2 First Handbook for Teachers.* Available at: https://www.cambridgeenglish.org/Images/CER_6168_V1_APR19_Cambridge_English_First_Handbook_WEB_v3.PDF [Accessed 15 July 2020].

University of Cambridge Local Examinations Syndicate (2019b). *B2 First for Schools Handbook for Teachers.* Available at: https://www.cambridgeenglish.org/Images/CER_2387_V1_APR19_First_for_Schools_Handbook_Update_2018_WEB_v3.PDF [Accessed 21 November 2020].

University of Cambridge Local Examinations Syndicate (2019c). *C2 Proficiency Handbook for teachers*. Available at: https://www.cambridgeenglish.org/Images/168194-c2-proficiency-teachers-handbook.pdf [Accessed 31 August 2020].

University of Cambridge Local Examinations Syndicate (2020a). *Learning Oriented Assessment* [online]. Cambridge English. Research and Validation. Fitness for Purpose. Available at: www.cambridgeenglish.org/research-and-validation/fitness-for-purpose/loa/ [Accessed 19 September 2020].

University of Cambridge Local Examinations Syndicate (2020b). *The Science behind the Test*. [online]. Linguaskill. Available at: https://www.cambridgeenglish.org/exams-and-tests/linguaskill/information-about-the-test/the-science-behind-the-test/20UCLES [Accessed 20 September 2020].

University of Cambridge Local Examinations Syndicate (2020c). *Why Choose us?* [online]. Cambridge English. Available at: https://www.cambridgeenglish.org/why-choose-us/ [Accessed 31 August 2020].

University of Cambridge Local Examinations Syndicate (2020d). *The Write Criteria* [online]. Available at: https://www.youtube.com/watch?v=KdnVXTkQYUc [Accessed 10 October 2020].

University of Cambridge Local Examinations Syndicate (2020e). *Reading and Use of English at C1 and C2*. [online]. Available at: https://www.youtube.com/watch?v=KdnVXTkQYUc [Accessed 10 October 2020].

Urquhart, S. & Weir, C. (1998). *Reading in a Second Language: Process, Product and Practice*. London: Addison Wesley / Longman.

van den Berg, Y. H. M., Segers, E. & Cillessen, A. H. N. (2012). Changing peer perceptions and victimization through classroom arrangements: A field experiment, *Journal of Abnormal Child Psychology, 40*: 403-412. DOI:10.1007/s10802-011-9567-6.

Van Lier, L. (1989). Ethnography: bandaid, bandwagon or contraband? In: C. Brumfit & R. Mitchell, eds., *Research in the Language Classroom.* Hong Kong: Modern English Publications and The British Council, 33-53.

Vandergrift, L. (1999). Facilitating Second language listening comprehension: acquiring successful strategies, *ELT Journal, 53*(3): 168-176.

Vandergrift, L. (2004). Listening to learn or learning to listen?, *Annual Review of Applied Linguistics, 24*, 3-25. Available at: http://dx.doi.org/10.1017/S02267190504000017 [Accessed 10 February 2013].

Vikiru, L. I. (2011). From assessment to learning: the teaching of English beyond examinations, *The Educational Forum, 75*(2): 129-142.

Wall, D. (n.d.). *Examining Washback: What do we know and what is there to explore?* Available at: https://www.yumpu.com/en/document/view/30870210/examining-washback-what-do-we-know-and-what-is-there-alte) [Accessed 14 September 2020].

Wall, D. (1996). Introducing new tests into traditional systems: insights from general education and from innovation theory, *Language Testing 13*(3): 334-354.

Wall, D. (1997). Impact and washback in language testing. In: C. Caplam & D. Corson, eds., *Language Testing and Assessment*. Amsterdam: Kluwer Academic Publishers.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled?, *System, 28*: 499-509.

Wall, D. (2005). *Studies in Language Testing: Vol. 22. The Impact of High-stakes Testing on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*. Cambridge: Cambridge University Press.

Wall, D. & Horak, T. (2006). The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe: phase 1, the baseline study, *TOEFL Monograph Series. Report Number: RR-06-18, TOEFL-MS-34*. Princeton, NJ: Educational Testing Service.

Wallace, M. J. (1991). *Training Foreign Language Teachers: A Reflective Approach*. Cambridge: Cambridge University Press.

Watanabe, Y. (1996). Does Grammar-Translation Come from the Entrance Examination? Preliminary Findings from Classroom-Based Research, *Language Testing, 13*(3): 318-333.

Watanabe, Y. (1997). *The washback effects of the Japanese University entrance examinations of English-classroom-based research*. Unpublished PhD thesis. Lancaster University.

Watanabe, Y. (2001). Does the University Entrance Examination Motivate Learners? A Case Study of Learner Interviews. In: *Trans-Equator Exchanges: A Collection of Academic Papers in Honour of Professor David Ingram* (Vol. 100-110). Akita University.

Watanabe, Y. (2004). Methodology in washback studies. In: L.Y. Cheng, Y. Watanabe & A. Curtis, eds., *Washback in Language Testing: Research Contents and Methods.* Mahwah, New Jersey: Lawrence Erlbaum Associates, 19-36.

Wei, W. (2017). A critical review of washback studies: Hypothesis and evidence. In: R. Al-Mahrooqi, C. Coombe, F. Al-Maamari & V. Thakur, eds., *Revisiting EFL Assessment*. New York: Springer.

Weir, C. J. (1981). Reaction to Morrow paper (1). In: J. C. Alderson & A. Hughes, eds., *Issues in Language Testing. ELT Document 111*. London: The British Council, 26-37.

Weir, C. J. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basignstoke: Palgrave Macmillan.

Weir, C. J., Vidakovic, I. & Galaczi, E. D. (2013). *Measured Constructs: A History of Cambridge English Language Examinations 1913-2012*. Cambridge: Cambridge University Press.

Wenyuan, Z. (2017). The Washback Effect of CET Spoken English Test Upon College English Teaching, *Canadian Social Science, 13*(1): 62-68. DOI:10.3968/9241.

Wesdorp, H. (1982). Backwash effects of language-testing in primary and secondary education, *Journal of Applied Language Study, 1*(1): 40-55.

Wigfield, A. & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation, *Contemporary Educational Psychology, 25*(1): 68-81.

Williams, P. (2014). Squaring the circle: A new alternative to alternative assessment, *Teaching in Higher Education, 19*(5): 565-577.

Wiseman, S. (1961). *Examinations and English Education*. Manchester, England: Manchester University Press.

Xie, Q. (2010). *Test design and use, preparation, and performance: A structural equation modeling study of consequential validity*. Unpublished doctoral dissertation. University of Hong Kong.

Xie, Q. (2013). Does test Preparation work? Implications for score validity, *Language Assessment Quarterly, 10*(2): 196-218.

Xie, Q. & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling, *Language Testing, 30*(1): 49-70.

Xue, G. & Nation, P. (1983). A University word list, *Language Learning and Communication, 3*(2): 215-229.

Zapata, G. (2016). University students' perceptions of integrated performance assessment and the connection between classroom learning and assessment, *American Council on the Teaching of Foreign Languages, 49*(1): 93-104.

Zhan, Y. & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: insights from possible self theories, *Assessment in Education: Principles, Policy and Practice, 21*(1): 71-89.

Zhan, Y. & Wan, Z. H. (2013). Dynamic nature of washback on individual learners: The role of possible selves, *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2013.872769.

Zhan, Y. & Wan, Z. H. (2016). Test takers' beliefs and experiences of a high-stakes computer-based English listening and speaking test, *RELC Journal, 47*(3): 363-376.

9. APPENDIXES

Appendix 1

Cuestionario inicial DNI: _____

Por favor, rellene con sinceridad este cuestionario sobre el curso de inglés al que usted asiste. La información que proporcione será utilizada de forma anónima por Victoria Peña Jaenes para la elaboración de su Tesis Doctoral sobre el efecto que pueden provocar los exámenes de acreditación en los estudiantes de idiomas.

Sección 1
Esta sección me servirá para conocer un poco más sobre su perfil como estudiante de inglés.

1. Señale su edad con un círculo.

12-17 18-20 21-25 ≥ 26

2. Formación (señale con un círculo el máximo nivel académico que ha superado)

a) Educación Primaria
b) Educación Secundaria Obligatoria (E.S.O.)
c) Educación Secundaria no Obligatoria (Bachillerato)
d) Grado / Diplomatura / Licenciatura
e) Máster o doctorado

3. ¿Cuál es su ocupación principal?

a) trabajando
b) estudiando
c) no active / desempleado

4. ¿A qué edad empezó usted a aprender inglés? (en años). Señale la opción que más se ajusta a usted.

< 6 6 -11 12-17 ≥ 18

5. ¿Tiene algún título oficial que acredite su nivel de inglés? En caso afirmativo indique cuál(es) y el año en el que lo(s) obtuvo.

SÍ Título:_____ Año:_____ NO

6. ¿Por qué está aprendiendo inglés? Señale su principal razón (solo una).

a) Necesito tener un título para acabar la carrera o para continuar mi formación
b) Necesito tener un título para acceder al mercado laboral en España o para mejorar mi trabajo
c) Necesito aprender inglés y además me gusta
d) Quiero aprender inglés para estudiar o trabajar en el extranjero
e) Quiero aprender inglés para viajar y por interés personal

7. ¿Cómo de importante es para usted tener un buen dominio del inglés? Señale la opción que más se ajusta a usted en una escala de 1 a 4 donde 1 es "nada importante" y 4 "muy importante".

1	2	3	4

8. ¿Cómo de importante es para usted este curso? Señale la opción que más se ajusta a usted en una escala de 1 a 4 donde 1 es "nada importante" y 4 "muy importante".

1	2	3	4

9.a ¿Tiene usted como objetivo aprobar un examen de acreditación del nivel B2 este año?

SÍ NO

9.b ¿Cómo de importante es para usted aprobar el examen de acreditación de B2? Señale la opción que más se ajusta a su perfil en una escala de 1 a 4 donde 1 es "nada importante" y 4 es "muy importante".

1	2	3	4

9.c En caso de responder afirmativamente a la pregunta 9a, indique cuál.

a) "Cambridge English: First"
b) Examen de acreditación de la Universidad de Jaén CertAcles-UJA
c) APTIS
d) Examen de acreditación de la Escuela Oficial de Idiomas
e) IELTS
f) Otro _____

9.d ¿Por qué ha elegido este examen? (solo una)

a) Porque es el único que conozco
b) Porque me lo han recomendado
c) Porque tiene reconocido prestigio y las principales instituciones lo admiten
d) Porque preparándome para este examen también aprendo inglés
e) Otra _____

10. Señale el número que mejor representa su percepción de la dificultad del examen de "Cambridge English: First". Señale la opción que más se ajusta a usted en una escala de 1 a 4 donde 1 indica que es imposible aprobar el examen en 18 meses de preparación y 4 es que está seguro de que aprobará el examen después de un máximo de 18 meses de preparación.

1	2	3	4

11. ¿Qué aspectos lingüísticos cree que son los más necesarios para aprobar el examen de B2? (rodee solo una opción)

a) las cinco destrezas (*listening, speaking, writing, reading, interacting*).
b) gramática y vocabulario
c) aspectos culturales
d) otros. Indique cuál(es) _____

12. ¿Cuántas horas semanales de media dedica al inglés? (sin contar las clases)

< 1 1-2 3-4 4-5 ≥ 5

13. ¿Es capaz de identificar sus puntos fuertes y sus puntos débiles en inglés?

SÍ NO

14. ¿Sabe cómo puede mejorar su nivel de inglés?

SÍ NO

15.a ¿Cree que puede interactuar con personas no hispanohablantes a través del inglés?

SÍ NO

15.b En caso de que la respuesta a la pregunta 15a sea afirmativa, ¿podría hacerlo de forma adecuada –con suficiente corrección gramatical y léxica y con un estilo apropiado–?

SÍ NO

Sección 2

Esta sección me permitirá conocer sus expectativas respecto a este curso de inglés

16. ¿Qué busca en sus clases de inglés? Señale con una cruz la casilla que representa la opción que más se ajusta a su opinión.

CUESTIÓN	Muy de acuerdo	De acuerdo	No estoy de acuerdo	Total desacuerdo	No sé
1. Quiero que en las clases se practiquen principalmente las destrezas orales: *listening* y *speaking*.					
2. Quiero que en las clases haya tanta práctica del examen como sea posible.					
3. Quiero que en las clases se explique y se practique gramática y vocabulario principalmente.					
4. Quiero que en las clases se enseñen todos los aspectos de la lengua -pronunciación y fonología, gramática y vocabulario, cultura, y las 5 destrezas.					
5. No estoy interesado/a en hacer pruebas periódicamente, busco aprender inglés a mi ritmo.					

17. ¿Conoce los objetivos de este curso?

SÍ NO

18. ¿Conoce los criterios y métodos de evaluación que se van a seguir?

SÍ NO

¡Gracias!

Appendix 2

Cuestionario fin de curso DNI _____

Por favor, rellene con sinceridad este cuestionario sobre el curso de inglés al que usted asiste. La información que proporcione será utilizada de forma anónima por Victoria Peña Jaenes para la elaboración de su Tesis Doctoral sobre el efecto que pueden provocar los exámenes de acreditación en los estudiantes de idiomas.

Sección 1

Esta sección me servirá para conocer su percepción sobre el curso de idiomas al que asiste, sobre sus resultados y su progreso

1. ¿Conoce los objetivos de este curso?

SÍ NO

2. ¿Conoce los criterios y métodos de evaluación que se han seguido?

SÍ NO

3. En su opinión, ¿van en consonancia con las actividades realizadas en clase?

SÍ NO

4. ¿Cuáles de estas actividades se han incluido en sus clases? Puede elegir más de una opción.
   a) Práctica de examen y explicación del formato de un examen de acreditación de B2 (FCE, CertAcles-UJA, otro)
   b) Comentarios y recomendaciones sobre estrategias para realizar el examen de acreditación de B2
   c) Comentarios sobre cómo se han realizado las tareas –*feedback* sobre errores y aspectos positivos–
   d) Actividades para practicar las cinco destrezas en general (*listening, reading, speaking, writing, interacting*)
   e) Ejercicios de gramática y vocabulario basados en el libro de texto
   f) Actividades en los que se utiliza material real (artículos de periódico, vídeos en *youtube* sin modificar…)
   g) Otras. Por favor, indique cuál(es) _____

5. ¿Qué dificultades encuentra al enfrentarse a actividades de comprensión auditiva (*listening*)?

_____
_____

6. ¿Qué dificultades encuentra al enfrentarse a actividades de comprensión lectora (*reading*)?

_____
_____

7. ¿Qué dificultades encuentra al enfrentarse a actividades de expresión e interacción oral (*speaking*)?

_____
_____

8. ¿Qué dificultades encuentra al enfrentarse a actividades de expresión e interacción escrita (*writing*)?

_____
_____

9. ¿Qué dificultades encuentra al enfrentarse a actividades de gramática y vocabulario (*Use of English)*?

_____
_____

10. ¿Sabe cómo puede mejorar su nivel de inglés?

SÍ NO

11. ¿Cree que es capaz de identificar sus puntos fuertes y débiles?

SÍ NO

12.a ¿Cree que es más capaz de interactuar con personas no hispanohablantes a través del inglés que antes de este curso?

SÍ NO

12.b En caso de que la respuesta a la pregunta 13a sea afirmativa, ¿podría hacerlo de forma más adecuada –con mayor corrección gramatical y léxica y con un estilo más apropiado– que antes de este curso?

SÍ NO

13.a ¿En qué medida ha cumplido este curso con sus expectativas? Señale la opción que más se ajusta a usted en una escala de 1 a 4 donde 1 es que no las ha cumplido en absoluto y 4 que las ha superado.

1	2	3	4

13.b Por favor, explique su respuesta

_____
_____

14 . En caso afirmativo, ¿cuántas horas dedica de media a la semana? (sin contar las horas de clase)

< 1 1-2 3-4 4-5 ≥ 5

15.a ¿Prepara de forma diferente el examen “Cambridge English: First” respecto a otros exámenes de inglés?

SÍ NO

15.b En caso afirmativo, ¿en que se diferencia la preparación para este examen?

_____
_____

16.a ¿Se ha preparado de forma diferente para la segunda prueba del examen de “Cambridge English: First”?

SÍ NO

16.b En caso afirmativo, ¿en que se diferencia la preparación para este examen?

_____
_____

17. Señale el número que mejor representa su percepción de la dificultad del examen de “Cambridge English: First” en una escala de 1 a 4 donde 1 es que es imposible aprobar el examen en 18 meses de preparación y 4 es que está seguro de que aprobará después de un máximo de 18 meses de preparación.

1	2	3	4

18.a ¿Tiene intención de presentarse a un examen de acreditación de B2 este año?

SÍ NO

18.b En caso de responder afirmativamente a la pregunta 19a, indique cuál.

g) “Cambridge English: First” (FCE)
h) Examen de acreditación de la Universidad de Jaén CertAcles-UJA
i) APTIS
j) Examen de acreditación de la Escuela Oficial de Idiomas
k) IELTS
l) Otro _____

18.c En caso de que su respuesta a la pregunta 18a haya sido afirmativa, ¿cree que el hecho de presentarse al examen le ha animado a trabajar de forma más intensa?

SÍ NO

18.d En caso de que su respuesta a la pregunta 18a haya sido afirmativa, ¿cómo de importante es para usted aprobar el examen de acreditación de B2? Señale la opción que más se ajusta a usted en una escala de 1 a 4 donde 1 es “nada importante” y 4 es “muy importante”.

1	2	3	4

¡Gracias!

Sección 2

Por favor, rellene esta sección <u>únicamente</u> si va a presentarse a un examen de acreditación de B2 (FCE, CertAcles-UJA, …). Esta sección me permitirá saber cómo se ha preparado para el examen de acreditación de B2, cuál es su percepción hacia él y a usted le permitirá reflexionar sobre su estudio y preparación.

19.a En su opinión ¿Le ha indicado su profesor(a) cómo puede estudiar y preparar este examen?

SÍ NO

19.b En caso afirmativo, ¿cree que estos consejos le ayudarán a mejorar su rendimiento en el examen? SÍ NO

19.c ¿Cree que estos consejos le ayudarán a mejorar su inglés en general?

SÍ NO

20. Por favor, ordene estas actividades de mayor (7) a menor eficacia (1) para aprobar el examen de acreditación de B2.

a) Práctica de examen y explicación del formato del examen de acreditación de B2
b) Comentarios y recomendaciones sobre estrategias para realizar el examen de B2
c) Comentarios sobre cómo se han realizado las tareas –*feedback* sobre errores y aspectos positivos–
d) Actividades para practicar las cinco destrezas en general (*listening, reading, speaking, writing, interacting*)
e) Ejercicios de gramática y vocabulario basados en el libro de texto
f) Actividades en los que se utiliza material real (artículos de periódico, vídeos en *youtube* sin modificar…)
g) Otras. Por favor, indique cuál(es) _____

21. ¿Cómo se siente cuando se realizan actividades tipo antes de un examen de acreditación de B2? Marque una sola opción.

a) Me sirven bastante para afianzar mi preparación y sacar el máximo partido de mis conocimientos
b) Me sirven porque me permiten sentirme más seguro(a) ya que sé lo que voy a encontrar el día del examen
c) No me sirven porque son más fáciles que el examen y me aburren
d) Me resultan estresantes y me ponen nervioso(a)

22. En su opinión ¿tienen los exámenes de acreditación de B2 y las prácticas de examen que se realizan en clase un grado de dificultad similar?

SÍ NO

23.a ¿Ha superado las pruebas de "Cambridge English: First" realizadas a lo largo del curso?

a) Todas o la mayoría
b) Algunas pero la progresión ha sido positiva
c) Algunas pero la progresión no siempre ha sido positiva
d) Ninguna

23.b ¿Cómo se siente ante resultado adversos? Por favor, elija la opción que más se ajusta a su situación

a) No me preocupa en exceso porque no estudié lo suficiente
b) Las pruebas no son importantes, lo importante es el examen oficial
c) Me hace ser consciente de cómo puedo mejorar
d) Me anima a trabajar más
e) Me desanima y me hace sentir que puedo suspender
f) Otra. Por favor, indique cuál_____

23.c Si usted ha elegido las opciones c o d en la pregunta 23a ¿A qué cree que se debe?

a) Falta de estudio o de asistencia a clase
b) Suerte
c) Estado físico (cansancio, enfermedad)
d) Carga de estudio en el colegio, instituto o universidad o carga de trabajo
e) Diferentes grados de dificultad del examen
f) Otra(s). Por favor indique cuál(es) _____

24. Por favor rodee la(s) forma(s) en que se prepara para el examen de acreditación de B2.

a) Hago prácticas de examen (libro de texto, páginas web con práctica y consejos, etc.)
b) Estudio gramática y vocabulario que creo que van a entrar en el examen y realizo ejercicios
c) Estudio inglés de forma general: contenidos vistos en clase, leo libros o prensa, veo series o películas en inglés, escucho la radio, estoy en contacto con hablantes nativos (tándem, intercambios…)
d) Leo consejos útiles sobre qué hacer y no hacer en el examen
e) Analizo y reflexiono sobre el formato de examen (qué se espera de mí en cada parte, el tiempo que tengo y cómo gestionarlo, aspectos que mejoran mi calificación)
f) Estudio el *feedback* que me da mi profesor(a) y el que yo obtengo de los ejercicios que hago
g) Hago mis deberes y estudio para los exámenes del curso
h) Otra(s) _____

25.a ¿Sabe cómo puede mejorar su puntuación/ rendimiento en el examen?

SÍ NO

25.b En caso de que la respuesta a la pregunta 25a haya sido afirmativa, ¿de dónde ha obtenido esta información?

a) Del profesor(a)
b) De información oficial de la institución examinadora (Cambridge, British Council, Universidad de Jaén, EOI)
c) De compañeros o amigos
d) De mi reflexión y experiencia personal

e) De libros o páginas con consejos sobre el examen

26. ¿Continuará estudiando inglés después de aprobar el examen de acreditación de B2?
SÍ NO

27. ¿Cree que sus resultados en el examen de acreditación de B2 reflejarán su esfuerzo durante este curso? SÍ NO

¡Gracias!

Appendix 3

Questionnaire for teachers

Please answer the following questions honestly. The information you provide in this questionnaire will be data used anonymously in the PhD Thesis of Victoria Peña Jaenes.

Section 1

In this section I would like to learn about your teaching qualifications and experience

1. What is your primary degree?

_____
_____

2. Do you possess any of these English-language teaching qualifications? Please circle.
   a. TEFL
   b. CELTA
   c. DELTA
   d. MA in Education
3. Have you received Cambridge training?
   a) Yes, seminars organised every year
   b) Yes, in-house training at the beginning of the school year
   c) Yes, ongoing in house training
   d) No, but I have received training for other official exams
   e) No, I haven't received any training for official accreditation exams
4. If your answer to question 3 was positive, what did you receive training on?
   a) Marking and assessment
   b) Format of the exam and test-taking strategies
   c) Exam practice activities
   d) Resources available
   e) How to motivate your students
   f) Other, please specify _____-
5. How many teaching hours do you have per week? Please circle.

< 10 11-15 16-20 21-25 26-30 > 30

6. What levels do you teach according to the Common European Framework of Languages? Please circle.

A1 A2 B1 B2 C1 C2

7. What type of courses are you teaching at the moment? Please circle.
    a. Extensive General English courses (≥60 teaching hours)
    b. Intensive General English courses (< 60 teaching hours)
    c. Extensive Cambridge English: First test preparation course (≥60 teaching hours)
    d. Intensive Cambridge English: First test preparation course (< 60 teaching hours)

8. How old are your B2 students? Please circle.
    a. Teenagers
    b. Adults
    c. I don't teach B2 level

Section 2

In this section I would like to learn about your opinion about language courses and "Cambridge English: First Test".

9. How are test preparation courses different from general English courses?

_____
_____
_____
_____

10. The educational goals of this course are connected to the assessment criteria?
    a) I strongly disagree
    b) I disagree
    c) I agree
    d) I strongly agree
    e) I don't know

11. If a student passes Cambridge First test, it means that s/he has a B2 level. Please circle.
    a) I strongly disagree
    b) I disagree
    c) I agree
    d) I strongly agree
    e) I don't know

Why?

_____
_____

12. From your point of view and on the basis of your experience, how important are these factors for success in Cambridge First test? Please circle.

5 Very Important 1 Not important at all 2 I don't know

1. Student aptitude and ability	1	3	4	5	2
2. Educational experience	1	3	4	5	2
3. Openness to instruction and willingness to follow the teachers guidance	1	3	4	5	2
4. Maturity	1	3	4	5	2
5. Motivation	1	3	4	5	2
6. Age	1	3	4	5	2
7. Class attendance	1	3	4	5	2
8. Participation in class	1	3	4	5	2
9. Personal work	1	3	4	5	2
10. Exposure to English outside the class	1	3	4	5	2
11. Exam preparation (attending a B2 preparation course/ lessons)	1	3	4	5	2

13. In your experience, how long does a student need to pass from a B1 level to a B2 level? Please circle.

31-60 teaching hours
60-120 teaching hours
120-180 teaching hours
>180 teaching hours

14. If you are teaching Cambridge preparation courses, do you feel the exam is the most important factor influencing your lessons?

a) No, it does not influence my lessons at all
b) Yes, it is one more aspect together with discipline, syllabus, textbook, interests
c) Yes, it is the most relevant factor and my lessons depend on the proximity of the test
d) I don't know

15. How does "Cambridge English: First Test" affect your students?

_____
_____
_____
_____

16. Do you feel your professional value is judged on the basis of (you can choose more than one)

a. Your students' performance on tests throughout the year?
b. Your students' performance on achievement tests?
c. Students' and parents' opinions
d. B2 First
e. Other. Please specify _____

17. How do you feel about it?

_____
_____
_____
_____

Section 3

In this section I would like to learn about your lessons and the activities carried out in class.

18. Have you explained the objectives of the course you are teaching?
    a. Yes
    b. No
    c. I don't know
19. Have you explained the marking criteria and methods?
    a. Yes
    b. No
    c. I don't know
20. Have you explained the connection between the tasks carried out in class or for homework and the exam?
    a. Yes
    b. No
    c. I don't know
21. Do you foster self-assessment?
    a. Yes,
    b. No
    c. I don't know

    If so, how?

_____
_____

22. How do you motivate your students for the exam?

_____
_____

23. Which of these activities have you carried out in class?

a. Exam practice and explanation of the format of "Cambridge English: First Test"
b. Comments and recommendations about test taking strategies
c. Feedback
d. Listening, reading, speaking, writing and interaction, not connected to Cambridge test
e. Grammar and vocabulary activities based on the textbook
f. Activities using authentic material
g. Other. Please specify

_____

24. Please order the activities in question 22 from the least effective (1) to the most effective (7) to pass "Cambridge English: First Test".

25. How often do you do this in your classes?

	Every lesson	Every week	Every month	I don't do it
Work on speaking/oral interaction in general				
Work on writing in general				
Work on listening in general				
Work on reading in general				
Work on grammar/vocab				
Do exam practice and explain the format of Cambridge English: First Test				
Use the textbook (SB, WB, CD-ROM)				
Use resource books (Grammar in Use…)				
Use exam preparation material				
Use authentic material				

Regarding feedback (please, answer only if you have circled this option in question 23

26. When do you give feedback? You can choose more than one
    a. While the students are doing the activity (I interrupt and correct the mistakes)
    b. Just after finishing the task
    c. Later (in the following lesson, following days, when we correct it)

27. What role do grammar and vocabulary play in your classes?
    a. It is the main aspect in my lessons
    b. It is used as support for skills practice
    c. It has a very small part in my classes
    d. Other _____

28. Do you give homework?
    a. Yes, please specify how often _____
    b. No
    c. I don't know

29. If the answer is positive please give an example of the type of homework you give your students for a weekend

    _____

30. If the answer to question 28 is positive. How do you correct…?
    a. A writing activity (e.g. an essay) _____
    b. An activity from the WB _____
    c. A reading activity _____

Section 4

In this section I will obtain information about your students and their learning process.

31. What are the main problems your students encounter in the reading paper?

_____

32. What are the main problems your students encounter in the listening paper?

_____

33. What are the main problems your students encounter in the use of English paper?

_____

34. What are the main problems your students encounter in the speaking paper?

_____

35. What are the main problems your students encounter in the writing paper?

_____

36. Do they encounter the same problems when working on these skills in general? Why?

    _____

    _____

_____

_____

37. My students are more self-directed and autonomous.

1 I totally disagree 4 I totally agree

1	2	3	4

38. My students are aware of their strengths and weaknesses

1 I totally disagree 4 I totally agree

1	2	3	4

39. What has your students' performance been throughout the year?
    a. Most of them have passed most of the mock exams
    b. Most of them have had a positive progression and have passed the last mock exam
    c. Most of them have passed some mock tests but the progression was not always positive
    d. Most of them did not pass any of the mock tests
40. What are the causes of your students' results
    a. They need to study more and attend lessons more regularly
    b. Luck
    c. Physical condition (they were tired or sick)
    d. Workload (outside the lessons)
    e. Some exams were more difficult than others
    f. other. Please specify _____
41. What percentage of your students are taking the official exam this year?

    _____

42. Do you think that the fact that they are taking the official exam encourages them to work harder?
    a. Yes
    b. No
    c. I don't know
43. What aspects do you think your students have especially improved? (speaking, listening…)

_____

_____

_____

Thank you!

Appendix 4

Exam conditions regulations[9]

Please follow these guidelines. The data obtained from the exam results will be used anonymously in the Ph D Thesis of Victoria Peña Jaenes. It will also offer interesting information about your students' profiles and performance.

Before the exam starts

1. Students must be informed about the format of the test and the instructions should be read aloud before the exam starts.
2. There must be a clock that all students can see and the invigilator must write the start and the finish time.
3. The invigilator cannot give any information about the answers or any hints.
4. Students must sit in lines at a distance from each other that prevents them from viewing the work of others, intentionally or otherwise.
5. Candidates are not permitted to keep any electronic items, such as mobile phones, recording devices, mp3 players, smart watches etc in the exam room during the exam.
6. Students are not permitted to have any unauthorised material (pencil case, books, notebooks…) on their desk. Everything must be kept in their bags, which must be closed under their chairs.

During the exam

7. Once the exam has started, doubts cannot be solved.
8. Students are not allowed to speak during the test.
9. There must be an invigilator (the teacher) in the room, who supervises the students at all times.
10. Students cannot leave the room before finishing the test.
11. Students must answer directly on the exam paper (grammar and vocabulary test) or on the answer sheet (Cambridge Exams).
12. Students must use a pen.

After the exam

[9] Based on:
Cambridge International Examinations (2014). *Cambridge handbook 2015 (International). Regulations for conducting Cambridge examinations* [online]. Available at http://www.cie.org.uk/images/178636-cambridge-handbook-2015-international.pdf [Retrieved 1st September 2017].
University of Cambridge Local Examinations Syndicate (2016). *Cambridge English. Regulations*. [online]. Available at http://www.cambridgeenglish.org/images/84609-faq-regulations.pdf [Retrieved 1st September 2017].

13. The invigilator must collect, count and check that all the exams have been collected and that the code gap has been filled in.
14. The invigilator must put all the exam together using a elastic band and write the date of the exam and the group code.

Thank you very much for your cooperation.

Appendix 5

Cambridge English: First Test

A: Format

Paper	General Description	Timing
Reading and Use of English (25%)	**7 parts** **Part 1 – multiple choice (1 mark per correct question)** **Part 2 – open cloze (1 mark per correct question)** **Part 3 – word formation (1 mark per correct question)** **Part 4 – key word transformation (2 marks per correct sentence*)** **Part 5 – multiple choice (2 marks per correct question)** **Part 6 – gapped text (2 marks per correct question)** **Part 7 – multiple matching (1 mark per correct question)**	**1 h. 15 min.**
Writing (25%)	**2 parts** **Part 1 – essay** **Part 2 – article, email, essay, letter, report, review (you must complete one from a choice of three)**	**1h 20 min.**
Listening (25%)	**4 parts** **Part 1 – multiple choice (1 mark per correct question)** **Part 2 – sentence completion (1 mark per correct question)** **Part 3 – multiple matching (1 mark per correct question)** **Part 4 – multiple choice (1 mark per correct question)**	**40 min.**
Speaking (25%)	**4 parts** **Part 1 – conversation with interlocutor** **Part 2 – individual long turn + response from other candidate** **Part 3 – conversation between candidates** **Part 4 – discussion**	**14 min.**

B: Tips and recommendations

B.1 Reading and Use of English

Do

- Read the titles and subtitles of the texts.
- Read each text carefully before answering the questions to get an overall impression of it.
- Check the words around the gap carefully (Parts 1 and 2).

* Each sentence is divided into two parts; if the part is totally correct, one mark is awarded.

- Remember that the words you need to write in Part 3 might have to change into a negative or a plural.
- Check your spelling in all parts of the test.
- Write between two and five words as your answer in Part 4.
- Check that the completed paragraph makes sense in the passage as a whole (Part 6).
- Keep the development of the text in mind.
- Read the questions carefully and check each option against the text before rejecting it. (Part 6)

Don't

- Don't try to answer any questions without **referring** carefully to the text.
- Don't spend too much **time** on any one part of the paper.
- Don't assume that if the same word appears in the text as well as in an option, this means you have located the answer. The correct answer usually paraphrases the content of the text.
- Don't leave the base word in Part 3 unchanged

B.2 Writing

Do

- Read the whole question thoroughly and underline important parts.
- Make a plan for each answer, including ALL points.
- Expand the points, using relevant reasons and examples.
- Write in paragraphs.
- Use a range of grammar and vocabulary.
- Check tense endings, irregular past tenses, plural forms and word order in sentences.
- Use language that is appropriately formal or informal for the task.
- Choose a Part 2 question that you feel confident you can write about.
- Write clearly, so that the examiner can read your answer.

Don't

- Don't misspell key words which appear on the question paper.
- Don't use the exact words from the question paper too much.
- Don't waste time writing addresses for a letter. They are not required.
- Don't worry if you run slightly over the word limit.

B.3 Listening

Do

- Listen to and read the instructions.
- Use the preparation time to read through the question and think about the context.
- Look carefully at what is printed before and after the gap in Part 2 and think about the kind of information that you are listening for.
- Check your answers when the recording is played a second time.
- Answer all the questions – even if you're not sure.

Don't

- Don't rephrase what you hear in Part 2; write down the exact word(s) or figure(s) that you hear on the recording.
- Don't spend too much time on a question you are having difficulty with as you may miss the next question.
- Don't rush to choose an answer just because you hear one word or phrase – concentrate on the overall meaning. (Parts 1, 3 and 4). Remember that the correct answer usually paraphrases what you hear.

B.4 Speaking

Do

- Listen carefully to the instructions and questions during the test and respond appropriately.
- Speak clearly, so that both the interlocutor and assessor can hear you.
- Use all the opportunities you are given to speak in the test –but do not dominate your partner or interrupt them abruptly– and extend your responses whenever possible.
- Ask for clarification of instructions or a question if you're not sure.
- Initiate discussion as well as responding to what your partner says.
- Make full use of the time so that the examiner who is listening hears plenty of your English.

Don't

- Don't prepare long answers in advance, or learn and practise speeches.
- Don't leave long or frequent pauses.
- Don't worry about being interrupted by the examiner. This shows you have spoken enough.

Sources:

Cambridge English De (2012). Cambridge English First DOs and DON'Ts. [online]. Available at https://cambridgeesolde.wordpress.com/2012/11/12/cambridge-english-first-dos-and-donts/ [Retrieved: 31st August 2017].

University of Cambridge Local Examinations Syndicate (2013a). Cambridge English First. Cambridge: University of Cambridge Local Examinations Syndicate.

University of Cambridge Local Examinations Syndicate (2017). *Cambridge English First (FCE). Exam Format*. [online]. Available at http://www.cambridgeenglish.org/exams/first/exam-format/ [Retrieved: 30th August 2017].

Appendix 6

Entry Vocabulary Test[10]

DNI _____

La información que proporcione será utilizada de forma anónima por Victoria Peña Jaenes para la elaboración de su Tesis Doctoral sobre el efecto que pueden provocar los exámenes de acreditación en los estudiantes de idiomas. Además, le ayudará a comprobar su nivel de inglés y los aspectos en los que puede mejorar.

Instructions

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

l	business	
2	clock	_____ part of a house
3	horse	_____ animal with four legs
4	pencil	_____ something used for writing
5	shoe	
6	wall	

You answer it in the following way.

l	business	
2	clock	___6__ part of a house
3	horse	___3__ animal with four legs
4	pencil	___4__ something used for writing
5	shoe	
6	wall	

Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are business, clock, and shoe.

If you have no idea about the meaning of a word, do not guess. But if you think you might know the meaning, then you should try to find the answer.

Time : 20 minutes

[10] Source:
Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.
Xue, G. and Nation, P. (1983). "A University word list", *Language Learning and Communication, 3*(2): 215-229.

Part 1

Question 1

1 belt
2 climate
3 executive
4 notion
5 palm
6 victim

_____ idea
_____ inner surface of your hand
_____ strip of leather worn around the waist

Question 2

1 acid
2 bishop
3 chill
4 ox
5 ridge
6 structure

_____ cold feeling
_____ farm animal
_____ organization or framework

Question 3

1 bench
2 charity
3 jar
4 mate
5 mirror
6 province

_____ long seat
_____ help to the poor
_____ part of a country

Question 4

1 boot
2 device
3 lieutenant
4 marble
5 phase
6 vein

_____ army officer
_____ a kind of stone
_____ tube through which blood flows

Question 5

1 apartment
2 candle
3 draft
4 horror
5 prospect
6 timber

_____ a place to live
_____ chance of something happening
_____ first rough form of something written

Question 6

1 betray
2 dispose
3 embrace
4 injure
5 proclaim
6 scare

_____ frighten
_____ say publicly
_____ hurt seriously

Question 7

1 encounter
2 illustrate
3 inspire
4 plead
5 seal
6 shift

_____ meet
_____ beg for help
_____ close completely

Question 8

1 assist
2 bother
3 condemn
4 erect
5 trim
6 whirl

_____ help
_____ cut neatly
_____ spin around quickly

Question 9

1 annual
2 concealed
3 definite
4 mental
5 previous
6 savage

_____ wild
_____ clear and certain
____ happening once a year

Question 10

1 dim
2 junior
3 magnificent
4 maternal
5 odd
6 weary

_____ strange
_____ wonderful
_____ not clearly lit

Part 2

Question 11

1 balloon
2 federation
3 novelty
4 pail
5 veteran
6 ward

_____ bucket
_____ unusual interesting thing
_____ rubber bag that is filled with air

Question 12

1 alcohol
2 apron
3 hip
4 lure
5 mess
6 phase

_____ stage of development
_____ state of untidiness or dirtiness
_____ cloth worn in front to protect your clothes

Question 13

1 apparatus
2 compliment
3 ledge
4 revenue
5 scrap
6 tile

_____ expression of admiration
_____ set of instruments or machinery
_____ money received by the Government

Question 14

1 bulb
2 document
3 legion
4 mare
5 pulse
6. tub

_____ female horse
_____ large group of soldiers or people
_____ a paper that provides information

Question 15

1 blend
2 devise
3 hug
4 lease
5 plague
6 reject

_____ mix together
_____ plan or invent
_____ hold tightly in your arms

Question 16

1 gloomy
2 gross
3 infinite
4 limp
5 slim
6 vacant

_____ empty
_____ dark or sad
_____ without end

Question 17

1 abolish
2 drip
3 insert
4 predict
5 soothe
6 thrive

_____ bring to an end by law
_____ guess about the future
_____ calm or comfort someone

Question 18

1 concrete
2 era
3 fiber
4 loop
5 plank
6 summit

_____ circular shape
_____ top of a mountain
_____ a long period of time

Question 19

1 bleed
2 collapse
3 precede
4 reject
5 skip
6 tease

_____ come before
_____ fall down suddenly
_____ move with quick steps and jumps

Question 20

1 casual
2 desolate
3 fragrant
4 radical
5 unique
6 wholesome

_____ sweet-smelling
_____ only one of its kind
____ good for your health

Part 3

Question 21

1 antics
2 batch _____ foolish behavior
3 connoisseur _____ a group of things
4 foreboding _____ person with a good knowledge of art or music
5 haunch
6 scaffold

Question 22

1 auspices
2 dregs _____ confused mixture
3 hostage _____ natural liquid present in the mouth
4 jumble _____ worst and most useless parts of anything
5 saliva
6 truce

Question 23

1 casualty
2 flurry _____ someone killed or injured
3 froth _____ being away from other people
4 revelry _____ noisy and happy celebration
5 nut
6 seclusion

Question 24

1 apparition
2 botany _____ ghost
3 expulsion _____ study of plants
4 insolence _____ small pool of water
5 leash
6 puddle

Question 25

1 arsenal
2 barracks _____ happiness
3 deacon _____ difficult situation
4 felicity _____ minister in a church
5 predicament
6 spore

Question 26

1 acquiesce
2 bask _____ to accept without protest
3 crease _____ sit or lie enjoying warmth
4 demolish _____ make a fold on cloth or paper
5 overhaul
6 rape

Question 27

1 blaspheme
2 endorse _____ slip or slide
3 nurture _____ give care and food to
4 skid _____ speak badly about God
5 squint
6 straggle

Question 28

1 clinch
2 jot _____ move very fast
3 mutilate _____ injure or damage
4 smolder _____ burn slowly without flame
5 topple
6 whiz

Question 29

1 auxiliary
2 candid _____ bad-tempered
3 luscious _____ full of self-importance
4 morose _____ helping, adding support
5 pallid
6 pompous

Question 30

1 dubious	
2 impudent	_____ rude
3 languid	_____ very ancient
4 motley	_____ of many different
5 opaque	kinds
6 primeval	

Appendix 7

End-of-Course Vocabulary Test[11]

Código _____

La información que proporcione será utilizada de forma anónima por Victoria Peña Jaenes para la elaboración de su Tesis Doctoral sobre el efecto que pueden provocar los exámenes de acreditación en los estudiantes de idiomas. Además, le ayudará a comprobar su nivel de inglés y los aspectos en los que ha mejorado.

Instructions

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

l	business	
2	clock	_____ part of a house
3	horse	_____ animal with four legs
4	pencil	_____ something used for writing
5	shoe	
6	wall	

You answer it in the following way.

l	business	
2	clock	___6__ part of a house
3	horse	___3__ animal with four legs
4	pencil	___4__ something used for writing
5	shoe	
6	wall	

Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are business, clock, and shoe.

If you have no idea about the meaning of a word, do not guess. But if you think you might know the meaning, then you should try to find the answer.

Time: 20 minutes

[11] Source:
Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.
Xue, G. & Nation, P. (1983). "A University word list", *Language Learning and Communication, 3*(2): 215-229.

Part 1

Question 1

1 copy
2 event _____ end or highest point
3 motor
4 pity _____ this moves a car
5 profit _____ thing to be like another
6 tip

Question 2

1 blanket
2 excel _____ holiday
3 bridge _____ good quality
4 merit _____ wool covering used on beds
5 pillow
6 vacation

Question 3

1 accident
2 debt _____ loud deep sound
3 fortune _____ something you must pay
4 pride _____ having a high opinion of yourself
5 roar
6 thread

Question 4

1 arrange
2 develop _____ to bend or move from a vertical position
3 lean
4 owe _____ managing business and affairs
5 prefer
6 seize _____ grow

Question 5

1 burst
2 concern _____ break open
3 deliver _____ make better
4 fold _____ take something to someone
5 improve
6 urge

Question 6

1 abandon
2 dwell _____ live somewhere
3 survive _____ follow in order to catch
4 pursue _____ leave something for good
5 stump
6 cease

Question 7

1 bake
2 connect _____ keep within a certain size
3 inquire _____ walk without a purpose
4 limit _____ join together
5 recognize
6 wander

Question 8

1 overcome
2 endure _____ suffer patiently
3 grasp _____ join wool threads together
4 knit _____ hold firmly with your hands
5 combine
6 catch

Question 9

1 chubby
2 stripped _____ slim
3 normal _____ steady
4 naked _____ nude
5 slender
6 stable

Question 10

1 aware
2 random _____ average
3 extraordinary _____ best or most important
4 normal _____ knowing or realizing something
5 confident
6 supreme

Part 2

Question 11

1 analysis
2 curb _____ eagerness
3 gravel _____ loan to buy a house
4 mortgage _____ small stones mixed with sand
5 scar
6 zeal

Question 12

1 cavalry
2 eve _____ small hill
3 ham _____ day or night before a holiday
4 mound
5 steak _____ soldiers who fight from horses
6 switch

Question 13

1 guitar
2 desk _____ string instrument
3 lecture _____ seat without a back or arms
4 sermon
5 stool _____ speech given by a priest
6 trumpet

Question 14

1 artillery
2 creed _____ a kind of tree
3 hydrogen _____ system of belief
4 maple _____ large gun on wheels
5 pork
6 streak

Question 15

1 chart
2 forge _____ map
3 mansion _____ large beautiful house
4 outfit _____ place where metals are made and shaped
5 sample
6 volunteer

Question 16

1 contemplate
2 extract _____ reflect deeply
3 gamble _____ bring back to consciousness
4 launch
5 provoke _____ make someone get cross
6 revive

Question 17

1 demonstrate
2 embarrass _____ have a rest
3 heave _____ break suddenly into small pieces
4 obscure
5 relax _____ make someone feel shy or nervous
6 shatter

Question 18

1 correspond
2 send _____ exchange letters
3 lurk _____ hide and wait for someone
4 seek
5 furious _____ feel angry about something
6 resent

Question 19

1 decent
2 frail _____ weak
3 harsh _____ concerning a city
4 incredible _____ difficult to believe
5 municipal
6 specific

Question 20

1 achieve
2 conceive _____ making people obey a rule
3 grant _____ finish successfully
4 link _____ money paid for services
5 enforcement
6 fee

Part 3

Question 21

1 alabaster
2 sand _____ small barrel
3 jar _____ soft white stone
4 keg _____ tool for shaping wood
5 rasp
6 drill

Question 22

1 benevolence
2 convoy _____ kindness
3 lien _____ set of musical notes
4 octave _____ speed control for an
5 stint engine
6 throttle

Question 23

1 bourgeois
2 brocade _____ middle class people
3 consonant _____ row or level of something
4 prelude _____ cloth with a pattern or gold
5 stupor or silver threads
6 tier

Question 24

1 cure
2 alibi _____ priest
3 bandage _____ release from prison early
4 parole _____ medicine to put on wounds
5 salve
6 vicar

Question 25

1 alkali
2 banter _____ light joking talk
3 coop _____ a rank of British
4 mossaic nobility
5 stealth _____ picture made of small
6 viscount pieces of glass or stone

Question 26

1 dissipate
2 flaunt _____ steal
3 impede _____ scatter or vanish
4 loot _____ twist the body about
5 squirm uncomfortably
6 vie

Question 27

1 contaminate
2 cringe _____ write carelessly
3 immerse _____ move back because of fear
4 peek _____ put something under water
5 relay
6 scrawl

Question 28

1 blurt
2 dabble _____ walk in a proud way
3 dent _____ kill by squeezing
4 pacify someone's throat
5 strangle _____ say suddenly without
6 swagger thinking

Question 29

1 illicit
2 lewd _____ immense
3 mammoth _____ against the law
4 slick _____ wanting revenge
5 temporal
6 vindictive

Question 30

1 indolent
2 nocturnal _____ lazy
3 obsolete _____ no longer used
4 torrid _____ clever and tricky
5 translucent
6 wily

Appendix 8

Entry Grammar test[12]

DNI _____

La información que proporcione será utilizada de forma anónima por Victoria Peña Jaenes para la elaboración de su Tesis Doctoral sobre el efecto que pueden provocar los exámenes de acreditación en los estudiantes de idiomas. Además, le ayudará a comprobar su nivel de inglés y los aspectos en los que puede mejorar.

Instructions

This is a grammar test. You must mark the correct letter. There is only one correct answer per question.

Example

I don't know _____ about him.

a. nothing
b. anything
c. something
d. no thing

Time: 12 minutes

[12] Based on:

Bell, J. & Thomas, A. (2014). *Gold First Coursebook*. Essex: Pearson Education Limited.
Brook-Hart, G. & Owen, D. (2011). *Complete First Certificate*. Madrid: Cambridge University Press.
Cambridge University Press (2017). *Cambridge English. Resources* [online]. Available at Cambridge [Retrieved 8th August 2017].
Capel, A. & Sharp, W. (2014). *Objective First 4th Edition*. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.
Hewings, M. (1999). *Advanced Grammar in Use*. Cambridge: Cambridge University Press.
Latham-Koenig, C. & Oxenden, C. (2014*). English File Upper Intermediate Student's Book.* Oxford: Oxford University Press.
Latham-Koenig, C. & Oxenden, C. (2015). *English File Advanced.* Oxford: Oxford University Press.
Murphy, R. (2004). *English Grammar in Use*. Cambridge: Cambridge University Press.
Vince, M. (2014). *Language Practice for First*. London: Macmillan Education.

1. Chinese will be more important in the future, _____ it?
    a. doesn't
    b. isn't
    c. wasn't
    d. won't
2. You _____!
    a. always complain
    b. complain always
    c. are always complaining
    d. always are complaining
3. By the time we arrived, they _____.
    a. left
    b. leaving
    c. have left
    d. had left
4. In 20 years' time what _____?
    a. are you doing
    b. will you do
    c. are you going to do
    d. will you be doing
5. They _____ the house if they'd known about the noisy neighbours.
    a. wouldn't have bought
    b. hadn't bought
    c. wouldn't buy
    d. won't have bought
6. Matthew _____ in Glasgow.
    a. used to live
    b. use to living
    c. use to lived
    d. was use to live
7. William had _____ yesterday.
    a. cut his hair
    b. cutting his hair
    c. his hair cut
    d. been cutting his hair
8. Mike _____ his car last week.
    a. crashed
    b. was crash
    c. was crashing
    d. crash
9. Alicia _____ seven o'clock.
    a. seldom get up before
    b. gets seldom up before
    c. seldom gets up before
    d. seldom before gets up
10. It was _____ a rainy day we decided to go to the cinema.
    a. very
    b. so
    c. such
    d. absolutely

11. He's the same age _____ my sister.
    a. as
    b. more
    c. than
    d. to
12. We bought _____ for our house.
    a. new furniture
    b. a new furniture
    c. new furnitures
    d. a new pieces of furniture
13. A web browser is _____ you use when you want to search the Internet.
    a. that
    b. which
    c. what
    d. –
14. _____ the plane departed late, we arrived in Saint Petersburg on time.
    a. In spite
    b. Although
    c. Despite
    d. Despite of
15. I wish you wouldn't always insist _____ sitting at the back of the cinema.
    a. at
    b. in
    c. on
    d. with
16. Peter asked me if I _____ Nat that morning.
    a. have seen
    b. had seen
    c. saw
    d. will see
17. I really miss _____ in the city.
    a. live
    b. to live
    c. living
    d. lives
18. Look! He's fallen asleep. He _____ have been tired.
    a. can't
    b. may
    c. might
    d. must
19. I think _____ dogs are humans' best friends.
    a. the
    b. –
    c. a
    d. a few
20. Do you know what time _____?
    a. starts the football match
    b. the football match starts
    c. does the football match start
    d. will the football match start

21. The two cars for sale were in poor condition, so I didn't buy _____ of them.
    a. neither of them
    b. both of them
    c. either of them
    d. each of them
22. Too much rubbish is being dumped in _____.
    a. sea
    b. the sea
    c. a sea
    d. some sea

Appendix 9

End-of-course Grammar Test[13]

Código _____

La información que proporcione será utilizada de forma anónima por Victoria Peña Jaenes para la elaboración de su Tesis Doctoral sobre el efecto que pueden provocar los exámenes de acreditación en los estudiantes de idiomas. Además, le ayudará a comprobar su nivel de inglés y los aspectos en los ha mejorado.

Instructions

This is a grammar test. You must mark the correct letter. There is only one correct answer per question.

If you have no idea about the answer, do not guess. But if you think you might know the meaning, then you should try to find the answer.

Example

I don't know _____ about him.

e. Nothing
f. Anything
g. Something
h. No thing

Time: 15 minutes

[13]Based on:

Bell, J. & Thomas, A. (2014). *Gold First Coursebook*. Essex: Pearson Education Limited.
Brook-Hart, G. & Owen, D. (2011). *Complete First Certificate*. Madrid: Cambridge University Press.
Cambridge University Press (2017). *Cambridge English. Resources* [online]. Available at Cambridge [Retrieved 8th August 2017].
Capel, A. & Sharp, W. (2014). *Objective First 4th Edition*. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.
Hewings, M. (1999). *Advanced Grammar in Use*. Cambridge: Cambridge University Press.
Latham-Koenig, C. & Oxenden, C. (2014*). English File Upper Intermediate Student's Book.* Oxford: Oxford University Press.
Latham-Koenig, C. & Oxenden, C. (2015). *English File Advanced.* Oxford: Oxford University Press.
Murphy, R. (2004). *English Grammar in Use*. Cambridge: Cambridge University Press.
Vince, M. (2014). *Language Practice for First*. London: Macmillan Education.

1. We should study more, _____?
    a. Isn't it?
    b. Doesn't we
    c. Shouldn't we?
    d. Couldn't I
2. Excuse me! _____ for the bus to Glasgow?
    a. Do you wait
    b. Are you wait
    c. Will you be waiting
    d. Are you waiting
3. We _____ when Paul called
    a. didn't left
    b. hadn't leave
    c. hadn't left
    d. haven't left
4. Next time we talk I'll _____ from my holiday
    a. return
    b. have returned
    c. have been returned
    d. returned
5. You _____ if I hadn't told you.
    a. wouldn't knew
    b. wouldn't known
    c. hadn't known
    d. wouldn't have known
6. I wish you _____ me about the accident yesterday.
    a. tell
    b. wouldn't tell
    c. hadn't told
    d. haven't told
7. _____ board games when you were young?
    a. Did you use to play
    b. Have you used to play
    c. Did you use to playing
    d. Did you use to played
8. I _____ tomorrow
    a. am having painted my daughter's bedroom
    b. have my daughter's bedroom painted
    c. am having my daughter's bedroom paint
    d. am having my daughter's bedroom painted
9. That's very _____ news
    a. depress
    b. depressive
    c. depressed
    d. depressing
10. Brad _____ very good at dancing
    a. has been never

b. never has been
c. has never been
d. been has never

11. I really think that apologizing is _____ you can do.
    a. not as much as
    b. a little
    c. the least
    d. as far as
12. This dress is _____. I can't afford it.
    a. enough expensive
    b. expensive enough
    c. too much expensive
    d. too expensive
13. RAM, _____ is Random Access Memory, is an important thing to consider when buying a computer.
    a. that
    b. which
    c. who
    d. what
14. Owning a motorbike has several advantages. _____, you can go wherever you want.
    a. First of all
    b. As a result
    c. Personally
    d. Besides this
15. You really need to be responsible _____ your actions.
    a. on
    b. for
    c. of
    d. in
16. Can you let me _____ her?
    a. to tell
    b. telling
    c. tell
    d. told
17. You can't _____ left your wallet in the pub. You paid for the taxi to come home.
    a. have
    b. haven't
    c. had
    d. be
18. Can you tell me _____?
    a. where can I change some money
    b. where I can change some money
    c. where I will change some money
    d. where did I change some money
19. They told us they'd arrived an hour before

a. “we arrived an hour ago” they said
b. “we arrived an hour before” they say
c. “we’d arrive an hour before” they said
d. “we’d arrived an hour ago” they said

20. My father used to work in _____ prison as a cleaner.
a. –
b. the
c. a
d. these

21. My friends gave us _____ great news
a. a
b. several
c. –
d. a few

22. I hope I will be _____ my grandfather, he was a very talented man
a. as
b. like
c. such as
d. -

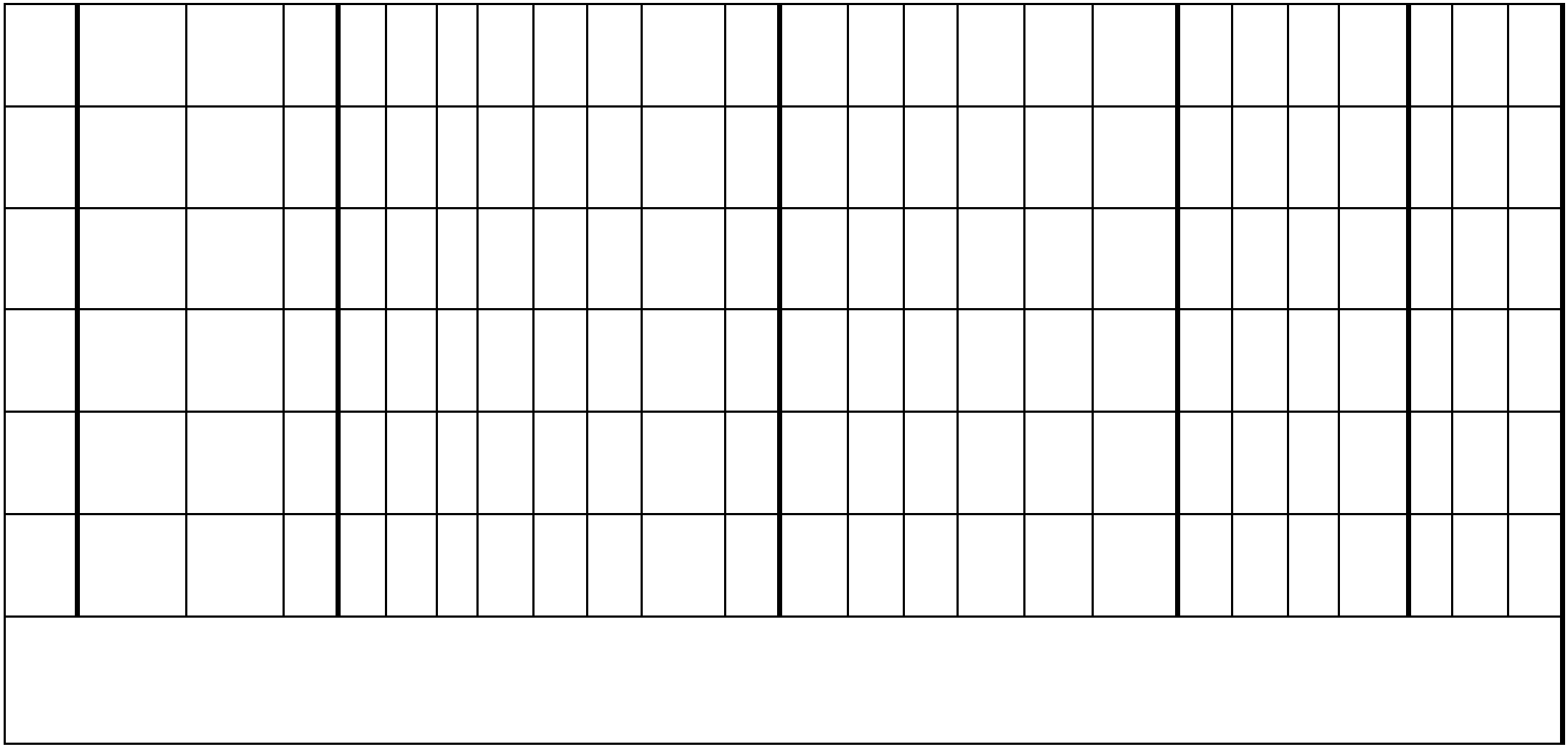
Appendix 10

Cambridge-General English Observation Schedule: Classroom Observation[14]

Sheet number: _____ **Institution:** CEB/ CEALM **Level:** B2.1/ B2.2 **Group:** _____ **Date:** _____

Time	**Activities and episodes**			**Content**								**Test**						**Materials**				**HW**		
	Teacher actions	Student actions	Approach	Writing	Translatio	Pron/phon	Speaking	Listening	Reading	Grammar/ vocabular	Culture	Cambridge	Other	Format	Strategies	Feedback	Practice	Textbook	Authentic	Exam prep	Extra support	Textbook	Exam prep	Other

[14] Based on:
Spada and Fröhlich (1995), Flanders (1970), Moskowitz (1971), Green (2007), Mickan and Moterram (2009:17)



Instructions

1. **Activities:** warm up, homework correction, grammar/vocabulary explanation, drilling activities, setting homework, house keeping (talking about dates for the exam, attendance, justifications for absences, discipline…), leave blank if it is related to exam
2. **Episodes:** instructions, clarification, asking for clarification, observation, correction, feedback (in a specific activity carried out in class, specify type of feedback), performance (teacher keeps record of students' performance), response (choral, specific, open-ended), asking questions, jokes, laughter
3. **Approach:** whole class activity, pairs, individual, groups, process, product
4. **Writing:** essay, article, letter, report, email, summary, review, other
5. **Translation**
6. **Pronunciation/phonetics:** when it is an activity only devoted to that. If it is part of feedback, this box shouldn't be ticked
7. **Speaking:** debates, interaction st-teacher, interaction st-st
8. **Listening**
9. **Reading**
10. **Grammar and vocabulary:** state content point
11. **Culture**
12. **Format:** parts of the exam, marking…
13. **Strategies:** read before listening, looking for paraphrases
14. **Feedback**
15. **Practice:** exam tasks
16. **Textbook:** student book, workbook, CD-ROM, specify page number and activity…
17. **Authentic material:** non adapted material from any source
18. **Exam preparation:** activities obtained from an exam preparation resource book or website
19. **Extra support:** photocopies with additional grammar, vocabulary
20. **HW exam preparation:** can be for Cambridge or for achievement test at the end of the course. Specify.